

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METİNDEN GÖRÜNTÜ ÜRETİMİ

Melike Nur YEĞİN

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman

Prof. Dr. Mehmet Fatih AMASYALI

Haziran, 2025

**T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

METİNDEN GÖRÜNTÜ ÜRETİMİ

Melike Nur YEĞİN tarafından hazırlanan tez çalışması 24.06.2025 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Mehmet Fatih AMASYALI
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

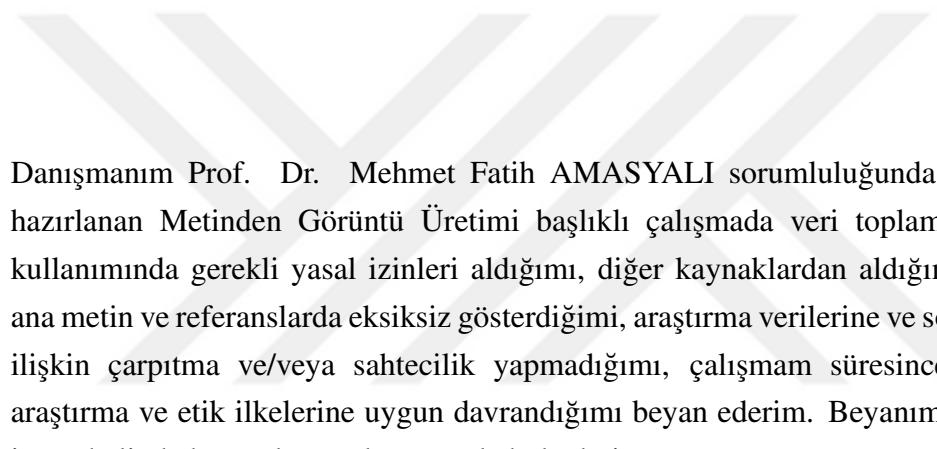
Prof. Dr. Mehmet Fatih AMASYALI, Danışman
Yıldız Teknik Üniversitesi

Prof. Dr. Sırma ÇEKİRDEK YAVUZ, Üye
Yıldız Teknik Üniversitesi

Doç. Dr. Akhtar JAMIL, Üye
FAST National University

Prof. Dr. Songül VARLI, Üye
Yıldız Teknik Üniversitesi

Dr. Öğr. Üyesi Fatma GÜMÜŞ, Üye
Milli Savunma Üniversitesi



Danışmanım Prof. Dr. Mehmet Fatih AMASYALI sorumluluğunda tarafimca hazırlanan Metinden Görüntü Üretimi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğim, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Melike Nur YEĞİN

İmza



Bu çalışma, TÜBİTAK Bilim İnsanı Destek Programları Başkanlığı tarafından yürütülen 2211/A - Yurt İçi Doktora Burs Programı ile desteklenmiştir.

*"İlim ilim bilmektir, ilim kendin bilmektir
Sen kendini bilmez isen, ya nice okumaktır."*

Yunus Emre



TEŞEKKÜR

Doktora yıllarım, kendimi yeniden keşfetmekle geçti. Bir yandan araştırma konumun diğer yandan çocukların hızla geliştiği bu süreç benim için bilimsel anlamanın ötesinde bir karakter inşası olarak anlam kazandı. Bu bağlamda kattığı tüm değerler için, kıymetli hocam Prof. Dr. Mehmet Fatih AMASYALI'ya sonsuz teşekkürlerimi sunuyorum. Onun öğrencisi olmakla hep gurur duyacağım ve onun aktarımlarını geleceğe taşımak adına elimden geleni yapacağım.

Tez çalışmam boyunca zamanını ve emeğini harcayan tez komitemin değerli üyeleri Prof. Dr. Sırma YAVUZ ve Doç. Dr. Akhtar JAMIL hocalarımı da içten teşekkürlerimi sunuyorum. Aynı zamanda tez çalışmam süresince sağlamış oldukları desteklerden ötürü TÜBİTAK Bilim İnsanı Destekleme Daire Başkanlığı'na da teşekkür ediyorum.

Varlığıyla bana güç veren kıymetli eşim Erdem YEĞİN'e ve kızlarımız Ayşe Sare ile Fatma Zehra'ya bu süreçte bana ilham ve motivasyon kaynağı oldukları için teşekkür ediyorum. Ayrıca desteklerini ve dualarını her zaman yanımdaya hissettiğim değerli aileme ve arkadaşlarımı da teşekkürlerimi sunuyorum.

Melike Nur YEĞİN

İÇİNDEKİLER

KISALTMA LİSTESİ	ix
ŞEKİL LİSTESİ	xi
TABLO LİSTESİ	xiii
ÖZET	xiv
ABSTRACT	xv
1 GİRİŞ	1
1.1 Tezin Konusu	1
1.2 Tezin Amacı	2
1.3 Tezin İçeriği	2
2 LİTERATÜR İNCELEMESİ	4
2.1 Temel Kavramlar	4
2.2 Üretken Modeller	5
2.2.1 Varyasyonel Otokodlayıcı	6
2.2.2 Üretken Çekişmeli Ağ	10
2.2.3 Otoregresif Model	12
2.2.4 Akış Tabanlı Model	12
2.2.5 Enerji Tabanlı Model	13
2.2.6 Dönüştürücü Modeli	14
2.2.7 Difüzyon Modeli	14
3 ÇOK KİPLİ VARYASYONEL OTOKODLAYICILARLA METİNDEN GÖRÜNTÜ ÜRETİMİ	16
3.1 Giriş	16
3.2 İlgili Çalışmalar	17
3.3 Yöntem	20
3.3.1 Sabit Katsayılarla Ağırlıklandırma	22
3.3.2 Otomatik Ağırlıklandırma	22
3.4 Deneyler	24

3.4.1	Gaussian Kiplerle Denemeler	24
3.4.2	MNIST / FashionMNIST Denemeleri	26
3.5	Sonuç	30
4	ÜRETKEN DİFÜZYON MODELLERİ: GÜNCEL TEORİK GELİŞMELERİN BİR İNCELEMESİ	33
4.1	Giriş	33
4.2	Difüzyon Modellerinin Temel Çalışmaları	34
4.2.1	Gürültü Arındırın Olasılıksal Difüzyon Modelleri	35
4.2.2	Gürültü Şartlı Skor Ağları	39
4.2.3	Stokastik Diferansiyel Denklemler ile Skor-tabanlı Modelleme	43
4.2.4	Temel Çalışmaların Aralarındaki İlişki ve Sınırlamaları . . .	46
4.3	Difüzyon Modellerinin Teorik Gelişmeleri	48
4.3.1	Eğitim Tabanlı Gelişmeler	49
4.3.2	Eğitimden Bağımsız Gelişmeler	64
4.3.3	Gelişmelerin Değerlendirilmesi	73
4.4	Sonuç	78
5	DİFÜZYON MODELLERİ İLE METİNDEN GÖRÜNTÜ ÜRETİMİ	81
5.1	İlgili Çalışmalar	81
5.1.1	Stable Diffusion	82
5.1.2	Anlamsal Uyumun Çıkarım Zamanında İyileştirilmesi . . .	83
5.2	Sonuçlanmayan Yöntemler	86
5.2.1	Çapraz Dikkat Haritaları ile Nesneleri Yerleştirme	86
5.2.2	Yerleşim Bilgisi için Büyük Dil Modeli İnce Ayarlama . . .	89
5.2.3	Sahneyi Planlama	91
5.3	Çok Kipli Büyük Dil Modeli Rehberliği	92
5.3.1	Rehberlik Adımları	92
5.3.2	Nitel Deneyler	94
5.3.3	Nicel Deneyler	97
6	ÇOK KİPLİ BÜYÜK DİL MODELİ REHBERLİĞİNDE METİNDEN GÖRÜNTÜ ÜRETİMİ	100
6.1	Giriş	101
6.2	İlgili Çalışmalar	103
6.3	Yöntem	104
6.3.1	Yerleştirme	105
6.3.2	Kendini Düzeltme	105
6.3.3	Akıl Yürütmeye	105

6.3.4	Hizalama	105
6.4	Deneysel	106
6.4.1	Deney Kurulumu	106
6.4.2	Nitel Deneysel	108
6.4.3	Nicel Deneysel	109
6.5	Analiz	110
6.5.1	Ablasyon Çalışması	110
6.5.2	Genelleme Yeteneği	111
6.6	Sonuç	112
7	SONUÇ	114
KAYNAKÇA		116
A	BÖLÜM 6'DA ÖNERİLEN KOMUTLAR	133
A.1	Yerleştirme için MLLM Komutu	133
A.2	Kendini Düzeltme için MLLM Komutu	134
A.3	Akıl Yürütmeye için MLLM Komutu	135
A.4	Hizalama için MLLM Komutu	135
B	BÖLÜM 6'DA ÖNERİLEN VERİ KÜMESİ	137
TEZDEN ÜRETİLMİŞ YAYINLAR		144

KISALTMA LİSTESİ

CoT	Düşünce zinciri
DDPM	Gürültü arındırılan olasılıksal difüzyon modelleri
EBM	Enerji tabanlı model
ELBO	Kanıt alt sınır optimizasyonu
FID	Frechet başlangıç mesafesi
GAN	Üretken çekişmeli ağ
GPT	Üretken öneğitimli dönüştürücü
GPU	Grafik işleme birimi
IS	Başlangıç puanı
JS	Jensen-Shannon
KL	Kullback-Leibler
LDM	Gizli difüzyon modeli
LLM	Büyük dil modeli
MCMC	Markov zinciri monte carlo
MoE	Uzman karışımı
NCSN	Gürültü şartlı skor ağıları
NLL	Negatif log-olasılık
ODE	Adi diferansiyel denklem
PoE	Uzman çarpımı
RNN	Tekrarlayan sinir ağıları
SD	Stable Diffusion
SDE	Stokastik diferansiyel denklem
SGD	Stokastik gradyan düşümü

VAE	Varyasyonel otokodlayıcı
VI	Varyasyonel çıkışım
VLB	Varyasyonel alt sınır
VQ	Vektör nicelemeli



ŞEKİL LİSTESİ

Şekil 2.1	Varyasyonel Otokodlayıcı	8
Şekil 2.2	Yeniden parametrelendirme hilesi	9
Şekil 2.3	Üretken Çekişmeli Ağ	11
Şekil 2.4	Otoregresif model	12
Şekil 2.5	Akış tabanlı model	13
Şekil 2.6	Enerji tabanlı model	14
Şekil 2.7	Dönüştürücü modeli	14
Şekil 2.8	Difüzyon modeli	15
Şekil 3.1	Ağırlıklandırılmış uzman karışıntılarında μ ve Σ 'ların bulunması .	22
Şekil 3.2	Kiplerin üretebilirliğinin hesaplanması	23
Şekil 3.3	Kiplerin üretilebilirliğinin hesaplanması	24
Şekil 4.1	Teorik gelişmelerin kategorileri	48
Şekil 4.2	Farklı gürültü dağılımları için sonuçlar[45].	49
Şekil 4.3	Farklı anahtarlama noktaları(β) için DAED'den örnekler[75].	54
Şekil 4.4	Gizli Skor Tabanlı Üretken Model (LSGM) 'in mekanizması[101].	58
Şekil 4.5	Gizli Difüzyon Modeli (LDM)'in mekanizması [102].	58
Şekil 4.6	Düzelte akışı [117] ($N \geq 2$) için görüntü oluşturma görevinde çok az sayıda adımla iyi örnekler üretir.	60
Şekil 4.7	Hoogeboom ve diğerleri[120] tarafından önerilen Argmax akışları	61
Şekil 4.8	Hoogeboom ve diğerleri[120] tarafından önerilen Çok terimli(multinomial) difüzyon	62
Şekil 5.1	Stable Diffusion'ın çıkarım zamanı [214]	82
Şekil 5.2	Stable Diffusion'ın zorlanması.	83
Şekil 5.3	GLIGEN'in kapılı öz dikkat katmanı[215]	85
Şekil 5.4	GLIGEN ile yerleştirme[215]	86
Şekil 5.5	Çapraz dikkat haritaları	87
Şekil 5.6	Çapraz dikkat haritaları ile ilk nesnenin yerini bulma	88
Şekil 5.7	Çapraz dikkat haritaları ile yerleştirme	88
Şekil 5.8	Tek adımda tüm nesneleri yerleştirme	92
Şekil 5.9	Nesneleri adım adım yerleştirme	92
Şekil 5.10	Çok Kipli Büyük Dil Modeli Rehberliğine Genel Bakış	93
Şekil 5.11	LLM-yerleştirme ile görüntü üretimi	93

Şekil 5.12	LLM-kendini-düzelme ile görüntü üretimi	93
Şekil 5.13	LLM-hizalama ile görüntü üretimi	94
Şekil 5.14	LLM-akıl-yürütmeye ile görüntüyü hizalama	95
Şekil 5.15	Mevcut yöntemlerle karşılaştırma	96
Şekil 5.16	Hizalama ve Akıl yürütme ile hizalamanın sonuçları	97
Şekil 6.1	MLLM Rehberliğinin son teknoloji devlerini geride bırakması . .	100
Şekil 6.2	MLLM Rehberliğinin 4 adımı	102
Şekil 6.3	MLLM Rehberliğinin SDXL ve benzer çalışmaları geride bırakması	108
Şekil 6.4	MLLM Rehberliğinin konumsal ve 3 boyutlu konumsal ilişkileri yakalamadaki üstünlüğü	109
Şekil 6.5	İkinci ve üçüncü adımların etkisini gösteren ablasyon çalışması .	111
Şekil 6.6	MLLM rehberliğinin çeşitli rehberler ve metinden görüntü üreten modellere genelleştirilmesi	111

TABLO LİSTESİ

Tablo 3.1	PoE ile tüm kiplerin üretilmesi	25
Tablo 3.2	MoE ile tüm kiplerin üretilmesi	26
Tablo 3.3	PoE ile MNIST veri kümesinin sonuçları	28
Tablo 3.4	PoE ile FashionMNIST veri kümesinin sonuçları	29
Tablo 3.5	MoE ile MNIST veri kümesinin sonuçları	31
Tablo 3.6	MoE ile FashionMNIST veri kümesinin sonuçları	32
Tablo 4.1	DDPM örnekleme algoritması	39
Tablo 4.2	NCSN örnekleme algoritması	43
Tablo 4.3	VE-SDE için tahmin edici-düzeltilci örnekleme algoritması	46
Tablo 4.4	VP-SDE için tahmin edici-düzeltilci örnekleme algoritması	47
Tablo 4.5	CIFAR-10 görüntüyü üretimi	75
Tablo 4.5	CIFAR-10 görüntüyü üretimi (devamı)	76
Tablo 4.6	CelebA(64X64) görüntüyü üretimi	77
Tablo 4.7	ImageNet(64x64) görüntüyü üretimi	78
Tablo 5.1	LLM-yerleştirme ile görüntüyü üretimi algoritması	94
Tablo 5.2	LLM-kendini-düzeltilme ile görüntüyü üretimi algoritması	94
Tablo 5.3	LLM-hizalama ile görüntüyü üretimi algoritması	95
Tablo 5.4	LLM-akıl-yürütmeye ile görüntüyü hizalama algoritması	95
Tablo 5.5	Farklı metriklerin nicel değerlendirme puanları	98
Tablo 5.6	Yerleştirme ve düzeltme yöntemlerinin kıyaslamaları	99
Tablo 5.7	Hizalama ve akıl yürüterek hizalama yöntemlerinin kıyaslamaları	99
Tablo 6.1	MLLM rehberliği algoritması	106
Tablo 6.2	Metriklerin insan değerlendirme ile korelasyonları	108
Tablo 6.3	Kıyaslama sonuçları	110

ÖZET

Metinden Görüntü Üretimi

Melike Nur YEĞİN

Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi

Danışman: Prof. Dr. Mehmet Fatih AMASYALI

Metinden görüntü üretimi, yapay öğrenme tekniklerini kullanarak verilen metin şartı ile uyumlu görüntüler üretmeyi hedefleyen bir araştırma alanıdır. Günümüzde iş süreçlerine verimlilik katması nedeniyle oldukça önemli bir konu haline gelmiştir. Bu görev için başlarda varyasyonel otokodlayıcılar ve üretken çekişmeli ağlar kullanılırken günümüzde çoğunlukla difüzyon modelleri kullanılmaktadır. Bu tez kapsamında da öncelikle varyasyonel otokodlayıcılarla ardından difüzyon modelleri ile çalışmalar yapılmıştır.

Metinden görüntü üreten modeller genel olarak iki farklı açıdan değerlendirilir. Birincisi görüntü kalitesi, ikincisi ise görüntü-metin hizalamasıdır. Yapay öğrenme alanının genelinde olduğu gibi bu görev için de veriyi ve eğitilen modelin parametre sayısını artırarak daha iyi sonuçlara ulaşmak mümkündür. Ancak bu yaklaşım modelin yeniden eğitilmesi için uzun bir süreç gerektirmekle beraber görüntü-metin hizalamasını sağlamayı da garanti etmemektedir. Bu tez kapsamında önerilen yeni yöntemlerle modelleri yeniden eğitmeden görüntü-metin hizalamasının iyileştirileceği gösterilmiştir.

Anahtar Kelimeler: Difüzyon modelleri, Büyük dil modelleri, Görüntü-metin hizalaması.

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ABSTRACT

Text-to-Image Generation

Melike Nur YEĞİN

Department of Computer Engineering
Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Mehmet Fatih AMASYALI

Text-to-image generation is a research topic that aims to develop new methods using machine learning techniques to generate images that are compatible with a given text condition. Nowadays, it has gained a great importance due to adding efficiency to many business areas. While this task was initially performed with variational autoencoders and generative adversarial networks, in these days diffusion models are mostly used. In this thesis, we carried out our first studies with variational autoencoders and then continued with diffusion models.

Text-to-image generation models are generally evaluated from two different perspectives. The first is image fidelity, and the second is image-text alignment. As with machine learning in general, it is possible to obtain better results by increasing the amount of data and model parameters. However this approach, comes with a long process which is required for retraining the model and also does not guarantee image-text alignment. In this thesis, we showed that image-text alignment can be improved with the proposed methods without retraining the models.

Keywords: Diffusion models, Large language models, Image-text alignment.

**YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING**

1 GİRİŞ

Bu bölümde öncelikle tezin konusu hakkında kısa bir giriş yapılacak ardından tezin amacı belirtilecektir. Son kısmında ise tezin içeriği ve organizasyonu hakkında bilgi verilecektir.

1.1 Tezin Konusu

Metin koşullu görüntü sentezleme ya da kısaca metinden görüntü üretimi, yapay öğrenme tekniklerini kullanarak verilen metin girdisine uyumlu görüntü elde etmeyi hedefleyen bir alandır. Doğal dilde verilen tanımlara karşılık gelen görsellerin üretilmesi 2010’ların ortalarında gündeme gelmiş ve günümüzde gerçek fotoğraf veya insan eserinden ayırt edilemeyecek görseller üretebilen modeller geliştirilmiştir.

Yapay öğrenme ile metinden görüntü üreten modeller pek çok alanda iş süreçlerini daha hızlı ve verimli hale getirmeleri nedeniyle oldukça önemli hale gelmiştir. Bu modeller, yaratıcı düşünme gerektiren tüm tasarım alanlarında ve fikirlerini görsellerle açıklayarak akılda kalıcı olmayı hedefleyen reklamcılık ve pazarlama alanlarında çalışanların ilham bulmaları için vazgeçilmez araçlardır.

Metinden görüntü üretimi konusu son yıllarda oldukça popüler hale gelmiş ve bu görevi daha iyi gerçekleştirmek için bir çok yöntem ortaya çıkmıştır. Bu görev için başlangıçta varyasyonel otokodlayıcılar ve üretken çekişmeli ağlar kullanılırken günümüzde metinden görüntü üreten modeller çoğunlukla difüzyon modelleridir. Bu tez kapsamında, öncelikle varyasyonel otokodlayıcılar üzerinde çalışmalar gerçekleştirilmiş ardından difüzyon modellerinin çıkışları ile bu alandaki başarısı göz önüne alınarak difüzyon modelleri ile çalışmalara devam edilmiştir.

1.2 Tezin Amacı

İnternet teknolojisinin yaygınlaşması ile elde edilen büyük veri, donanım teknolojisinin gelişmesi ile ortaya çıkan güçlü grafik işlemcilerle birleşince tarihte ilk defa makinelerin insan zekasına yaklaşabilmesi gündeme gelmiştir. Sadece metinden görüntü üretimi konusu özelinde değil yapay öğrenme alanının genelinde veriyi çoğaltarak ve/veya modelin parametre sayısını artırarak daha iyi çıktılara ulaşmak mümkündür. Ancak bu yaklaşım donanımsal imkanlara sahip olmayı gerektirmekle beraber modelin yeniden eğitilmesi de uzun bir süreç almaktadır. Diğer yandan yöntemlerin başka parametrelerini iyileştirmek başarıyı artırmak da mümkündür.

Metinden görüntü üreten modellerin başarısı başlıca 2 kategoride ele alınmaktadır. Birincisi üretilen görüntünün kalitesi, ikincisi ise üretilen görüntünün istenen metin ile ne kadar uyumlu olduğunu ölçen görüntü-metin hizalamasıdır. Daha yüksek çözünürlüklü veriler ve daha büyük çaplı modeller kullanılarak üretilen görüntünün kalitesini iyileştirebilmek mümkündür ancak görüntü-metin hizalamasını iyileştirmek için bu yaklaşım yeterli olmamaktadır.

Bu tez kapsamında metinden görüntü üreten modellerin başarısında görüntü-metin hizalaması dikkate alınarak istenen metin girdileri ile daha uyumlu görüntü çıktılarının elde edilmesi hedeflenmiştir. Bu bağlamda önerilen yeni yöntemlerle görüntü-metin hizalamasının daha çok veri ve/veya daha büyük model kullanmadan da iyileştirilebileceği gösterilmiştir.

1.3 Tezin İçeriği

Günümüzde pek çok yapay öğrenme modeli metinden görüntü üretimini başarıyla gerçekleştirebilmektedir. Bu tez çalışmasında ilk olarak çok kipli varyasyonel otokodlayıcılarda kiplerin ağırlıklandırılmasının metinden görüntü üretimi performansına etkisi ile ilgili bir çalışma yapılmıştır. İlerleyen süreçte difüzyon modellerinin tanıtılması, metinden görüntü üretimi görevinde bir dönüm noktası olmuş ve bu modellerin teorik çalışmalarının araştırıldığı bir inceleme makalesi üzerinde çalışılmıştır. İnceleme makalesinin ardından yapılan çalışmada ise difüzyon modellerine büyük dil modelleri tarafından rehberlik yapılması önerilmiş ve bu sayede metinden görüntü üretiminde önemli bir konu olan görüntü-metin hizalamasının iyileştirilmesi sağlanmıştır.

Aşağıda verilen tez organizasyonunda tezin ilerleyen bölümleri kısaca açıklanmıştır:

- **Bölüm 2:** Literatür incelemesi bölümündür. Üretken modellemenin temel kavramları açıklanmakta ve başlıca üretken modellerden bahsedilmektedir.
- **Bölüm 3:** Çok kipli varyasyonel otokodlayıcılarla metinden görüntü üretiminde kiplerin ağırlıklandırılmasının etkisi araştırılmaktadır.
- **Bölüm 4:** Üretken difüzyon modellerinin temel çalışmaları ve güncel teorik gelişmeleri detaylı bir şekilde incelenmektedir.
- **Bölüm 5:** Difüzyon modelleri ile metinden görüntü üretimine giriş yapılarak deneme-yanılma sürecine kısaca değinilmektedir.
- **Bölüm 6:** Görüntü-metin hizalamasını iyileştirmek için Çok-kipli büyük dil modellerinin metinden görüntü üretimine rehberlik yaptığı bir yöntem önerilmektedir.
- **Bölüm 7:** Sonuç bölümündür. Tez kapsamında yapılan çalışmalar tartışılmakta ve gelecek çalışmalar için öneriler yer almaktadır.

2 LİTERATÜR İNCELEMESİ

Literatürde yer alan çoğu üretken model Bayesçi öğrenmeye dayanır. Bu nedenle literatür çalışmasında öncelikle Bayesçi öğrenmenin temel kavramları hikayeleştirilerek açıklanacak ardından başlıca üretken modellerden bahsedilecektir.

2.1 Temel Kavamlar

Bir okulun bahçesinde utangaç bir arkadaş gördünüz ve bu okulda 2 bölüm var. Matematik-Doktora ve İşletme. Sizce bu arkadaş hangi bölümde olabilir? İlk bakışta hemen Matematik-Doktora öğrencisi olduğunu düşündünüz. Ancak bir şeyleri ihmali ediyorsunuz. Sadece utangaç olma bilgisine göre karar veriyorsunuz. Okulda toplasanız kaç tane Matematik-Doktora öğrencisi vardır ki? Bu bölümün öğrencilerinden çok daha fazla İşletme öğrencisi vardır. Bu durumda arkadaşın İşletme öğrencisi olma ihtimali artar. İşte öğrencilerin böümlere dağılımı ile ilgili bu ön bilgiye Öncül(Prior) diyoruz. Bir de hepimizin bildiği gibi Matematik-Doktora öğrencilerinin utangaç olma oranı ve İşletme öğrencilerinin utangaç olma oranı var. Bu bilgiye de Olasılık(Likelihood) diyoruz. Marjinal olasılık(Marginal likelihood) ise gözlemlenen durumun koşulsuz olasılığı yani bu okuldaki tüm öğrencilerin utangaç olma oranıdır. Öncül ve olasılık oranlarının çarpımının marjinal olasılığa oranı utangaç olduğunu bildiğimiz bir arkadaşın Matematik-Doktora öğrencisi olma ihtimalini gösterir. Buna da Sonsal(Posterior) denir.

Bayes kuralı Eşitlik 2.1'de verilmektedir. Burada $p_\theta(z|x)$ sonsal, $p_\theta(x|z)$ olasılık, $p_\theta(z)$ öncül ve $p_\theta(x)$ marjinal olasılık olarak ifade edilir.

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)} \quad (2.1)$$

Bayesçi öğrenmede parametrelerin birlikte gelme olasılığı dikkate alınarak buna göre bir dağılım çıkartılır. Marjinal olasılık oranı, özellik sayısı ve özelliklerin alabileceği değerlerin sınırlı olduğu durumlarda kolaylıkla hesaplanabilmektedir. Ancak özelliklerin çok sayıda olduğu ve sürekli değerler alabildiği durumlarda işler daha karmaşık bir hale gelmektedir. Marjinal olasılık Eşitlik 2.2'de verilmektedir.

$$p_{\theta}(x) = \int p_{\theta}(z)p_{\theta}(x|z)dz \quad (2.2)$$

Bir şey üretmek için öncelikle o şeyin özelliklerini çıkartıp onu iyice tanıtmamız gereklidir. Örneğin, bir masa üretmek istiyoruz, tüm masaların sahip olabileceği özelliklerden oluşan bir uzayı var olduğunu düşünürsek bu uzaydaki her bir özellik (ör. malzemesi, bacaklarının sayısı/şekli v.b.) gizli değişkenlere karşılık gelir. Eşitlik 2.1'de verilen Bayes kuralını yeniden ifade edersek, x modellemek istediğimiz veri, z gizli değişkenler, $p_{\theta}(x)$ verinin olasılıksal dağılımı, $p_{\theta}(z)$ gizli değişkenlerin olasılıksal dağılımı, $p_{\theta}(x|z)$ verilen gizli değişkenler ile istenen veriyi üretmenin dağılımı, $p_{\theta}(z|x)$ ise istenen veriyi üretebilecek gizli değişkenlerin dağılımıdır.

Örneğin elimizdeki x örneklerinin insan yüzü verisi ve z gizli değişkenlerinin kaşların şekli, kulakların yeri, burun büyülüüğü v.b. yüz özellikleri olduğunu varsayıyalım. Bayesçi çıkarım yapabilmek için bu özelliklerin birlikte gelme olasılıklarını hesaplamak gereklidir ancak bu değişkenler sürekli değerler alabiliyorken bu hesabı yapmak oldukça zordur. Marjinal olasılık kavramının gözlemlenen durumu sağlayacak tüm olasılıkların toplamı olduğundan bahsetmiştim. Yani bu problem için "bu insan yüzünü oluşturabilecek tüm olası değerlerin toplamı nedir?" gibi içinden çıkışmaz bir soru ile karşılaşmaktayız. Bu hesaplamanın yapılması için önerilen yöntemlerden bazıları ise Markov Chain Monte Carlo ile örnekleme veya Varyasyonel çıkarım yapmaktadır.

2.2 Üretken Modeller

Bu bölümde temel üretken modellerden bahsedilecektir. Bunlar sırasıyla Varyasyonel Otokodlayıcı, Üretken Çekişmeli Ağ, Otoregresif model, Dönüşürücü Model, Akış tabanlı model ve Enerji tabanlı model olarak ele alınmıştır. Ayrıca bu bölümde Difüzyon modellerine giriş yapılarak önceki modellere göre avantajlarına değinilmiştir.

2.2.1 Varyasyonel Otokodlayıcı

Bu bölümde öncelikle Bayesçi öğrenmenin hesaplama çözümlerinden biri olan Varyasyonel çıkarım yöntemi açıklanacaktır. Sonrasında ise bu çözümü uygulamak için geliştirilen Varyasyonel otokodlayıcı (Variational autoencoder) modeli açıklanacaktır.

2.2.1.1 Varyasyonel Çıkarım

Özelliklerin çok sayıda ve sürekli (continuous) olduğu durumlarda marjinal olasılık hesaplamasının zorlu (intractable) bir integral gerektirdiğinden önceki bölümde bahsedilmiştir. Bu ifade ile başa çıkabilmek için Markov Chain Monte Carlo(MCMC) gibi örneklemeye tabanlı çözümler önerilmiştir ancak bu yöntemler her bir veri için maliyetli bir örneklemeye döngüsü içerdikinden büyük veri kümelerinde verimli olmamaktadır. Bunun yerine marjinal olasılık hesaplaması için Kingma ve Welling tarafından varyasyonel alt sınırın optimizasyonu [1] önerilmiştir. Varyasyonel Çıkarım (VI) yöntemi ile değerlendirmesi daha kolay olan basit bir dağılım ör. Gaussian, kullanarak gerçek dağılım modellenir ve Kullback-Leibler (KL) iraksaması metriğini (dağılımlar arasında ne kadar fark olduğunu söyleyen bir metrik) kullanarak bu iki dağılım arasındaki fark en aza indirilmeye çalışılır.

Çözülmlesi zor olan sonsalı tahminleyecek bir tanıma modelimiz $q_\phi(z|x)$ olduğunu varsayıyalım. Bu bölümde gizli değişkenleri ifade eden z 'ye kısaca “kod” $q_\phi(z|x)$ 'ye de olasılıksal “kodlayıcı” (probabilistic encoder) diyeceğiz. Olasılıksal kodlayıcımız verilen bir x verisini üretetebilmek için z kodlarının olası değerleri üzerine bir dağılım (Gaussian v.b.) üretir. Benzer yaklaşımla $p_\theta(z|x)$ 'e de olasılıksal “kodçözcü” (probabilistic decoder) diyeceğiz. Olasılıksal kodçözcümüz ise verilen bir z koduna karşılık gelebilecek x verilerinin olası değerleri üzerine bir dağılım üretir. Burada daha basit ifade edilebilen $q_\phi(z|x)$ 'i kullanarak $p_\theta(z|x)$ 'i çıkarmak istiyoruz. KL-iraksaması Eşitlik 2.3 'te verilmektedir.

$$\begin{aligned} D_{KL}(q_\phi(z|x) \| p_\theta(z | x)) &= \sum_z q_\phi(z) \log \frac{q_\phi(z)}{p_\theta(z | x)} \\ &= E \left[\log \frac{q_\phi(z)}{p_\theta(z)} \right] \\ &= E [\log q_\phi(z) - \log p_\theta(z)] \end{aligned} \quad (2.3)$$

Bayes kuralı ile $p_\theta(x)$, $p_\theta(z|x)$, ve $p_\theta(z)$ 'yi denklemde görünür hale getirebiliriz.

$$\begin{aligned}
D_{KL}(q_\phi(x) \| p_\theta(z | x)) &= E \left[\log q_\phi(x) - \log p_\theta(z) \frac{p_\theta(z)}{p_\theta(x)} \right] \\
&= E [\log q_\phi(x) - p_\theta(z) + \log p_\theta(z) - \log p_\theta(x)] \\
&= E [\log q_\phi(x) - \log p_\theta(z) - \log p_\theta(z) + \log p_\theta(x)]
\end{aligned} \tag{2.4}$$

Burada beklenen fonksiyonunun z üzerinde olduğunu ve $p_\theta(x)$ 'in z 'ye bağlı olmadığını düşünürsek $p_\theta(x)$ 'i beklenenin dışına alabiliriz.

$$\begin{aligned}
D_{KL}(q_\phi(x) \| p_\theta(z | x)) &= E [\log q_\phi(x) - \log p_\theta(z) - \log p_\theta(z)] + \log p_\theta(x) \\
D_{KL}(q_\phi(x) \| p_\theta(z | x)) - \log p_\theta(x) &= E [\log q_\phi(x) - \log p_\theta(z) - \log p_\theta(z)]
\end{aligned} \tag{2.5}$$

Eşitliğin sağ tarafına dikkatlice bakarsak buranın yeni bir KL-ıraksaması olarak yazılabileceğini görebiliriz. İşaretleri düzenleyerek başlayalım:

$$\begin{aligned}
D_{KL}(q_\phi(x) \| p_\theta(x)) - \log p_\theta(x) &= E [\log q_\phi(x) - \log p_\theta(z) - \log p_\theta(z)] \\
\log p_\theta(x) - D_{KL}(q_\phi(x) \| p_\theta(z | x)) &= E [\log p_\theta(z) - (\log q_\phi(x) - \log p_\theta(z))] \\
&= E [p_\theta(z)] - E [\log q_\phi(x) - \log p_\theta(z)] \\
&= E [p_\theta(z)] - D_{KL}(q_\phi(x) \| p_\theta(z))
\end{aligned} \tag{2.6}$$

Böylece varyasyonel çıkarım formülasyonunu elde etmiş oluyoruz.

$$\log p_\theta(x) = E [p_\theta(z)] - D_{KL}(q_\phi(x) \| p_\theta(z)) + D_{KL}(q_\phi(x) \| p_\theta(x)) \tag{2.7}$$

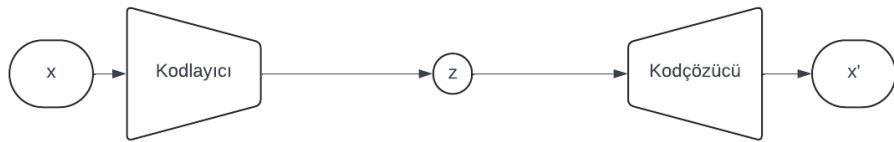
Burada elimizdeki terimlere bir bakalım. Eşitliğin sağ tarafındaki en son terim gerçek sonsal dağılım ile kodlayıcı tarafından üretilen tahmini dağılımın farkını ifade eden KL-ıraksaması terimidir. Yukarıda bahsettiğimiz gibi $p_\theta(x)$ 'in zorlu bir integral içermesi probleminden dolayı bu terimi hesaplamak mümkün değildir ancak KL-ıraksamasının her zaman 0'dan büyük olduğunu biliyoruz. KL-ıraksaması negatif olmadığından kalan ifade x verisinin marjinal olasılığı için varyasyonel alt sınır $L(\theta, \phi; x)$ olarak adlandırılır ve aşağıdaki şekilde yazılır:

$$\begin{aligned}
\log p_\theta(x) &\geq L(\theta, \phi; x) \\
L(\theta, \phi; x) &= -D_{KL}(q_\phi(x) \| p_\theta(z)) + E_{q_\phi(x)} [\log p_\theta(x | z)]
\end{aligned} \tag{2.8}$$

Varyasyonel alt sınır eşitliğinde sağ taraftaki ilk terim kodlayıcı ve z 'nin dağılımları arasındaki KL-ıraksamasını ifade eder ve çözümü mümkünür. İkinci terim ise kodçözücü ağda örneklemeye yapılarak hesaplanabilir. Yani bu terim türevi alınıp optimize edilebilecek bir alt sınır vermektedir. Varyasyonel otokodlayıcıların bu temel optimizasyon yöntemine Kanıt alt sınır optimizasyonu (Evidence lower bound optimization - ELBO) denir.

$$\theta^*, \phi^* = \arg \max_{\theta, \phi} \sum_{i=1}^N L(\theta, \phi; x^{(i)}) \quad (2.9)$$

Bu noktada, verimiz x 'i gizli değişkenler uzayına taşıyan $q_\phi(z|x)$ 'ı kodlayıcı ağ, gizli değişkenlerimiz z 'yi kodlanmış gösterim, verilen gizli değişkenlerden veri üreten $p_\theta(x|z)$ 'yi ise kodçözücü ağ olarak modelleyen Varyasyonel otokodlayıcı(Variational autoencoder - VAE) Şekil 2.1'de gösterilmektedir.



Şekil 2.1 Varyasyonel Otokodlayıcı

2.2.1.2 Varyasyonel Alt Sınırın Otokodlayıcılarla Çözümü

Otokodlayıcılar verinin daha düşük boyutlu temsili özelliklerini öğrenmek için ortaya atılan gözetmensiz(unsupervised) bir yapay öğrenme yaklaşımıdır.

Otokodlayıcılarda kodlayıcı ve kod çözücü ağlar bulunur. Varyasyonel alt sınırı sınır ağılarıyla çözebilme için rastgele örnekler üzerinde geri yayılım yapmak gereklidir. Kodlayıcı ağ $z \sim q_\phi(x^{(i)})$ çıkışında parametrelerini belirlediğimiz bir Gaussian'dan z 'yi örnekleyebiliriz ancak direkt örneklemeye yapılrsa, otokodlayıcı gradyan düşümü ile eğitildiğinden örneklemeye işlevinin gradyanı olmayacağı.

Bunun üstesinden gelmek için yeniden parametrelendirme hilesine başvurulur. Bu hile ile ağı farklılıkabilir(differentiable) hale getireceğiz. Yeniden parametrelendirme hilesi farklılıkamayan(non-differentiable) işlemi ağına çıkartır, böylece hala farklılıkamayan bir terim içermemize rağmen, bu terim ağından outside olduğundan, ağı eğitilebilir.

Yeniden parametrelendirme hilesi şu şekilde işler: elimizde $x \sim N(\mu, \Sigma)$ olduğunu varsayıyalım ve bunu standartlaştıralım, sonuçta $\mu = 0$, $\Sigma = 1$ olur.

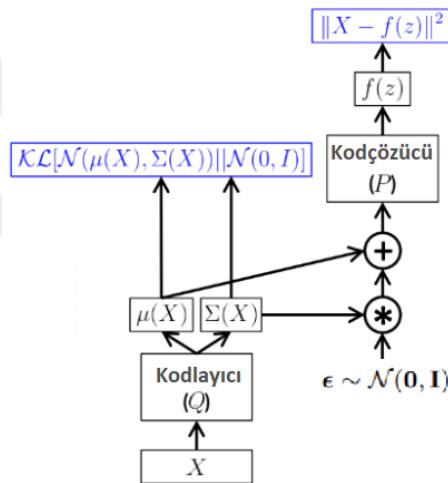
Standardizasyon sürecini geri alarak orijinal dağılıma geri dönebiliriz.

$$x = \Sigma^{\frac{1}{2}} x_{std} \quad (2.10)$$

Bunu genişletirsek standart normal dağılımdan aldığımız bir örneği, ortalamasını ve varyansını bildiğimiz bir Gaussa dönüştürebiliriz. Bu nedenle $\epsilon \sim N(0, 1)$ olduğu durumda z 'nin örneklenmesi şu şekildedir:

$$z = \mu(X) + \Sigma^{\frac{1}{2}}(X)\epsilon \quad (2.11)$$

Böylelikle örnekleme süreci ağıın dışında olduğu için kodlayıcı ağıın çıkışındaki μ ve Σ eğitilebilir. Varyasyonel alt sınırın yeniden parametrelendirme hilesini kullanarak otokodlayıcılarla çözümü Şekil 2.2'de gösterilmiştir.



Şekil 2.2 Yeniden parametrelendirme hilesi

2.2.1.3 KL Hesaplaması

$D_{KL}(q_\phi(x) \| p_\theta(z))$ terimindeki $p_\theta(z)$ gizli değişken dağılımıdır. $p_\theta(z)$ 'den örnekleme yapmak için öncelikle standart normal dağılım $N(0, I)$ 'dan örnekleme yaptığımızdan bahsettik. Örneklemeye işlemini kolaylaştırmak için $q_\phi(x)$ in de $N(0, I)$ e olabildiğince benzemesini istiyoruz.

$p_\theta(z) = N(0, I)$ değerine sahip olması ve $q_\phi(x)$ 'in $\mu(x)$ ve $\Sigma(x)$ parametreleriyle Gaussian olması durumunda bu iki dağılım arasındaki KL ıraksaması kapalı formda hesaplanabilir. k Gaussumuzun boyutu ve $\text{tr}(x)$ iz(trace) fonksiyonu, yani x matrisinin köşegenindeki elemanların toplamı olmak üzere KL-ıraksaması

aşağıdaki gibi yazılabilir.

$$D_{KL}(N(\mu(x), \Sigma(x)) \| N(0, I)) = \frac{1}{2} (\text{tr}(\Sigma(x)) + \mu(x)^T \mu(x) - k - \log \det(\Sigma(x))) \quad (2.12)$$

Köşegen bir matrisin determinantı, köşegenindeki elemanların çarpımı olduğundan $\Sigma(x)$ i bir vektör olarak da uygulayabiliriz:

$$\begin{aligned} D_{KL}(N(\mu(x), \Sigma(x)) \| N(0, I)) &= \frac{1}{2} \left(\sum_k \Sigma(x) + \sum_k \mu^2(x) - \sum_k 1 - \log \prod_k \Sigma(x) \right) \\ &= \frac{1}{2} \left(\sum_k \Sigma(x) + \sum_k \mu^2(x) - \sum_k 1 - \sum_k \log \Sigma(x) \right) \\ &= \frac{1}{2} \sum_k (\Sigma(x) + \mu^2(x) - 1 - \log \Sigma(x)) \end{aligned} \quad (2.13)$$

Uygulamada $\Sigma(X)$ i modellemek $\log \Sigma(X)$ i modellemekten daha iyidir. Çünkü \log hesaplamak yerine üs almak sayısal olarak daha kararlıdır. Sonuçta KL-ıraksaması terimimizin kapalı formda hesaplanması aşağıdaki gibidir:

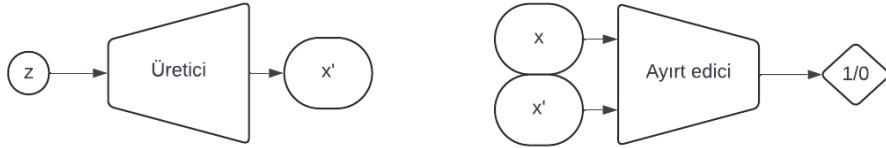
$$D_{KL}(N(\mu(x), \Sigma(x)) \| N(0, 1)) = \frac{1}{2} \sum_k (\exp(\Sigma(x)) + \mu^2(x) - 1 - \Sigma(x)) \quad (2.14)$$

2.2.2 Üretken Çekişmeli Ağ

Verinin olasılık dağılımı $p_\theta(x)$ çoğu zaman çok karmaşık bir dağılımdır ve bu dağılımdan çıkarımda bulunmak çok zor olabilir. Bu nedenle, karmaşık olasılık dağılımını modellemek zorunda kalmadan $p_\theta(x)$ 'den örnek üretebilen bir üretken ağ düşünücsü çok mantıklıdır. Bu amaçla eğitilen bir üretken ağ ile $p_\theta(x)$ dağılımına benzeyen örnekler daha pratik bir şekilde üretilebilir.

Bu bağlamda Goodfellow ve ark. [2] iki rakip sinir ağının modelini beraber eğiterek gerçek dağılımdaki örneklerle benzeyen yapay örnekler üretebilmisti. Bu ağlardan biri girdi olarak "gürültü" alır ve örnekler üretir. Buna üretici(generator) ağ denir. Diğer model ise hem üreticiden hem de eğitim verilerinden örnekler alır ve iki kaynağı birbirinden ayırt etmeye çalışır. Bu da ayırt edici(discriminator)

olarak adlandırılır. Bu iki ağ bir oyun oynamaya başlar ve bu süreçte ayırt edici, üretilen verileri gerçek verilerden ayırmada daha iyi olmayı öğrenirken, üretici de daha gerçekçi örnekler üretmeyi öğrenir. İki ağın aynı anda eğitilmesi ile yapay örneklerin gerçek verilerden ayırt edilemez hale gelmesi beklenir. Üretken Çekişmeli Ağ (Generative Adversarial Network-GAN) Şekil 2.3'te gösterilmiştir.



Şekil 2.3 Üretken Çekişmeli Ağ

Bir benzetme ile bu oyunu aklımızda kalıcı hale getirelim. Sahte para basan bir suçlu ile sahte ve gerçek parayı ayırt etmeye çalışan bir polisin arasındaki ilişkiyi düşünelim. Sahte para basan suçlunun gerçeğine mümkün olduğunda benzeyen paralar üretmesi gerekir ki polisler paranın sahte mi gerçek mi olduğunu anlasın. Öte yandan iyi bir polisin de sahte parayı mümkün olduğunda iyi tespit etmesi gerekir. Polis sahte parayı ayırt edebildikçe suçlu ayırt edici noktaları öğrenerek gerçeğine daha yakın paralar üretir. Burada üretici, ürettiği verinin iyi olup olmadığını bildiren bir ödül sinyali aldığından pekiştirmeli öğrenme(reinforcement learning) akla gelir. Ancak burada temel fark, gradyan bilgisinin ayırt edici ağdan üretici ağa geri yayılımla aktarılmasıdır. Böylece üretici, ayırt ediciyi kandırabilecek örnekleri üretmek için parametrelerini nasıl ayarlayacağını bilir. Bu tür ilişkilere minimax oyunu adı verilir ve bu süreç çekişmeli(adversarial) süreç olarak adlandırılır.

Üretken Çekişmeli Ağ, çekişmeli süreçlerin özel bir durumudur. Bu modellerde birinci ağ veri üretirken ikinci ağ birinci ağın ürettiği veriyi gerçek veriden ayırt etmeye çalışır. İkinci ağ verinin gerçek olma olasılığını $[0,1]$ aralığında sayısal bir çıktı olarak verir.

Üretici ağı G , ayırt edici ağı ise D olarak adlandırırsak GAN'ların hedef fonksiyonunu Eşitlik 2.15 'teki gibi yazabiliriz. Burada $D_{\theta_d}(x)$ ayırt edicinin gerçek veriler için tahmin ettiği olasılığı, $D_{\theta_d}(G_{\theta_g}(z))$ ise yapay veriler için tahmin ettiği olasılığı ifade etmektedir.

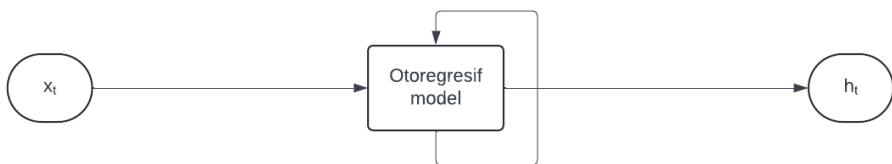
$$\mathcal{L}_{GAN} = \min_{\theta_g} \max_{\theta_d} \left[E_{x \sim p_{\text{data}}} \log D_{\theta_d}(x) + E_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right] \quad (2.15)$$

Denge noktası minimax oyununun en iyi noktasıdır. Bu noktada üretici ağ veriyi öyle bir modeller ki ayırt edici ağ her veri için 0.5 olasılık üretir, yani üretici ağın ürettiği yapay verileri gerçek veri zanneder.

Bu çekişmeli eğitim bazı dezavantajlara neden olur. Öncelikle, GAN'lar hiperparametre seçimlerine karşı aşırı hassastır. Üretici ve ayırt edici arasındaki dengesizlik, model parametrelerinin tutarsız, yakınsak olmayan(non-convergent) veya aşırı uyumlu(overfit) hale gelmesine neden olabilir. Çok yaygın bir diğer durum ise üreticinin sınırlı çeşitliliğe sahip örnekler ürettiği mod çökmesidir(mode collapse).

2.2.3 Otoregresif Model

Otoregresif model [3], ardışık olarak çalışan ve önceki verilerle bir sonraki veriyi tahmin eden işlenebilir(tractable) yoğunluk modelidir. Yalnızca bir sonraki durumun polinomsal zaman olasılığını hesaplayarak başarılı tahminlemeler yaparlar. Otoregresif modeller sınıfına giren Tekrarlayan sinir ağları (Recurrent Neural Network-RNN) [4], ve alt modelleri olan Uzun-kısa süreli bellek (Long-short term memory-LSTM) [5] ve Kapılı tekrarlayan birim (Gated Recurrent Unit-GRU) [6] gibi modeller dil modelleme ve zaman serisi tahmini gibi alanlarda oldukça sık kullanılmaktadır. Üretken öneğitimli dönüştürücü (Generative Pretrained Transformer-GPT) [7] gibi bazı modeller yapısal olarak dönüştürücü(transformer) olup veriyi birim(token) bazında adım adım tahmin ettiği için çalışma prensibi bakımından otoregresiftir. Otoregresif modellerin görüntü üretimi için de bazı uygulamaları mevcuttur [8, 9]. Bu modellerin dezavantajı bir sonraki durumun olasılığı karmaşık olduğunda hesaplamada zorlukların ortaya çıkmasıdır [10]. Otoregresif model Şekil 2.4'te gösterilmiştir.

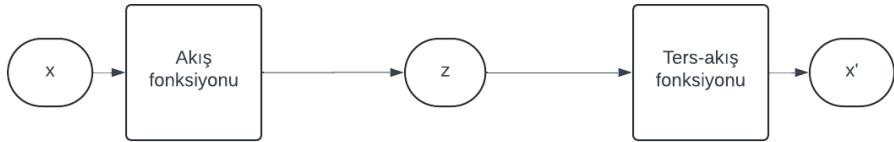


Şekil 2.4 Otoregresif model

2.2.4 Akış Tabanlı Model

Gizli değişkenli üretken modeller, hesaplama verimliliği için sıkılıkla Gauss dağılımını kullanır. Ancak gerçek dünyada, çoğu dağılım Gauss dağılımından çok daha karmaşıktır.

Akış tabanlı (Flow-based) model [11] ve Normalize eden akışlar (Normalizing flows) [12, 13], her adımda bir dizi bijektif dönüşüm(transformation) fonksiyonu uygulayarak basit bir dağılımı karmaşık bir dağılıma dönüştürür. Normalize etmek, dönüşüm uygulandıktan sonra normalize edilmiş bir yoğunluğun(density) elde edilmesi anlamına gelir. Giriş verilerinin tam log-olasılıkları hesaplanabilir ve optimizasyon doğrudan negatif log-olasılık üzerinden gerçekleştirilebilir. Akış tabanlı modeller Şekil 2.5'te gösterilmiştir.



Şekil 2.5 Akış tabanlı model

Normalize eden akış modelleriyle ilgili zorluklardan biri veri hacmidir. Gizli uzay, dönüşümler sırasında çok yüksek boyutlu hale gelir ve bu da yorumlamayı zorlaştırır. Ayrıca, bu modellerle koşullu üretim görevlerini uygulamak çok zordur. Üretilen örneklerin kalitesi de GAN ve VAE'lere kıyasla daha düşüktür.

2.2.5 Enerji Tabanlı Model

Bir diğer üretken model ailesi enerji tabanlı modellerdir (Energy based models-EBM) [14, 15]. Enerji tabanlı model, enerji fonksiyonu olarak adlandırılan normalleştirilmemiş negatif log-olasılıkları bulur. Herhangi bir doğrusal olmayan regresyon fonksiyonu enerji fonksiyonu olarak seçilebilir. Olasılık dağılımı enerji fonksiyonunun hacmine bölünerek normalleştirilir. Enerji tabanlı model Şekil 2.6'da gösterilmiştir. Her iterasyonda gradyan tabanlı bir MCMC yöntemi ile örnekleme yapılır ve eğitim örnekleri ile üretilenler arasındaki farka dayanarak parametreler güncellenir. Bu şekilde model bir fonksiyonu öğrenir ve düşük enerjileri doğru değerlerle, yüksek enerjileri ise yanlış değerlerle ilişkilendirir. Eğitimin sonunda elde edilen enerji modelinden Metropolis–Hastings algoritması ile yeni örnekler üretilebilir.

İlk enerji tabanlı üretken sinir ağı, 2016 yılında önerilen ve evrişimli bir sinir ağı kullanan üretken ConvNet'tir [16]. EBM'lerin esnekliği modellemede önemli faydalar sağlar ancak karmaşık modellerden alınan örneklerin kesin olasılığını hesaplamak ve üretmek genellikle zordur. Ayrıca MCMC tabanlı örnekleme yapıldığından bu yöntemle büyük veri setlerinde çalışmak zordur.

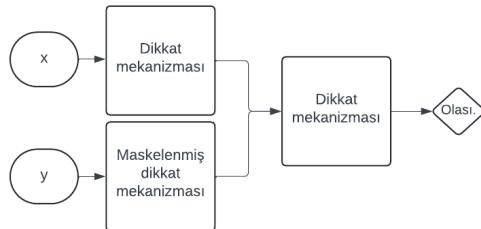


Şekil 2.6 Enerji tabanlı model

2.2.6 Dönüştürücü Modeli

2017 yılında önerilen çok başlı dikkat(multi-head attention) mekanizmasına [17] dayanan derin öğrenme mimarisidir. Dönüştürüçüler, tekrarlayan(recurrent) birimlerin yerine giriş ve çıkış arasındaki küresel bağımlılıklara önem veren bir dikkat mekanizmasına güvenir. Bu sayede önemli ölçüde daha fazla paralellik sağlayarak otoregresif modellere göre daha az eğitim süresi gerektirir.

Dönüştürücü modeli Şekil 2.7'de verilmiştir. Dönüştürüçüler metni, birim(token) adı verilen sayısal gösterimlere dönüştürür ve her birim, bir kelime yerleştirme(word embedding) tablosunda aranarak bir vektöre dönüstürülür. Daha sonra her katmanda paralel çok başlı dikkat mekanizması aracılığıyla her birim diğer birimlerle bağlam(context) penceresinin kapsamı içinde bağlamsallaştırılır ve bu sayede önemli birimler için sinyalin yükseltilmesi, daha az önemli birimler içinse azaltılmasına olanak sağlanır.



Şekil 2.7 Dönüştürücü modeli

2.2.7 Difüzyon Modeli

Üretken difüzyon modelleri [18, 19], VAE'lerde yaklaşık olasılık hesaplamasını, GAN'lardaki çekici-mavili eğitimi, otoregresif modellerdeki ardışık öğrenme gereksinimini, akış tabanlı modellerdeki hacim büyümeyi ve EBM'lerdeki örneklemeye zorluğunu ortadan kaldırın yeni bir üretken model ailesidir. Bu modellerde iki süreç vardır. Birincisi, ileri süreç; Veri dağılımına birden fazla ölçekte gürültü ekler ve kademeli olarak rastgele gürültüye dönüştürür. İkincisi, geri süreç; Difüzyon işlemini adım adım tersine çevirerek orjinal verinin benzerini üretir. Difüzyon modeli Şekil 2.8'de gösterilmiştir.



Şekil 2.8 Difüzyon modeli



3 ÇOK KİPLİ VARYASYONEL OTOKODLAYICILARLA METİNDEN GÖRÜNTÜ ÜRETİMİ

Yapay öğrenme modellerinin aynı konuyu farklı bilgi kaynaklarından öğrenmesi literatürde çok kipli(multimodal) öğrenme olarak bilinmektedir. Bu tür bir öğrenme sürecinde kiplerden bazıları diğerlerinden daha güvenilir olabilir ve daha iyi öğrenilmesi istenebilir. Bu bölümde çok kipli öğrenmede kiplerin gerek sabit katsayılarla gerekse otomatik olarak ağırlıklandırılması için çeşitli yaklaşım önerilmiştir. Önerilen yöntemler yapay ve gerçek veri kümeleri üzerinde klasik yöntemlerle karşılaştırılmıştır.

3.1 Giriş

Doğal öğrenme sürecinde aynı konuyu farklı kaynaklardan öğrenmek bilginin pekiştirilmesini sağlar. İnsan beyni farklı kaynaklardan edindiği bilgileri kendi içerisinde harmanlayıp öğrendiği konuya dair soyut bir anlayış oluşturur. Tıpkı bu doğal süreçteki gibi bir konuyu farklı kaynaklardan (kiplerden) öğrenip kendi içinde soyut bir anlam oluşturmak yapay öğrenme modellerinden varyasyonlu otokodlayıcıların(VAE) [1] yapısına oldukça uygundur ve bu konuda pek çok çalışma mevcuttur [20–23].

Gerçek hayatta bazı bilgi kaynakları diğerlerinden daha önemlidir ve onlara daha çok odaklanmak gerekebilir. Bazı bilgi kaynakları ise daha az bilgi içerir ve daha az odaklanmak yeterli olur. Bu durumda her bilgi kaynağının eşit odaklanmak eğitim verimini düşürebilir. Buradan hareketle, bazı bilgi kaynakları daha güvenilir olarak kabul edilip ortak dağılımı bulurken onların daha çok katkı yapması ile daha başarılı bir model elde etmek hedeflenmiştir. Bu bölümün katkıları aşağıdaki maddelerde özetlenmiştir:

- Çok kipli VAE’leri eğitirken kipleri güvenilirliğine göre ağırlıklandırma yaklaşımı önerilmiştir.
- Çok kipli VAE’lerde ortak sonsal dağılımı bulan uzman yaklaşımının kipleri ağırlıklandırarak da kullanılabileceği gösterilmiştir.
- Güvenilir kipin bilindiği durumlar için sabit katsayılarla ağırlıklandırma önerilmiştir.
- Güvenilir kipin bilinmediği durumlarda otomatik ağırlıklandırma için farklı yaklaşım önerilmiştir.
- Önerilen yöntemler yapay ve gerçek veri kümeleri üzerinde orijinal yöntemlerle karşılaştırılmış ve anlamlı sonuçlar elde edilmiştir.

3.2 İlgili Çalışmalar

Bir kavramı öğrenirken farklı bilgi türleri ile girdi sağlamak insanlarda olduğu gibi üretken yapay öğrenme modellerinde de etkili olabilir. Bilgi türlerine kısaca kip denildiğinde; Örneğin, "kuş" kavramını öğrenirken kuşun görüntüsü bir kip, sesi başka bir kip, onun hakkında açıklayıcı bilgiler içeren bir metin ayrı bir kip olarak düşünülebilir. Bu şekilde farklı bilgi türleri ile çalışabilen yapay öğrenme modellerine çok kipli modeller(multimodal models) denir. Çok kipli modeller kipler arasında ortak bir dağılım bulup öğrendiği kavramı bu gizli ve soyut dağılıma göre yorumlar. Bulunan ortak dağılım yeni ve bilinmeyen örnekler üretmek için kullanılır. Çok kipli Varyasyonel Otokodlayıcılar (Multimodal VAE) bu tür modellerin en popülerlerinden biridir.

Suzuki ve ark. [20] çok kipli varyasyonel otokodlayıcılarla kipleri çift yönlü olarak üretebilmek için bir yöntem önermiştir. Bu yöntemle eksik olarak verilen bazı kiplerin üretilmesi istendiğinde çapraz üretim gerçekleştirilebilir. Bu çalışmadan sonra kiplerin birleştirilmesi için uzman yaklaşımı önerilmiş ve çok kipli varyasyonel otokodlayıcılar ile kiplerin üretimi daha efektif hale getirilmiştir.

Varyasyonel Çıkarım bölümünde Kanıt Alt Sınır optimizasyonu (ELBO)’nun öğrenilen dağılımdan elde edilen örneklerin gerçek örneklerden farkı ile öğrenilen dağılımla gerçek dağılım arasındaki KL-ıraksaması toplanarak bulunduğundan bahsedilmiştir. Eşitlik 3.1’de ise çok kipli ELBO gösterilmektedir. Burada i kipinin örnekleri x_i kodlayıcıya q_ϕ verilerek kodlanır. Ardından kod çözücüye p_θ verilerek yeniden ürettirilir. Gerçek örneklerle üretilen örnekler arasındaki bu fark yeniden yapılandırma hatası (reconstruction loss) E olarak geçer. Kodlanan

dağılımla $q_\phi(z|x_i)$ önsel dağılım $p_\theta(z)$ arasındaki negatif KL-ıraksaması $-D_{KL}$ ise ELBO'nun bir diğer terimidir.

$$\mathcal{L}_i = -D_{KL}(q_\phi(z|x_i)||p_\theta(z)) + E_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] \quad (3.1)$$

Çok kipli VAE'lerde her bir kipin bireysel optimizasyonu için bir ELBO terimi alınır. Buna ek olarak ortak sonsal dağılımin optimizasyonu için bir ELBO terimi daha yer alır. Ortak dağılımin da tipki normal bir dağılım gibi ELBO optimize edilerek eğitilebilmesi için tek bir ortalama ve kovaryansının olması gereklidir. Bu ortak ortalama ve kovaryansı bulabilmek için uzman yaklaşımıları kullanılır. Uzman yaklaşımı tüm kiplerin kendi kodlayıcısında elde ettiği kodları(ortalama, kovaryans) birleştirir ve tek bir ortalama ve kovaryans elde eder. Bu ortalama ve kovaryans her kipin kod çözümüne verilerek kipler yeniden ürettirilir ve her kip için yeniden yapılandırma hataları(reconstruction loss) toplanır. Uzman yaklaşımı ile bulunan ortalama ve kovaryans aynı zamanda KL-ıraksamasını bulmak için de kullanılır.

Çok kipli VAE'lerin optimizasyonu Eşitlik 3.2'de gösterildiği gibi bireysel kiplerin optimizasyonuna (\mathcal{L}_i) ortak sonsal dağılımin optimizasyonu ($\mathcal{L}_{1:M}$) eklenecek yapılabilir.

$$\mathcal{L} = \sum_{i=1}^M \mathcal{L}_i + \mathcal{L}_{1:M} \quad (3.2)$$

Çok kipli VAE'ler tüm kipler aracılığıyla üst düzey kavramları yakalayan ortak bir temsil çıkartmaktadır. Ortak varyasyonel sonsal dağılımin çıkartılması için çeşitli uzman yaklaşımı önerilmiştir [21, 22]. Bu amaçla kullanılan iki temel yöntem uzman çarpımı (Product of Experts - PoE) ve uzman karışımı (Mixture of Experts - MoE) olarak bilinmektedir.

Uzman çarpımı (PoE). Hinton [24] tarafından farklı makine öğrenmesi modellerinin ortak kararını bulmak üzere önerilmiştir. Wu ve ark. [21] ise PoE'yi çok kipli VAE'lerde kiplerin birleştirilmesi için kullanmıştır. kiplerin her birinin kodlayıcısı $q_{\phi_i}(z|x_i) = (\mu_i(x), \Sigma_i(x))$, M tane kipin ortak dağılımı $q_\phi(z|x_{1:M}) \sim (\mu(x), \Sigma(x))$ olarak gösterilirse PoE, Eşitlik 3.3'teki gibi ifade edilir. Wu ve ark.[21] bu ifadeye bir de gizli önsel dağılımı $p(z)$ eklemiştir.

$$q_\phi(z|x_{1:M}) = p(z) \prod_{i=1}^M q_{\phi_i}(z|x_i) \quad (3.3)$$

Gaussian uzmanların çarpımının yine bir Gaussian olduğu ancak çarpımın bileşenlerden daha zengin yapıda olduğu Williams ve ark. [25] tarafından ifade edilmiştir. Uzmanların her birinin ortalaması $\mu_i(x)$ ve kovaryansı $\Sigma_i(x)$, uzman çarpımı için ortalama $\mu(x)$ ve kovaryans $\Sigma(x)$ 'nın PoE ile bulunması Eşitlik 3.4 ve 3.5'te verilmiştir. Burada $T_i(x)$, i uzmanın x verisi için hassasiyetini ifade eder ve Cao ve ark. [26] önerdiği şekilde $T_i(x) = \Sigma_i^{-1}(x)$ olarak alınır.

$$\mu(x) = \left(\sum_{i=1}^M \mu_i(x) T_i(x) \right) \cdot \left(\sum_{i=1}^M T_i(x) \right)^{-1} \quad (3.4)$$

$$\Sigma(x) = \left(\sum_{i=1}^M T_i(x) \right)^{-1} \quad (3.5)$$

Uzman çarpımında Gaussianların çarpımı, bileşeni olan Gaussianları birbirinden ayıramaz. Optimizasyon sırasında bireysel çıkarım ağları eğitilmezse test sırasında eksik veriler sunulduğunda ortak çıkarım ağı ne yapacağını bilemeyecektir. Örneğin 3 kip ile "kuş" kavramını öğretme görevinde, kuşun sesi ve görüntüsü verilip hakkında bir metin istenebilir veya metin ve ses verilip görüntü istenebilir. Bu gibi çapraz üretim görevleri için örnekler hem tamamen hem de kısmen gözlemlenmiş(combos) olarak ele alınır. Yani kuşun görüntüsü, sesi, metni ayrı ayrı ve görüntüsü-sesi-metninin ortak dağılımı optimize edilirken bir de görüntüsü-sesi, sesi-metni, görüntüsü-metni şeklindeki ortak dağılımlar için de birer optimizasyon terimi alınır. Eşitlik 3.6'da her bir kipin ELBO'su, \mathcal{L}_i , ortak ELBO, $\mathcal{L}_{1:M}$ ve kiplerin kombolarının ELBO'ları, $\mathcal{L}_{1:k}$ olarak gösterilmektedir.

$$\mathcal{L} = \sum_{i=1}^M \mathcal{L}_i + \mathcal{L}_{1:M} + \sum_{k=1}^K \mathcal{L}_{1:k} \quad (3.6)$$

Uzman karışımı (MoE). Jacobs ve ark. [27] tarafından bir tür karar birleştirme metodu olarak önerilmiştir. Shi ve ark. [22] MoE'yi çok kipli VAE'lerde kipleri birleştirmek üzere kullanmıştır. Bu yöntem ortak varyasyonel sonsal dağılımı kiplerin tekil sonsal dağılımlarının bir kombinasyonu olarak faktörlere ayırmaktadır. Farklı kiplerin yaklaşık güvenilirlikte olduğu varsayılarak, tüm kiplerin faktörü $1/M$ olarak alınır.

M kipin ortak dağılımı uzmanların bir karışımı olarak Eşitlik 3.7'deki gibi ifade edilir. Burada uzman karışımının ortalaması($\mu(x)$) Eşitlik 3.8 ve kovaryansı($\Sigma(x)$) Eşitlik 3.9'da verilmektedir. Dizileri birleştirmek için birleşim işaretini \cup

kullanılmıştır.

$$q_\phi(z|x_{1:M}) = \frac{1}{M} \sum_{i=1}^M q_{\phi_i}(z|x_i) \quad (3.7)$$

$$\mu(x) = \bigcup_{i=1}^M \mu_i(x) \left[\frac{N \cdot (i-1)}{M} : \frac{N \cdot i}{M} \right] \quad (3.8)$$

$$\Sigma(x) = \bigcup_{i=1}^M \Sigma_i(x) \left[\frac{N \cdot (i-1)}{M} : \frac{N \cdot i}{M} \right] \quad (3.9)$$

MoE'de ortak sonsal dağılımı bulmak için her kipin örnekleri o kipin kodlayıcısına vererek μ_i ve Σ_i dizileri elde edilir. Daha sonra ortak sonsal dağılımin ortalaması $\mu(x)$ 'i bulmak için her kipin ortalama dizisinden μ_i sırayla $\frac{1}{M}$ 'er örnek alınır. Aynı işlem $\Sigma(x)$ 'i bulmak için de yapılır. Elde edilen diziler ortak sonsal dağılımin ortalama ve kovaryans matrisi olarak kayıp fonksiyonunda kullanılır. MoE'de, tüm uzmanlar sonsal dağılıma eşit katılım yaptığı için, aşırı özgüvenli uzmanlardan etkilenmez ve bu sayede ortak dağılım tüm bireysel kiplerdeki bilgiye duyarlı hale gelir.

Hesaplama açısından bakıldığından MoE, PoE'ye göre bazı ek yükler neden olur. Çünkü her kipin kendi kodlama dağılımindan sağladığı örneklerin ortak üretici modelde değerlendirilmesi ilgili kod çözümlerinin üzerinden toplamda M^2 defa geçmeyi gerektirir. Ortak sonsal dağılımı bulmak için Uzman Karışımı (MoE) yaklaşımının eklemeli yapısı, bireysel uzmanların optimizasyonunu kolaylaştırır, ancak KL-ıraksamasını hesaplamak için kapalı form çözümü olmadığından hesaplama açısından etkili bir çözüm değildir.

3.3 Yöntem

Cao ve ark. [26] tarafından önerilen genelleştirilmiş uzman çarpımı (Generalized PoE - gPoE) yaklaşımı ile güvenilir kiplerin ağırlıklandırılması mümkündür. Uzman karışımı yaklaşımında da faktörler kullanılarak kiplere ağırlık verilebilir (gMoE).

Eşitlik 3.10'da gPoE verilmiştir. Burada M kip sayısı olmak üzere her bir $q_{\phi_i}(z|x_i) = (\mu_i(x), \Sigma_i(x))$ kipinin her bir x verisi için güvenilirliğini tutan ağırlık matrisi $a_i(x)$ olarak gösterilmektedir.

$$q_\phi(z|x_{1:M}) = \frac{1}{M} \prod_{i=1}^M q_{\phi_i}^{a_i(x)}(z|x_i) \quad (3.10)$$

Gaussian uzmanların ortak dağılımının $\mu(x)$ ve $\Sigma(x)$ değerleri gPoE yaklaşımıyla Eşitlik 3.11 ve 3.12'deki gibi bulunur. $a_i(x)$ güvenilirlik matrisi olmak üzere $T_i(x)$, i uzmanın x verisi için hassasiyetini ifade eder ve $T_i(x) = \Sigma_i^{-1}(x)$ olarak alınır[26]. Burada $\forall a_i(x) = a_i$ şeklinde tüm kiplere aynı güvenilirlik katsayısı verilmiştir.

$$\mu(x) = \left(\sum_{i=1}^M \mu_i(x) a_i(x) T_i(x) \right) \cdot \left(\sum_{i=1}^M a_i(x) T_i(x) \right)^{-1} \quad (3.11)$$

$$\Sigma(x) = \left(\sum_{i=1}^M a_i(x) T_i(x) \right)^{-1} \quad (3.12)$$

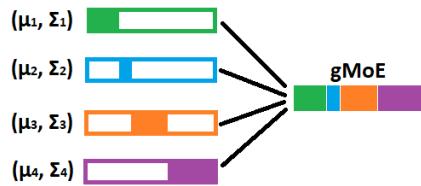
Uzman karışımlarında karışımı oluşturan kiplerden alınacak örnek sayısı o kipin karışımı ne kadar etkileyeceğini belirler. Uzmanların faktörleri a_i ile gösterildiğinde MoE'yi, Eşitlik 3.13'teki gibi ifade edebiliriz. Ortak sonsal dağılımın ortalamasının bulunması için önerilen yöntem Eşitlik 3.14'te verilmektedir. Burada N örnek sayısını göstermektedir ve $a_0 = 0$ olmak üzere a_i değerleri 0-1 aralığında normalize edilmiştir. Kovaryans $\Sigma(x)$ de benzer şekilde kiplerin covaryans dizilerinin birleşiminden oluşur.

$$q_\phi(z|x_{1:M}) = \sum_{i=1}^M a_i q_{\phi_i}(z|x_i) \quad (3.13)$$

$$\mu(x) = \bigcup_{i=1}^M \mu_i(x) \left[N \cdot \left(\sum_{j=1}^i a_{i-1} \right) : N \cdot \left(\sum_{j=1}^i a_i \right) \right] \quad (3.14)$$

Şekil 3.1'de, MoE yaklaşımının ağırlanılarak kullanılmasına bir örnek gösterilmiştir. Burada 2. kipin daha az ağırlıklı, 4. kipin ise daha çok ağırlıklı olduğu bir durumda ortalama ve covaryanslar bulunurken her kip ağırlığı kadar örnekle karışımı katılacak, bir kipin kaldığı yerden diğer kipin örnekleri alınarak devam edilecektir.

Kipleri nasıl ağırlıklandıracağımızı belirledikten sonra ağırlıkların nasıl belirleneceğine sıra gelmektedir. Bunun için sabit katsayı atama ve otomatik olarak bulma yöntemleri önerilmiştir.



Şekil 3.1 Ağırlıklandırılmış uzman karışımımlarında μ ve Σ 'ların bulunması

3.3.1 Sabit Katsayılarla Ağırlıklandırma

Hangi kipin güvenilir olduğunun bilindiği durumlarda kipler sabit katsayılarla ağırlıklandırılabilir. Ancak güvenilir kipe ne kadar ağırlık verileceği de bir problemdir. Deneme yanılma yöntemiyle hangi katsayının daha iyi sonuç verdiği tespit edilebilir.

3.3.2 Otomatik Ağırlıklandırma

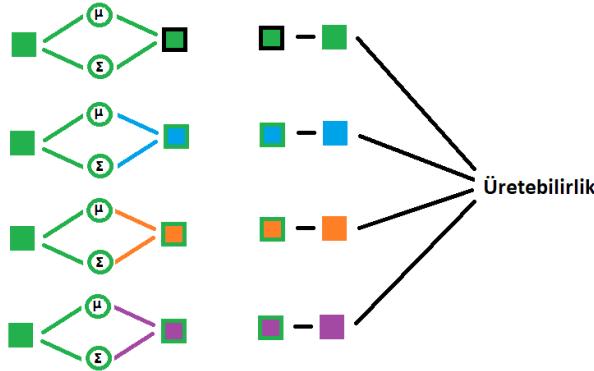
Gerçek hayatı kiplerden hangisini ağırlıklandıracığımızı bilmemiz zordur. Bu çalışmada hangi kipin güvenilir olduğunu eğitim sırasında tespit etmek üzere 2 farklı yöntem önerilmiştir.

3.3.2.1 Üretebilirliğe Göre Ağırlıklandırma

Bir kipin güncel ağırlığını bulmak için o kipin kendisini ve diğer kipleri üretebilme yeteneği göz önünde bulundurulur. Eşitlik 3.15'te i kipinin ağırlığı a_i hesaplanırken üretilen dağılımlar ve gerçek dağılımlar arasındaki JS-ıraksamaları toplanmaktadır. $p_\theta(z_j)$ j kipinin gerçek dağılımı $q_\phi(z_j|x_i)$ ise i kipinin j kipini üretmesi sonucunda elde edilen dağılımdir.

$$a_i+ = \sum_{j=1}^M D_{JS}(q_\phi(z_j|x_i) || p_\theta(z_j)) \quad (3.15)$$

Üretebilirliğe göre ağırlıklandırmayı açıklayan görsel Şekil 3.2 'de verilmektedir. Yeşil ile gösterilen 1. kipin üretebilirliğini bulmak için 1. kipin örnekleri kodlayıcısına verilerek kodlanır ve μ ve Σ kodları elde edilir. Daha sonra bu kodlar 1., 2., 3. ve 4. kiplerin kod çözüçülerine verilir ve bu kipler üretilir. Elde edilen çıktılar üreten kipin rengi(yeşil) ile çerçevelenmiştir(1. kipin kendini üretmesi siyah çerçevede verilmiştir.) Üretilen örneklerle kiplerin gerçek örnekleri arasındaki farkların toplamı 1. kipin üretebilirliğini vermektedir. Bu fark ne kadar düşük olursa o kadar üretken demektir. Farkın yüksek olması ise o kipin diğer kipleri üretmekte zorluk yaşadığını gösterir.



Şekil 3.2 Kiplerin üretemeliliğinin hesaplanması

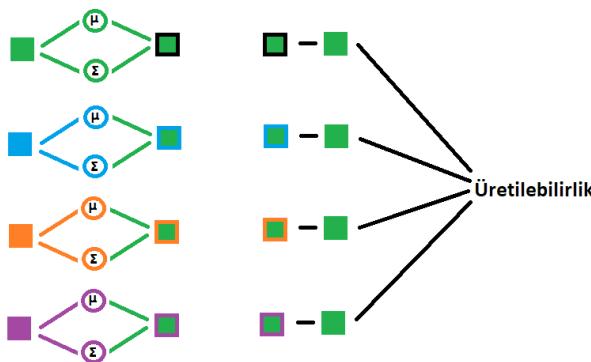
3.3.2.2 Üretilebilirliğe Göre Ağırlıklandırma

Bir kipin güncel ağırlığını bulmak için tek tek tüm kiplerden bu kipi üretmesi istenir. Eşitlik 3.16'da $p_\theta(z_i)$ i kipinin gerçek dağılımı $q_\phi(z_i|x_j)$ ise j kipinden i kipinin üretilmesi sonucunda elde edilen dağılımdir. Sonuçta i kipinin ağırlığı a_i bulunurken üretilen dağılımlar ve gerçek dağılımlar arasındaki JS-ıraksamaları toplanmaktadır.

$$a_i+ = \sum_{j=1}^M D_{JS}(q_\phi(z_i|x_j)||p_\theta(z_i)) \quad (3.16)$$

Üretilebilirliğe göre ağırlıklandırmayı açıklayan görsel Şekil 3.3 'te verilmektedir. 1. kipin diğer kipler tarafından üretilebilmesini bulmak için öncelikle her bir kip kendi kodlayıcısında kodlanır ve μ ve Σ kodları elde edilir. Sonrasında bu kodlar ayrı ayrı 1. kipin kod çözümüne verilerek her bir kipten 1. kip üretilmiş olur. Üretilen örnekler üreten kipin çerçevesinde gösterilmiştir. Üretilen örneklerle ve 1. kipin gerçek örnekleri arasındaki farkların toplamı 1. kipin üretilebilirliğini vermektedir. Bu fark ne kadar düşük olursa diğer kipler o kipi iyi üretiyor demektir. Farkın yüksek olması ise diğer kiplerin o kipi üretmekte zorluk yaşadığını gösterir.

Her kip için ağırlık faktörleri belirlendikten sonra bu faktörler normalize edilerek güvenilirlik katsayılarına eklenmiş ve güvenilirlik katsayıları da kendi aralarında normalize edilmiştir. Verilen eşitliklerdeki gibi eğitim gerçekleştirildiğinde üretilebilirliği/ürtebilirliği düşük olan kipe ağırlık verilmiş olur. Tam ters açıdan bakılırsa $a_i = 1 - a_i$ alınarak daha üretken/ürtebilir kipe de ağırlık verilebilir.



Şekil 3.3 Kiplerin üretilebilirliğinin hesaplanması

3.4 Deneyler

Kiplerin ağırlıklandırılması ile ilgili denemeler yapay (Gaussian) ve gerçek (MNIST/FashionMNIST) veri kümelerinde olmak üzere 2 başlık halinde verilecektir.

3.4.1 Gaussian Kiplerle Denemeler

Çok kipli öğrenmede tüm kiplerin ortak bir gizli dağılıma sahip olduğu varsayıldığından 3 adet 1-boyutlu koşullu bağımsız (conditionally independent), Gaussian kip üretilmiştir. Denemelerde ortak gizli dağılım standart normal dağılım olarak alınmış, μ ve Σ değerlerinin ikisi de 1. kip için 1, 2. kip için 2, 3. kip için 3 olarak alınmıştır. Dolayısıyla ortak gizli dağılıma en yakın kip 1. kiptir. Eğitim kümesi için her bir dağılımdan 10000'er örnek, test kümesi için ise 1000'er örnek alınmıştır.

Çok kipli VAE modelinin gizli katman boyutu 1 olarak alınmıştır. Kodlayıcı $q_{\phi_i}(z|x_i)$ ve kod çözücü $p_{\theta}(x_i|z)$ 5'er nöronluk tam bağlantılı (fully connected) ağlardır. Çok kipli VAE'lerin uzman yaklaşımlarından hem PoE hem de MoE için denemeler yapılmıştır. Kiplerin ağırlıklandırılması için uygulanan yöntemler aşağıda maddelenmiştir:

- Klasik PoE / MoE'de tüm kiplere eşit ağırlık verilmiştir.
- Sabit ağırlıklandırma yöntemlerinde sırayla 1., 2. ve 3. kiplere 2 kat ağırlık verilmiştir.
- Otomatik ağırlıklandırmada üretebilen/üretmemeyen ve üretilebilen/üretilemeyen kiplere ağırlık verilmiştir.

Test aşamasında üç kipin her biri ile tüm kipler tek tek üretilip toplam Jensen-Shannon (JS) ıraksamaları(daha düşük daha iyi) incelenmiştir. Klasik yöntemden daha iyi olan sonuçlar tablolarda kalın yazı ile gösterilmiştir.

3.4.1.1 Uzman Çarpımı Sonuçları

Tablo 3.1 kiplerin üretilme başarıları üzerinden yöntemlerin karşılaştırılmasını göstermektedir. Ağırlıklandırma yöntemleri, üretilme başarısı görece daha iyi olan 3. kipten biraz taviz verilerek diğer kiplerdeki üretilme başarısını iyileştirmeyi sağlamaktadır. Bu nedenle yöntemlerden biri hariç hepsi toplam üretme başarısı açısından klasik PoE'den daha iyidir. Gizli önsel dağılıma en yakın olan kipe ağırlık vermek(2.satır) diğer kiplerin bu kipi üretme başarısını düşürmektedir. En uzak olan kipe ağırlık vermek(4. satır) ise bu kipin üretme başarısını koruyarak diğer kiplerin üretme başarısını artırmaktadır. Otomatik ağırlıklandırma yöntemleri de genel olarak 1. ve 2. kipi üretme konusunda daha başarılıdır ancak 3. kipte zorluk çekmektedir. Bu nedenle PoE'de ağırlıklandırma yapmanın kiplerin üretim başarısını birbirine yaklaştırdığını söylemek mümkündür.

Tablo 3.1 PoE ile tüm kiplerin üretilmesi

PoE	Çıktılar			
	1	2	3	Toplam
$a_1 = a_2 = a_3$	0.4611	0.4644	0.1672	1.0927
$a_1 = 2a_2 = 2a_3$	0.4913	0.3878	0.2232	1.1023
$2a_1 = a_2 = 2a_3$	0.156	0.3909	0.1754	0.7223
$2a_1 = 2a_2 = a_3$	0.1701	0.3559	0.1682	0.6942
Üretebilen	0.1648	0.3592	0.1787	0.7027
Üretmemeyen	0.1511	0.3769	0.1743	0.7023
Üretilebilen	0.4188	0.4047	0.2227	1.0462
Üretilememeyen	0.1537	0.3645	0.1914	0.7096

3.4.1.2 Uzman Karışımı Sonuçları

MoE yaklaşımıyla kiplerin üretilme başarıları üzerinden yöntemlerin karşılaştırılması Tablo 3.2'de gösterilmektedir. MoE ile ağırlıklandırma zaten iyi üretilen kipin(2. kip) daha iyi üretilmesini sağlamıştır. Burada farklar daha minimal olmakla birlikte yine de ağırlıklandırma yöntemlerinden çögünün toplam üretme başarısı açısından klasik MoE'den daha iyi olduğu görülmektedir. MoE ile üretilme başarısı görece daha iyi olan 2. kipe ağırlık vermek(3. satır) performansı düşürmektedir. Gizli önsel dağılıma en uzak kipi ağırlıklandırmak(4. satır) ise diğer kiplerin üretilme başarısını artırmıştır.

Tablo 3.2 MoE ile tüm kiplerin üretilmesi

MoE	Çıktılar			
	1	2	3	Toplam
$a_1 = a_2 = a_3$	0.5223	0.1968	0.3652	1.0843
$a_1 = 2a_2 = 2a_3$	0.5194	0.1525	0.3576	1.0295
$2a_1 = a_2 = 2a_3$	0.6493	0.4270	0.4674	1.5437
$2a_1 = 2a_2 = a_3$	0.5074	0.1666	0.3803	1.0543
Üretebilen	0.4955	0.1679	0.3703	1.0337
Üretemeyen	0.5033	0.1758	0.3642	1.0433
Üretilebilen	0.5256	0.1803	0.3520	1.0579
Üretilemeyen	0.5323	0.1798	0.3878	1.0999

3.4.2 MNIST / FashionMNIST Denemeleri

MNIST / FashionMNIST veri kümeleri için görüntü 1. kip, sınıflar ise 2. kiptir. Sınıflar metne çevrilip kategoriler ikili(binary) olarak temsil edilmiştir(one-hot encoding). Yöntemler arasındaki farkın belirgin olabilmesi için eğitim kümesi 1000 örnekle sınırlandırılmıştır. Denemelerde yer alan yöntemler aşağıda maddelenmiştir:

- Klasik PoE / MoE
- Sabit ağırlıklandırmada görüntü kipine ve metin kipine ayrı ayrı 2 kat ve 3 kat ağırlık vermek
- Otomatik ağırlıklandırmada üretebilene / üretmemeyene ve üretilebilene / üretilemeyene ağırlık vermek

Yöntemleri karşılaştırmak için ikili çapraz entropi(binary cross-entropy) hatası ölçülmüştür. Ölçülen farklar aşağıda açıklanmıştır. Tüm hatalarda 10 farklı rastgele eğitim kümesi ile elde edilen değerler ortalama \pm standart sapma şeklinde verilmiştir. Klasik yöntemden daha iyi olan sonuçlar kalın yazı ile belirtilmiştir. T-testte p-değeri 0.05'in altında olan(istatistiksel açıdan anlamlı) karşılaştırmaların sonuçları italik olarak gösterilmiştir.

- **I2I farkı:** Görüntülerin yeniden üretilmesi sonucunda elde edilen yeni görüntülerle gerçek görüntüler arasındaki farktır.
- **I2T farkı:** Görüntülerin kodlandıktan sonra metin kod çözücüşüne verilmesi sonucunda üretilen metinlerle gerçek metinler arasındaki farktır.
- **T2I farkı:** Metinlerin kodlanması görüntü kod çözücüşüne verilmesi sonucunda elde edilen görüntülerle gerçek görüntüler arasındaki farktır.

- **T2T farkı:** Metinlerin yeniden üretilmesi sonucunda elde edilen yeni metinlerle gerçek metinler arasındaki farktır.
- **J2I farkı:** Ortak dağılımin kodlarının görüntü kod çözücüüsüne verilmesi ile üretilen görüntülerle gerçek görüntüler arasındaki farktır.
- **J2T farkı:** Ortak dağılımin kodlarının metin kod çözücüüsüne verilmesi ile üretilen yeni metinler ile gerçek metinler arasındaki farktır.

3.4.2.1 Uzman Çarpımı Sonuçları

Uzman çarpımı yaklaşımında ağırlıklandırma yöntemlerinin MNIST veri kümesindeki sonuçları Tablo 3.3'te ve FashionMNIST veri kümesindeki sonuçları ise Tablo 3.4'te verilmektedir. Sonuçlara genel olarak bakıldığında metin kipine güvenilirlik verildiğinde çapraz üretim, görüntü kipine güvenilirlik verildiğinde aynı kipi yeniden üretme başarılarının yükseldiği görülmektedir. Otomatik ağırlıklandırma yöntemleri incelendiğinde üretebilene ve üretilemeye ağırlık verme yöntemlerinin sonuçları görüntü kipine ağırlık veren yöntemlere benzemektedir. Üretmemeyen ve üretilebilene ağırlık verme yöntemlerinin sonuçları ise metin kipine ağırlık verilen yöntemlere benzemektedir.

Tablo 3.3 PoE ile MNIST veri kümelerinin sonuçları

Yöntemler	Farklar					
	J2I	J2T	I2I	I2T	T2I	T2T
$a_i = a_t$	0.2042±0.0022	0.8016±0.0444	0.1957±0.0028	0.9608±0.0391	0.2232±0.0006	0.7139±0.0535
$2a_i = a_t$	0.2158±0.0012	0.7793±0.0531	0.2031±0.0022	0.9138±0.0385	0.2222±0.0004	0.764±0.057
$3a_i = a_t$	0.2192±0.0006	0.7781±0.052	0.2055±0.0008	0.911±0.0445	0.2223±0.0005	0.7726±0.0534
$a_i = 2a_t$	0.1919±0.0014	0.8725±0.0529	0.1879±0.0017	1.0109±0.0465	0.2244±0.0009	0.656±0.0683
$a_i = 3a_t$	0.1889±0.0024	0.9174±0.0457	0.1869±0.0025	1.0309±0.0398	0.2251±0.0007	0.6139±0.0706
Üretebilen	0.2011±0.0028	0.8057±0.049	0.1933±0.0033	0.9619±0.0452	0.2234±0.0007	0.7065±0.0575
Üretmemeyen	0.2049±0.0022	0.7932±0.05	0.1957±0.003	0.9494±0.0405	0.2231±0.0007	0.7224±0.0588
Üretilen	0.2097±0.0021	0.7785±0.0579	0.197±0.0023	0.9294±0.0549	0.2227±0.0006	0.7422±0.0564
Üretilmemeyen	0.1988±0.0023	0.8428±0.0463	0.1937±0.0028	0.9878±0.0499	0.2238±0.001	0.6826±0.0598

Tablo 3.4 PoE ile FashionMNIST veri kümelerinin sonuçları

Yöntemler	Farklar					
	<i>J2I</i>	<i>J2T</i>	<i>I2I</i>	<i>I2T</i>	<i>T2I</i>	<i>T2T</i>
$a_i = a_t$	0.3483±0.0016	0.8287±0.0451	0.3388±0.0018	1.1539±0.037	0.3987±0.0011	0.6994±0.0504
$2a_i = a_t$	0.3622±0.0027	0.8044±0.0464	0.3436±0.0023	1.1523±0.0687	0.398±0.0008	0.7599±0.0436
$3a_i = a_t$	0.3705±0.003	0.7881±0.0454	0.3456±0.0016	1.1132±0.0696	0.3976±0.0005	0.7719±0.0444
$a_i = 2a_t$	0.3419±0.0014	0.9186±0.0511	0.3373±0.0017	1.2182±0.068	0.4006±0.0011	0.6289±0.0429
$a_i = 3a_t$	0.3391±0.0015	0.9889±0.0391	0.3367±0.0021	1.2585±0.0398	0.402±0.0015	0.5906±0.0509
Üretribilen	0.3475±0.0022	0.8404±0.0468	0.3385±0.0017	1.1802±0.0426	0.399±0.001	0.6935±0.0455
Üretmemeyen	0.3511±0.0025	0.8156±0.0457	0.3391±0.0025	1.1549±0.0463	0.3986±0.0007	0.7074±0.0497
Üretilabilen	0.365±0.0027	0.7912±0.0471	0.3427±0.003	1.1118±0.046	0.3972±0.0006	0.7645±0.0469
Üretilmemeyen	0.3418±0.0023	0.9591±0.0374	0.3388±0.003	1.2407±0.0379	0.4011±0.0009	0.6227±0.0506

3.4.2.2 Uzman Karışımı Sonuçları

Uzman karışımı yaklaşımında ağırlıklandırma yöntemlerinin sonuçları Tablo 3.5 ve 3.6'da verilmektedir. Uzman çarpımı yaklaşımına benzer şekilde metin kipine ağırlık verildiğinde çapraz üretim başarısının görüntü kipine ağırlık verildiğinde ise aynı kipi yeniden üretme başarısının arttığı görülmektedir. Üretmemeyen ve üretilebilene ağırlık verme yöntemleri metin kipine ağırlık veren yöntemlere benzerken üretebilene ve üretilemeyene ağırlık verme yöntemleri ise görüntü kipine ağırlık verilen yöntemlere büyük ölçüde benzemektedir.

3.5 Sonuç

Bu çalışmada çok kipli varyasyonel otokodlayıcıların optimizasyonunda güvenilir kiplerin gerek sabit katsayılarla gerekse otomatik yöntemlerle ağırlıklandırılması incelenmiştir. Yapay ve gerçek veri kümelerinde uzman çarpımı ve uzman karışımı yöntemlerinin ağırlıklı versiyonları kullanılarak denemeler gerçekleştirılmıştır.

Yapay veri kümelerinde ağırlıklandırmmanın etkisi ilgili uzman yaklaşımına göre değişmektedir. Uzman çarpımı yaklaşımı için gizli önsel dağılıma uzak olan kipin ağırlıklandırmasının sonuca olumlu katkı sağladığı görülmüştür. Uzman karışımlarında ortak sonsal dağılım tüm bireysel kiplerdeki bilgiye duyarlı olduğundan bir kip ne kadar ağırlıklandırsa da diğer kiplerin üretilme başarısı bu durumdan çok fazla etkilenmemektedir.

Gerçek veri kümelerinde her iki uzman yaklaşımının sonuçlarında bir korelasyon görülmektedir. Bu denemeler görüntü kipinin üretebilen ve üretmemeyen bir kip olduğunu metin kipinin ise üretebilen ve üretmemeyen bir kip olduğunu ortaya çıkarmıştır. Buradan görüntü kipinin daha baskın(dominant), metin kipinin ise daha çekinik(resesif) olduğunu çıkarsayarak; "*Baskın kipi ağırlıklandırmayan kiplerin kendini yeniden üretme başarısını artırduğunu, çekinik kipi ağırlıklandırmayan ise çapraz üretim başarısını artırduğunu*" söyleyebiliriz. Bu bağlamda, verilen bir görüntüyü yeniden oluşturma veya verilen bir metnin benzerini üretme gibi görevlerde görüntü kipine ağırlık vermek, görüntü verip açıklamasını isteme(image-to-text), metin verip görüntü oluşturma(text-to-image) gibi görevlerde ise metin kipine ağırlık vermek elverişli olacaktır.

Tablo 3.5 MoE ile MNIST veri kümelerinin sonuçları

Yöntemler	Farklar					
	J2I	J2T	I2I	I2T	T2I	T2T
$a_i = a_t$	0.204±0.0016	0.7965±0.0525	0.1931±0.0035	0.9655±0.0516	0.2232±0.0006	0.7055±0.0532
$2a_i = a_t$	0.2125±0.0041	0.7612±0.0542	0.195±0.0086	0.9409±0.0603	0.2229±0.0007	0.7237±0.0583
$3a_i = a_t$	0.2157±0.0036	0.7551±0.0523	0.1956±0.0099	0.9354±0.0637	0.2229±0.0008	0.7315±0.0575
$a_i = 2a_t$	0.1981±0.0038	0.8537±0.0444	0.1927±0.0027	0.986±0.0392	0.2235±0.0011	0.6746±0.0731
$a_i = 3_t$	0.1948±0.0066	0.8982±0.0491	0.1914±0.0057	1.0025±0.0479	0.2241±0.0012	0.6627±0.092
Üretebilen	0.2038±0.0015	0.8042±0.0517	0.1935±0.003	0.967±0.0466	0.2233±0.0005	0.7035±0.0579
Üretmemeyen	0.2055±0.0018	0.7885±0.0528	0.1937±0.0041	0.9616±0.0487	0.2229±0.0003	0.7066±0.0542
Üretilebilen	0.2095±0.0025	0.7742±0.0575	0.1948±0.0046	0.9525±0.0532	0.2231±0.0007	0.7201±0.0574
Üretilemeyen	0.2003±0.0023	0.832±0.0466	0.1929±0.0019	0.9802±0.0473	0.2232±0.0004	0.688±0.0624

Tablo 3.6 MoE ile FashionMNIST veri kümelerinin sonuçları

Yöntemler	Farklar					
	J2I	J2T	I2I	I2T	T2I	T2T
$a_i = a_t$	0.361±0.0015	0.8415±0.0434	0.3377±0.002	1.1528±0.0498	0.3975±0.0005	0.7012±0.0439
$2a_i = a_t$	0.3736±0.0014	0.7794±0.0495	0.3364±0.0013	1.1509±0.053	0.3982±0.0004	0.7057±0.0522
$3a_i = a_t$	0.3794±0.0013	0.7563±0.0477	0.3351±0.001	1.1403±0.0557	0.3982±0.0006	0.7009±0.0471
$a_i = 2a_t$	0.3519±0.0016	0.9206±0.0409	0.3403±0.0017	1.1463±0.0443	0.3975±0.0005	0.7068±0.0492
$a_i = 3a_t$	0.3494±0.002	0.977±0.0498	0.342±0.0018	1.1607±0.064	0.3974±0.0007	0.7086±0.0452
Üretebilen	0.3602±0.0016	0.8519±0.0528	0.3388±0.0016	1.1571±0.0629	0.3977±0.0005	0.7043±0.0494
Üretmemeyen	0.3635±0.0014	0.8291±0.0503	0.3379±0.0011	1.1518±0.0669	0.3976±0.0005	0.7017±0.0474
Üretilebilen	0.3776±0.001	0.7608±0.0512	0.3359±0.0013	1.1413±0.0562	0.3985±0.0007	0.7018±0.0486
Üretilemeyen	0.3492±0.0012	0.9441±0.0456	0.3401±0.001	1.1412±0.0627	0.3971±0.0005	0.7045±0.0489

4

ÜRETKEN DİFÜZYON MODELLERİ: GÜNCEL TEORİK GELİŞMELERİN BİR İNCELEMESİ

Veri dağılımını gürültüye dönüştüren ve sonrasında ise benzer bir dağılım elde etmek için gürültüden arındıran üretken difüzyon modelleri, güçlü bir teorik altyapıya sahip olup birçok alanda yüksek başarı göstermiştir. Literatürdeki mevcut incelemelerin çoğu, belirli uygulama alanlarına odaklanarak algoritmadaki geliştirmelere yoğunlaşmamıştır. Bu bölümde, üretken difüzyon modellerinin teorik gelişmeleri odaklandıkları konulara göre kategorize edilerek incelenmiştir. Bu kategorizasyon sayesinde gelecekte yeni geliştirmeler yapacak olan araştırmacılar için net bir anlayış elde edilmiştir.

4.1 Giriş

Üretken difüzyon modelleri, skor fonksiyonunu veya veri dağılımının yaklaşık alt sınırını yavaşça gürültüye dönüştüren, ardından benzer bir veri dağılımı elde etmek için ters işlemede gürültüyü geri alan bir üretken modeller ailesidir. Teorik geçmişine ek olarak, pratikte de yüksek başarı göstermiş ve literatürde dikkat çekmiştir.

Difüzyon modelleri literatüründe, 3 temel çalışmaya karşılaşılmaktadır [28–30]. Bu çalışmalar, teorik bir temel ve pratik sonuçlar sağlamaktadır, ancak aynı zamanda geliştirilmeye de ihtiyaçları vardır. Difüzyon modelleri, örnekleme sürecinin hesaplama maliyeti, daha yüksek log-olasılık değerleri ve farklı modalitelerle uyumluluk gibi bazı temel sorunlara sahiptir. Literatürde algoritmayı çeşitli bakış açılarından iyileştiren çok sayıda yaklaşım bulunmaktadır.

Bu bölümde, üretken difüzyon modellerinin teorik gelişmelerine genel bir bakış yapılacaktır. Mevcut incelemelerin çoğu, bilgisayarlı görme [31–33], doğal dil işleme [34], tıbbi görüntüleme [35], zaman serisi analizi [36], metinden görüntü üretimi [37] ve metinden konuşma sentezi [38] gibi belirli uygulama alanlarına odaklanmıştır. Hem teorik gelişmeleri hem de uygulamaları inceleyen diğer bazı

çalışmalar [39, 40] geniş bir bakış açısına sahip oldukları için yalnızca popüler araştırmalara odaklanmışlar ve birçok teorik gelişmeyi incelememişlerdir. Bir diğer inceleme [41], tasarım temelleri üzerine yapılan araştırmalara odaklanmış ve diğer bazı yaklaşımların yalnızca katkılarına değinmiştir.

Burada bahsedilecek çalışmanın en büyük farkı, üretken difüzyon modellerinin teorik gelişmelerini ayrıntılı olarak açıklayan ilk inceleme olmasıdır. Bir diğer fark ise, her gelişmenin tek bir kategoriye girmesiyle net bir anlayış getiren kategorizasyon perspektifimizdir. Odaklanılan gelişmelerin konuya bakış açıları dikkate alınarak, gelecekteki araştırma yönleri gösterilmiştir.

Bu çalışmada teorik gelişmeler, eğitim tabanlı ve eğitimden bağımsız yaklaşımalar olarak kategorize edilmiştir. Teorik gelişmeler, bu kategoriler altında odaklandıkları konulara göre sınıflandırılmıştır. Bu bölümün katkıları aşağıda özetlenmiştir:

- Difüzyon modellerinin teorik gelişmelerine derinlemesine odaklanan bu alandaki ilk incelemedir.
- Difüzyon modellerinin temel çalışmaları sistematik bir perspektifte incelenmiş, aralarındaki ilişkiler ve eksik noktaları açıklanmıştır.
- Teorik gelişmeler konularına göre kategorilere ayrılmıştır. Bu sayede her bir yaklaşımın tek bir kategoriye yerleştirilmesiyle net bir anlayış elde edilmiştir.
- Difüzyon modellerinin değerlendirme ölçütleri açıklanmış ve en bilinen veri kümeleri üzerindeki kıyaslama sonuçları raporlanmıştır.

4.2 Difüzyon Modellerinin Temel Çalışmaları

Literatürde üretken difüzyon modelleri hakkında üç ana çalışma bulunmaktadır. İlk çalışma, denge dışı(non-equilibrium) termodinamik teorisinden esinlenen Gürültü Arındırıran Olasılıksal Difüzyon Modelleri (Denoising Diffusion Probabilistic Models - DDPM)[28]'dır. DDPM'ler olasılık dağılımını tahmin etmek için gizli değişkenler kullanır. İkinci temel çalışma, Gürültü Şartlı Skor Ağları (Noise Conditional Score Networks - NCSN) [29]'dır. NCSN'lerde, paylaşımımlı bir ağ, farklı gürültü seviyelerinde veri dağılımlarının skor fonksiyonunu (log yoğunluğunun eğimi) tahmin etmek için skor-eşleştirme(score-matching) yöntemleriyle eğitilir. Üçüncü çalışma ise difüzyon sürecini ileri ve geri stokastik diferansiyel denklemlerle(SDE) çözen Skor SDE[30]'dır.

Bu bölümde, bu üç ana difüzyon modeli çalışması ayrıntılı olarak açıklanacak,

ayrıca DDPM'leri ve NCSN'leri bir araya getirmek için genel bir çerçeve gösteren çalışmalara değinilecektir.

4.2.1 Gürültü Arındırın Olasılıksal Difüzyon Modelleri

Tanımlar & Notasyonlar. Gürültü arındırın olasılıksal difüzyon modelleri[28] sonlu bir sürenin sonunda gerçek verilere benzeyen örnekler üretmek için varyasyonel çıkarım kullanılarak eğitilen parametrelî bir Markov zinciridir. Olasılıksal difüzyon modellerinde iki aşamalı bir süreç işler: birincisi ileri difüzyon süreci ikincisi ise ters difüzyon veya yeniden yapılandırma sürecidir. İleri difüzyon sürecinde, veriler tamamen gürültü olana kadar art arda Gaussian gürültü verilir. Ters difüzyon süreci ise bir sinir ağları modeli ile koşullu olasılık yoğunluklarını öğrenerek gürültüyü geri alır.

Gerçek bir veri dağılımından alınan bir veri noktası $x_0 \sim q(x)$ verildiğinde, bu örneğe T adım boyunca azar azar Gaussian gürültü ekleyerek bir dizi gürültülü örnek x_1, x_2, \dots, x_T üreten ileri difüzyon süreci Eşitlik 4.1'deki gibi tanımlanır. Burada t zamanındaki dağılımin tahmini, yalnızca hemen önceki $t - 1$ zamanındaki dağılıma ve dolayısıyla koşullu olasılık yoğunluğununa bağlıdır. Adım t büyükçe örnek x_0 , ayırt edilebilir özelliklerini kademeli olarak kaybeder. Sonunda $T \rightarrow \infty$ olduğu zaman, x_T izotropik Gaussian bir dağılıma eşdeğerdir.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (4.1)$$

Tüm sürecin dağılımı ise Eşitlik 4.2'deki gibi hesaplanır.

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) \quad (4.2)$$

Eşitlik 4.1'de $\beta_t \in (0, 1)_{t=1}^T$ ile gösterilen adım büyüklükleri süreç boyunca sabit olarak alınabilir, ardışık adımlarda kademeli olarak değiştirilebilir veya diferansiyel olarak belli bir aralık içinde parametrelendirilebilir. $\alpha_t = 1 - \beta_t$ ve $\bar{\alpha}_t = \prod_{i=1}^T \alpha_i$ olarak gösterilirse sonsal dağılım Eşitlik 4.3'teki gibi düzenlenlenebilir. Burada $z_{t-1}, z_{t-2}, \dots, \sim \mathcal{N}(0, I)$ ve \bar{z}_{t-2} iki Gaussian'ın birleşimidir. Farklı varyanslara sahip iki Gaussian'ın ($\text{ör}; \mathcal{N}(0, \sigma_1^2 I)$ ve $\mathcal{N}(0, \sigma_2^2 I)$) birleşimi $\mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)I)$ için standart sapma $\sqrt{(1 - \alpha_t) + \alpha_t(1 - \alpha_{t-1})} = \sqrt{1 - \alpha_t \alpha_{t-1}}$ olarak bulunur. Genellikle, örnek daha gürültülü hale geldiğinde daha büyük bir güncelleme adımı atılabilir $\beta_1 < \beta_2, \dots, \beta_T$, dolayısıyla $\bar{\alpha}_1 > \bar{\alpha}_2, \dots, \bar{\alpha}_T$ olur.

$$\begin{aligned}
x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}z_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{z}_{t-2} \\
&= \dots \\
&= \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}z
\end{aligned} \tag{4.3}$$

$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{a}_t}x_0, (1-\bar{a}_t)I)$

Ters difüzyon süreci için sistemin mevcut durumu verilir ve daha önceki adımdaki olasılık yoğunluğunun tahmin edilmesi gereklidir. Mevcut durumdan önceki durumun tahminini yapabilmek için önceki tüm gradyanların bilgisi gereklidir ve bu bilgi ancak bir öğrenme modeli ile elde edilebilir. Eşitlik 4.4'te verilen θ parametreli bir sınır ağı modeli $p_\theta(x_{t-1}|x_t)$ 'yi tahmin edebilir. Ters difüzyon süreci standart Gaussian dağılımı $p_\theta(x_T) = \mathcal{N}(x_T; 0, I)$ ile başlar.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4.4}$$

Ters difüzyonun tüm süreci ise Eşitlik 4.5'te verilmektedir.

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) \tag{4.5}$$

Eğitim hedefi. Gürültü arındırılan olasılıksal difüzyon modelleri, VAE'lere benzer şekilde eğitilir. Girdi katmanının boyutu, veri boyutlarıyla aynıdır. Orta katmanlar, ilgili aktivasyon fonksiyonlarına sahip doğrusal katmanlardır. Son katman, yine girdi katmanı ile aynı boyuttadır, böylece orijinal veriler yeniden yapılandırılır. DDPM'lerde son katman, olasılık yoğunluk dağılımının ortalaması ve varyansı için tahsis edilmiş iki ayrı çıktıdan oluşur. Ağ modelinin hedefi, Eşitlik 4.6'daki kayıp(loss) fonksiyonunu optimize etmektir.

$$\begin{aligned}
E[-\log p_\theta(x_0)] &\leq E_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] \\
&= E_q \left[-\log p_\theta(x_T) - \sum_{t \geq 1}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]
\end{aligned} \tag{4.6}$$

Varyansın azaltılması ile daha iyi hale gelen kayıp fonksiyonu Eşitlik 4.7'de yeniden yazılmıştır. Burada L_T ileri difüzyon sürecinin sondan bir önceki adımındaki dağılım ile son adımındaki rastgele gürültünün dağılımı arasındaki

farkı gösteren ileri hatadır ve bu değer varyans planlamasına(schedule) bağlı olan sabit bir değerdir. $L_{1:T-1}$ geri difüzyon sürecindeki her adımda ileri adım ve geri adım dağılımları arasındaki farkların toplamını gösteren hatadır. L_0 ise kod çözme(decoding) hatasıdır.

$$\begin{aligned} \mathcal{L} = E_q & \left[\underbrace{D_{KL}(q(x_T|x_0)||p(x_T)}_{L_T} \right. \\ & + \sum_{t>1} \underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \\ & \left. \underbrace{-\log p_\theta(x_0|x_1)}_{L_0} \right] \end{aligned} \quad (4.7)$$

Eşitlik 4.7, SGD ile minimize edilirken eğitilebilecek tek terim $L_{1:T-1}$ 'dir. Sonsal dağılım $q(x_{t-1}|x_t, x_0)$ 'yı Bayes kuralı ile yeniden parametrelendirirsek Eşitlik 4.8'i elde ederiz.

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \bar{\mu}_t(x_t, x_0), \bar{\beta}_t I) \quad (4.8)$$

Burada gösterilen ortalama ve varyans Eşitlik 4.9'daki gibi hesaplanabilir.

$$\begin{aligned} \bar{\mu}_t(x_t, x_0) &:= \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \\ \bar{\beta}_t &:= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \end{aligned} \quad (4.9)$$

Geri difüzyon sürecinde bulunan dağılım $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ için varyans fonksiyonu $\Sigma_\theta(x_t, t) = \sigma_t^2 I$ olarak ayarlanmış, $\sigma_t^2 = \beta_t$ ve $\sigma_t^2 = \bar{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ için deneysel olarak aynı sonuçlar elde edilmiştir. Ortalamayı $\mu_\theta(x_t, t)$ bulmak içinse özel bir parametreleştirme kullanılmıştır. Bunun için Eşitlik 4.7'de verilen L_{t-1} teriminde $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$ olarak yazılırsa L_{t-1} Eşitlik 4.10'da verildiği gibi iki ortalama katsayısı(mean coefficient) arasındaki L2-hatası olarak görülebilir.

$$L_{t-1} = E_q \left[\frac{1}{2\sigma_t^2} \|\bar{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] + C \quad (4.10)$$

Burada C , θ 'ya bağlı olmayan bir sabittir. Burada μ_θ için en makul parametreleştirme $\bar{\mu}_t$ 'yi tahmin edebilecek bir model olmalıdır. $\epsilon = \mathcal{N}(0, I)$ olmak

üzere Eşitlik 4.3, $x_t(x_0, \epsilon) = \sqrt{\bar{a}_t}x_0 + (1 - \bar{a}_t)\epsilon$ olarak yeniden parametreleştirilirse formül Eşitlik 4.11'deki hale gelir.

$$\begin{aligned} L_{t-1} - C &= E_{x_0, \epsilon} \left[\frac{1}{2\sigma_t^2} \left\| \bar{\mu}_t \left(x_t(x_0, \epsilon), \frac{1}{\sqrt{\alpha_t}} x_t(x_0, \epsilon - \sqrt{1 - \alpha_t}\epsilon) \right) - \mu_\theta(x_t(x_0, \epsilon), t) \right\|^2 \right] \\ &= E_{x_0, \epsilon} \left[\frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(x_t(x_0, \epsilon) - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon \right) - \mu_\theta(x_t(x_0, \epsilon), t) \right\|^2 \right] \end{aligned} \quad (4.11)$$

Burada x_t 'nin modele girdi olarak verildiğini düşünürsek μ_θ için seçilebilecek parametreleştirme Eşitlik 4.12'de verilmektedir. Burada ϵ_θ , x_t 'den ϵ 'yi tahmin etmeyi amaçlayan bir fonksiyon tahminleyicisidir(approximator).

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t) \right) \quad (4.12)$$

Son durumda basitleştirilmiş kayıp fonksiyonu Eşitlik 4.13'teki hale gelir.

$$L_{t-1} - C = E_{x_0, \epsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \alpha_t)} \|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, t)\|^2 \right] \quad (4.13)$$

Özetle, ters süreçteki ortalama(mean) fonksiyonu μ_θ parametrelestirmesi değiştirilerek, μ_t veya ϵ 'u tahmin etmek için eğitilebilir. ϵ 'u tahmin eden parametreleştirme, hem Langevin dinamiğine benzer hem de difüzyon modelinin varyasyonel sınırını(variational bound), gürültü arındıran skor eşleştirmesine benzer şekilde basitleştirmış olur.

Bununla birlikte, Ho ve ark. varyasyon sınırının Eşitlik 4.14 varyantı üzerinde eğitim yapılmasıının örnek kalitesi için daha yararlı olduğunu ve uygulanmasının daha basit olduğunu belirtmişlerdir.

$$L_{\text{simple}}(\theta) := E_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, t)\|^2 \right] \quad (4.14)$$

Örnekleme algoritması. Örnekleme süreci ϵ_θ ile veri yoğunluğunun gradyanını öğrenen bir Langevin dinamiği sürecine benzemektedir. Örneklemenin tüm süreci Tablo 4.1 'de verilmektedir. Burada $t = T$ olduğu durumda $x_T \sim \mathcal{N}(0, I)$ alınarak süreci başlanır. $z \sim \mathcal{N}(0, I)$ olmak üzere $t - 1$ anında $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ örneği

Eşitlik 4.12'deki gibi bir ortalama fonksiyonu olan dağılımdan çekilmektedir.

Tablo 4.1 DDPM örneklemme algoritması

```

 $x_T \sim \mathcal{N}(0, I)$ 
for  $t = T, \dots, 1$  do
    if  $t > 1$  then
         $z \sim \mathcal{N}(0, I)$ 
    else
         $z = 0$ 
    end if
     $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$ 
end for
return  $x_0$ 

```

4.2.2 Gürültü Şartlı Skor Ağları

Tanımlar & Notasyonlar. Gürültü şartlı skor ağları[29], olasılık yoğunluk fonksiyonunun kendisi yerine gradyanlarını modelleyen enerji tabanlı modellerdir. Enerji tabanlı modellerde olasılık dağılımı $p_\theta(x)$ Eşitlik 4.15'teki gibi ifade edilir. Burada E_θ 'yı enerji fonksiyonudur. Olasılığı yüksek olan veri noktalarının enerjisi düşükken, olasılığı düşük olan veri noktalarının enerjisi yüksektir. Bu nedenle E_θ 'nın önünde negatif bir işaret bulunmaktadır. Bu fonksiyon bir sinir ağı olarak ifade edildiğinde ağı parametreleri θ ve giriş verileri x ile gösterilmektedir. Bu ağıın çıkışı, $-\infty$ ve ∞ arasında skaler bir değerdir. Eksponansiyel işlemi, herhangi bir olası girdiye sıfırdan büyük bir olasılık atanmasını sağlar. Eşitlik 4.15'te z_θ θ 'ya bağlı normalizasyon sabitidir ve x 'in sürekli olduğu durumda $\int_x e^{-E_\theta(x)} dx$ olarak ifade edilir. z_θ yoğunluk toplamının 1 olmasını sağlar. ($\int_x p_\theta(x) dx = 1$)

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{z_\theta} \quad (4.15)$$

Eşitlik 4.15'te $p_\theta(x)$ 'i hesaplamak için normalizasyon sabiti z_θ 'yı bulmak gereklidir. Bu da girdilerin yüksek boyutlu olduğu durumlarda büyük sinir ağları için hesaplanamaz(intractactable) bir değerdir. Normalde enerji tabanlı modeller zıtlıklı ıraksama(contrastive divergence)[42] gibi bazı yöntemler ile eğitilir. Skor-tabanlı modellerde ise yoğunluk fonksiyonunun kendisi yerine skor fonksiyonunu modellemek normalizasyon sabitini hesaplama zorluğu ile uğraşmamayı sağlar.

Skor fonksiyonu $s_\theta(x)$, Eşitlik 4.16'daki gibi ifade edilir. Normalizasyon sabitinin gradyanı $\nabla_x \log z_\theta = 0$ olduğu için skor fonksiyonu z_θ 'dan bağımsızdır. Bu sayede

normalizasyon sabitini uygulanabilir hale getirmek için özel bir mimari kullanmak da gerekmez.

$$\begin{aligned} s_\theta(x) &\approx \nabla_x \log p_\theta(x) \\ s_\theta(x) &= \nabla_x E_\theta(x) - \nabla_x \log z_\theta \\ s_\theta(x) &= -\nabla_x E_\theta(x) \end{aligned} \quad (4.16)$$

Normalizasyon sabitini bulma problemi ortadan kalktıktan sonra $p_\theta(x)$ normalize edilmiş olasılık yoğunluk fonksiyonu olarak bulunur ve Eşitlik 4.17'deki gibi maksimum olasılıkla(maximum likelihood) eğitilebilir.

$$\max_{\theta} \sum_{i=1}^N \log p_\theta(x_i) \quad (4.17)$$

Skor-tabanlı modellerin teoride temel mantığı bu şekildedir ancak pratikte karşılaşılan örnek kalitesinin yetersizliği probleme difüzyon süreci ile çözüm bulunmuştur. Bu modellerde Langevin dinamiği ile örneklemeye yapılırken veri yüksek boyutlu uzayda olduğu zaman ilk örneğin düşük yoğunluklu bölgeden olma ihtimali yüksektir. Örneklemenin doğru çalışması için önerilen bir çözüm, difüzyon süreci ile dağılıma çoklu gürültü eklenmesidir. Bu çoklu gürültü veriyi yavaş yavaş rastgele gürültüye dönüştürür. Veri dağılımı farklı seviyelerde Gauss gürültüsü ile bozulur ve bu bozulmuş dağılıma karşılık gelen skorlar tahmin edilir. Bu mantıkla eğitilen modellere ise Gürültü Şartlı Skor Ağları denir.

Difüzyon prosesinde veri noktaları gürültüye maruz edilir ve bir sonraki eğitim gerçek veri yerine gürültülü veri üzerinden devam eder. Gürültü yeterince büyük olduğunda düşük yoğunluklu veri bölgeleri doldurulur ve tahmin edilen skorlar iyileşir. Büyük gürültüler eklemek veriyi orijinal halinden belirgin ölçüde farklılaştırır, küçük gürültüler eklemek ise düşük yoğunluklu bölgeleri tam olarak örtmeyebilir. Uygun gürültü büyüğünü ayarlamak için çoklu ölçekte gürültüler art arda kullanılır. Tüm gürültülerin izotropik Gaussian olduğu ve standart sapmalarının $\sigma_1 < \sigma_2 < \dots < \sigma_L$ şeklinde verildiği kabul edilirse gürültülü veri dağılımı $p_{\sigma_i}(x)$ Eşitlik 4.18'deki gibi gösterilir.

$$p_{\sigma_i}(x) = \int p(y) \mathcal{N}(x; y, \sigma_i^2 I) dy \quad (4.18)$$

Çoklu ölçekte gürültü eklemek, skor-tabanlı üretici modellerin başarısı için kritik öneme sahiptir. Gürültü ölçeklerinin sayısını sonsuza genellemek, hem daha yüksek kaliteli ve kontrol edilebilir üretim yapabilmeyi sağlar hem de tam log-olasılık hesaplamasını mümkün kılar.

Eğitim hedefi. Gürültü şartlı skor ağları, veri dağılımı ile model arasındaki Fisher mesafesini minimize ederek eğitilir. Fisher ıraksaması gerçek veri skoru ile skor-tabanlı model arasındaki l_2 mesafesinin karesini hesaplar. Optimizasyon terimi Eşitlik 4.19'da verilmektedir.

$$E_{p(x)}[||\nabla_x \log p(x) - s_\theta(x)||_2^2] \quad (4.19)$$

Bu ıraksamayı direkt hesaplamak mümkün değildir çünkü veri dağılıminin skoru $\nabla_x \log p(x)$ bilinmemektedir. Bunun için eğitilen skor eşleştirme ağları farklı gürültü seviyelerindeki verilerin karıştırılması yoluyla veri skorunu tahmin etmeye çalışır. Skor-eşleştirme ağları direkt olarak bir veri kümesi üzerinde stokastik gradyan düşümü ile optimize edilir. Gürültü şartlı skor ağlarını eğitmek için tek şart girdi ve çıktı boyutlarının eşit olmasıdır.

Eşitlik 4.19'da verilen optimizasyon terimi Eşitlik 4.20'deki gibi açıldığında l_2 mesafesi $p(x)$ katsayısı ile ağırlıklandırılır. Tahmin edilen skor fonksiyonları bu nedenle düşük yoğunluklu bölgelerde yanlış çalışmaktadır.

$$E_{p(x)}[||\nabla_x \log p(x) - s_\theta(x)||_2^2] = \int p(x) ||\nabla_x \log p(x) - s_\theta(x)||_2^2 dx \quad (4.20)$$

Düşük yoğunluklu bölgelerdeki örnekleme problemine Eşitlik 4.21'de verilen gürültü arındıran skor eşleştirme (Denoising Score Matching-DSM)[18] yöntemi efektif bir çözüm sunar. Bu yöntemde orijinal skor yoğunluğu gitgide artan bir gürültü dizisi ile bozulmaktadır. Burada $p_\sigma(\bar{x}|x)$ gürültüyle bozulmuş veri dağılımını ifade eder. Bu yöntem sadece $p_\sigma(\bar{x}|x) \approx p_{data(x)}$ olacak kadar küçük gürültüler için doğru sonuç verebilir.

$$\frac{1}{2} E_{p_\sigma(\bar{x}|x)p_{data(x)}}[||s_\theta(\bar{x}) - \nabla_{\bar{x}} \log p_\sigma(\bar{x}|x)||_2^2] \quad (4.21)$$

Burada gürültülü veri dağılımı $p_\sigma(\bar{x}|x) = \mathcal{N}(\bar{x}|x, \sigma^2 I)$ olarak belirlenirse, $\nabla_{\bar{x}} \log p_\sigma(\bar{x}|x) = -\frac{\bar{x}-x}{\sigma^2}$ olarak bulunur. Buna göre verilen bir σ değeri için gürültü arındıran skor-eşleştirme hedefi Eşitlik 4.22'deki gibi ifade edilir.

$$\ell(\theta; \sigma) = \frac{1}{2} E_{p_{data}(x)} E_{\bar{x} \sim \mathcal{N}(x, \sigma^2 I)} \left[\left\| s_\theta(\bar{x}) + \frac{\bar{x} - x}{\sigma^2} \right\|_2^2 \right] \quad (4.22)$$

Eşitlik 4.22 mevcut bütün $\sigma \in \{\sigma_i\}_{i=1}^L$ 'ler için birleştirilirse farklı gürültü ölçeklerindeki Fisher ıraksamalarının ağırlıklı toplamı Eşitlik 4.23'te verilmiştir. Burada $\lambda(\sigma_i) \in R_{>0}$ pozitif ağırlık fonksiyonudur ve skor-eşleştirmesi optimize edilirken genellikle $\lambda(\sigma_i) = \sigma_i^2$ olarak seçilir.

$$\mathcal{L}(\theta; \{\sigma_i\}_{i=1}^L) = \frac{1}{L} \sum_{i=1}^L \lambda(\sigma_i) \ell(\theta; \sigma_i) \quad (4.23)$$

Başka bir skor-eşleştirme yöntemi olan dilimli skor eşleştirme (Sliced Score Matching-SSM) [43] ise skoru rastgele vektöre yansıtarak(projecting) ileri mod otomatik farklılaşma(forward mode auto-differentiation) yoluyla gürültüyle bozulmamış gerçek skoru tahmin eder. Bu yöntem DSM ile benzer sonuçlar elde etmektedir ancak 4 kat daha fazla hesaplama maliyeti gerektirir.

Örnekleme algoritması. Gürültü şartlı skor ağlarından örnekleme yapmak için Langevin dinamiği kullanılır. Stokastik gradyan inişi ile birleştiğinde elde edilen stokastik gradyan Langevin dinamiği, Markov zincirindeki gradyanları kullanarak dağılımdan örnekler üretebilir. Standart SGD ile karşılaştırıldığında, stokastik gradyan Langevin dinamikleri, yerel minimumlara yakalanmayı önlemek için parametre güncellemelerine Gaussian gürültüyü ekler. Örneklemenin tüm süreci Tablo 4.2'de verilmektedir.

Sürecin başında yapılandırılmamış rastgele gürültü içeren bir önsel dağılımdan örnek \bar{x}_0 alınarak başlanır. Algoritma boyunca örneklerin üzerinde görülen tüm çizgiler örneklerin gürültülü örnek olduğunu göstermektedir. Alınan örnekler ve a_1 adım miktarı ile p_{σ_1} 'den örnek çekilir. Daha sonra bulunan örneklerle bir sonraki adıma geçilir. Her bir adımda, adım miktarı küçülmektedir. Örnekler bu şekilde modifiye edilerek gerçeğe yakın örnekler elde edilir.

Gürültü kademe kademe düştüğü için bu süreçce tavlanmış Langevin dinamiği denir. Bu sayede düşük yoğunluklu bölgelerdeki tutarsızlık problemi ortadan kalkmaktadır. σ_1 yeterince büyük seçildiğinde p_{σ_1} 'deki düşük yoğunluklu bölgeler oldukça azalmakta ve yüksek kalitede örnek çekilmesi mümkün olmaktadır. Bu örnekler bir sonraki adım için iyi bir başlangıç noktası olarak görülür ve son adımdaki dağılım p_{σ_L} 'den de iyi kalitede örnek elde edilmesini sağlar.

Tablo 4.2 NCSN örneklemme algoritması

Gereksinimler: $\{\sigma_i\}, \epsilon, T$
 \bar{x}_0 ’ı başlat
for $i = 1, \dots, L$ **do**
 $a_i = \epsilon \cdot \sigma_i^2 / \sigma_L^2$
for $t = 1, \dots, T$ **do**
 Örnek çek $z_t \sim \mathcal{N}(0, I)$
 $\bar{x}_t = \bar{x}_{t-1} + \frac{a_i}{2} s_\theta(\bar{x}_{t-1}, \sigma_i) + \sqrt{a_i} z_t$
end for
 $\bar{x}_0 \leftarrow \bar{x}_T$
end for
return \bar{x}_T

4.2.3 Stokastik Diferansiyel Denklemler ile Skor-tabanlı Modelleme

Tanımlar & Notasyonlar. Song ve Ermon [30] DDPM ve NCSN modellerindeki gürültü ekleme adımlarını sürekli zamanlı stokastik bir süreç olarak tanımlayarak adım sayısı sonsuz olacak şekilde genelleştirmiştir. Stokastik süreçler stokastik diferansiyel denklemelerin(SDE) çözümüdür. SDE’ler Eşitlik 4.24’teki gibi ifade edilir. Burada $f(\cdot, t) : R^d \rightarrow R^d$ sürüklendirme(drift) katsayısı adı verilen vektör değerli bir fonksiyonu, $g(t) \in R$ difüzyon katsayısı adı verilen gerçek değerli bir fonksiyonu, \mathbf{w} standart Brownian hareketinin sembolünü ve $d\mathbf{w}$ sonsuz küçük beyaz gürültüyü göstermektedir.

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \quad (4.24)$$

Gürültü eklemenin yüzlerce yolu olduğu gibi SDE’nin seçimi de hep aynı olmayabilir. DDPM ve NCSN’lerdeki gürültü ekleme süreçleri SDE’lerin farklı versiyonlarıdır. İlgili SDE’ler DDPM için Eşitlik 4.25 ’te ve NCSN için Eşitlik 4.26’dan verilmektedir. Burada T sonsuza gittikçe $\beta(\frac{t}{T}) = T\beta_t$ ve $\sigma(\frac{t}{T}) = \sigma_t$ olarak alınır.

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}dt + \sqrt{\beta(t)}d\mathbf{w} \quad (4.25)$$

$$d\mathbf{x} = \sqrt{\frac{d[\sigma(t)^2]}{dt}}d\mathbf{w} \quad (4.26)$$

Örnek üretmek için SDE’yi tersten çözmek gereklidir. Her SDE’ye karşılık, Eşitlik

4.27'de verildiği gibi bir ters-SDE vardır. Burada herhangi bir t zamanındaki örnek dağılımı $p_t(x)$ ile gösterilmektedir. SDE zamanda geriye doğru $t = T$ 'den Eşitlik 4.30'da Fisher mesafelerinin ağırlıklı kombinasyonu bulunurken gürültü-arındırın skor-eşleştirme kullanılmaktadır. Bu yöntem hesaplama açısından daha verimli olmakla birlikte dilimlenmiş skor eşleştirme gibi yöntemlerle aynı performansla optimizasyon yapılabilir[30]. $t = 0$ 'a çözülmesi gerektiği için dt negatif yönlü sonsuz küçük zaman adımlarıdır. Ters zamanlı SDE'nin çözümleri gürültüyü kademeli olarak verilere dönüştüren difüzyon süreçleridir.

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_x \log p_t(\mathbf{x})]dt + g(t)d\mathbf{w} \quad (4.27)$$

Tüm difüzyon süreçlerine karşılık, yörüngeleri aynı marjinal olasılık yoğunluklarını paylaşan bir deterministik süreç bulunmaktadır. Bu deterministik süreç bir adı diferansiyel denkleme(ODE) karşılık gelir. Eşitlik 4.28'de verilen olasılık akışı ODE(probability flow ODE) ters zamanlı SDE ile yörüngeleri aynı marjinallere sahip bir adı diferansiyel denklemidir. Hem ters zamanlı SDE hem de olasılık akışı ODE, yörüngeleri aynı marjinallere sahip olduğu için aynı veri dağılımından örneklemeye izin verir. Skor fonksiyonu zamana bağımlı skor tabanlı model(bir sinir ağı) tarafından tahmin edildiğinden Eşitlik 4.28 bir sinirsel ODE örneğidir.

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\nabla_x \log p_t(\mathbf{x}) \right]dt \quad (4.28)$$

Eğitim hedefi. Stokastik diferansiyel denklemlerin çözümü sürekli rastgele değişken koleksiyonu $\{x(t)\}_{t \in [0, T]}$ ileyidir. Bu rastgele değişkenler, zaman endeksi t başlangıç zamanı 0'dan bitiş zamanı T 'ye doğru büyütükçe stokastik yörüngeleri izler. $x(t)$ 'nin marjinal olasılık yoğunluk fonksiyonunun $p_t(x)$ ile gösterildiğini varsayıyalım. Buradaki $t \in [0, T]$, sonlu sayıda gürültümüz olduğu zamanki $i = 1, 2, \dots, L$ ile benzerdir. Buradaki $p_t(x)$ ise oradaki $p_{\sigma_i}(x)$ ile benzerdir. Hiç bir gürültü uygulanmadan önce $t = 0$ anındaki veri dağılımı gerçek dağılımdır $p_0(x) = p(x)$. $p(x)$ 'e yeteri kadar uzun bir T süresi boyunca stokastik süreç ile gürültü eklenmesi sonucunda elde edilen $p_T(x)$ açıklanabilir bir gürültü dağılımına $\pi(x)$ dönüşür, buna önsel dağılım adı verilir. Sonlu ölçekte gürültümüz olduğu zaman son ve en büyük gürültü σ_L 'yi eklediğimizde elde ettiğimiz dağılım $p_{\sigma_L}(x)$, buradaki $p_T(x)$ 'e benzerdir. Eşitlik 4.24'te verilen SDE tipki sonlu ölçekte gürültümüz olduğundaki $\sigma_1 < \sigma_2 < \dots < \sigma_L$ gibi elle belirlenir.

Skor-eşleştirme ile ters SDE'nin çözülmesi için zaman bağımlı skor-tabanlı bir model $s_\theta(x, t) \approx \nabla_x \log p_t(\mathbf{x})$ eğitilir. Bu model sonlu gürültü olduğundaki

$s_\theta(x, \sigma_i)$ ile benzerdir. Eşitlik 4.29'da verilen eğitim hedefi gerçek veri dağılımının skoru ile model arasındaki Fisher mesafelerinin sürekli ve ağırlıklı bir birleşimidir. Burada $U(0, T)$, $[0, T]$ zaman aralığında düzenli(uniform) bir dağılımı ifade eder ve $\lambda : R \rightarrow R_{>0}$ pozitif ağırlıklandırma fonksiyonudur. Farklı skor-eşleştirme kayıplarının büyülüüğünü zaman içinde dengelemek için genellikle $\lambda(t) \propto 1/E[\|\nabla_x \log p(x(t)|x(0))\|_2^2]$ olarak alınır.

$$E_{t \in U(0, T)} [E_{p_t(x)} \lambda(t) \|\nabla_x \log p_t(\mathbf{x}) - s_\theta(x, t)\|_2^2] \quad (4.29)$$

$$\theta^\star = \arg \min_{\theta} E_t \left\{ \lambda(t) E_{x(0)} E_{x(t)|x(0)} [\|s_\theta(x(t), t) - \nabla_{x(t)} \log p_{0t}(x(t)|x(0))\|_2^2] \right\} \quad (4.30)$$

Eşitlik 4.30'u çözmek için geçiş çekirdeği(transition kernel) $p_{0t}(x(t)|x(0))$ 'yi bilmek gereklidir. $f(\cdot, t)$ afin olduğunda, geçiş çekirdeği ortalama ve varyansın genellikle kapalı formda standart tekniklerle elde edilebileceği bir Gauss dağılımıdır.

DDPM ve NCSN'de kullanılan gürültü süreçlerinin SDE'lerin farklı versiyonlarına karşılık geldiğinden bahsetmiştik. Eşitlik 4.25, $t \rightarrow \infty$ 'a giderken başlangıç dağılımının birim varyansa(unit variance) sahip olduğu durumda sabit varyanslı olmaktadır. Eşitlik 4.26 ise $t \rightarrow \infty$ 'a giderken her zaman patlayan varyanslı(exploding variance) bir süreç olmaktadır. Bu farktan dolayı Eşitlik 4.25 varyans korumalı(VP) SDE, Eşitlik 4.26 ise varyans patlamalı(VE) SDE olarak adlandırılmaktadır. Bunlara ek olarak VP-SDE'den yola çıkarak olasılıkta(likelihood) iyi çalışan başka bir SDE önerilmiştir. Önerilen bu SDE'ye alt VP-SDE adı verilmiştir.

VE, VP ve alt VP-SDE'lerin hepsi afin sürekli katsayısına sahip olduğundan geçiş çekirdekleri $p_{0t}(x(t)|x(0))$ Gaussiandır ve kapalı formda hesaplanabilir.

Örnekleme algoritması. Zaman bağımlı skor tabanlı model s_θ eğitildikten sonra sayısal yaklaşımalarla ters zamanlı SDE'ler çözülerek p_0 'dan örnek üretilebilir. SDE'leri çözmek için birçok genel amaçlı sayısal çözümleyici bulunmaktadır. Örneğin Euler-Maruyama çözümleyici ve stokastik Runge-Kutta yöntemleri ters zamanlı SDE'lere uygulanarak örnek üretilebilir. DDPM'in örnekleme yöntemi ters zamanlı VP-SDE'nin özel bir ayrıklaştırmasına(discretization) karşılık gelmektedir. Ayrıca ters zamanlı SDE'yi ileri yönlü SDE ile aynı şekilde ayırttıran ve bu nedenle ileri ayrıklaştırma verildiğinde kolayca örnek türetilen ters difüzyon

örnekleyicileri [30] klasik örnekleyicilerden biraz daha iyi çalışmaktadır.

Tahmin edici-Düzeltilci(Predictor-Corrector - PC) örnekleyicilerde her bir zaman adımında öncelikle sayısal SDE çözümleyici bir sonraki adımdaki örnek için bir tahminde bulunur, buna tahmin edici(predictor) denir. Ardından skor-tabanlı MCMC yaklaşımıorneğin marjinal dağılımını düzelterek düzeltici(corrector) rolünü oynar. PC örnekleyiciler NCSN ve DDPM'in orijinal örneklemeye yöntemlerini genelleştirmektedir. NCSN, birim(identity) fonksiyonu tahmin edici ve tavlanmış Langevin dinamığını düzeltici olarak kullanırken, DDPM sayısal çözümleyicileri tahmin edici ve birim fonksiyonu düzeltici olarak kullanır. VE-SDE ve VP-SDE için tahmin edici-düzeltilci örnekleyiciler Tablo 4.3 ve 4.4'teki algoritmalarla verilmektedir. Algoritmalarla N parametresi ters zamanlı SDE içi ayıralaştırma sayısını, M ise düzelticinin adım sayısını belirtmektedir. Tahmin edici yöntem olarak ters difüzyon SDE çözümleyici, düzeltici yöntem olarak da tavlanmış Langevin dinamığı kullanılmıştır. Langevin dinamığının adım miktarı $\{\epsilon_i\}_{i=0}^{N-1}$ olarak gösterilmektedir.

Tablo 4.3 VE-SDE için tahmin edici-düzeltilci örneklemeye algoritması

```

 $x_N \sim \mathcal{N}(0, \sigma_{max}^2 I)$ 
for  $i = N - 1, \dots, 0$  do
     $x'_i \leftarrow x_{i+1} + (\sigma_{i+1}^2 - \sigma_i^2) s_{\theta^*}(x_{i+1}, \sigma_{i+1})$ 
     $z \sim \mathcal{N}(0, I)$ 
     $x_i \leftarrow x'_i + \sqrt{\sigma_{i+1}^2 - \sigma_i^2} z$ 
    for  $j = 1, \dots, M$  do
         $z \sim \mathcal{N}(0, I)$ 
         $x_i \leftarrow x_i + \epsilon_i s_{\theta^*}(x_i, \sigma_i) + \sqrt{2\epsilon_i} z$ 
    end for
end for
return  $x_0$ 

```

4.2.4 Temel Çalışmaların Aralarındaki İlişki ve Sınırlamaları

Gürültü şartlı skor ağları[29] ve gürültü arındırılan olasılıksal difüzyon modelleri[28] arasında bir kez temel farklılık vardır. NCSN'ler skor-eşleştirme ile eğitilir ve Langevin dinamigi ile örneklenirken, DDPM'ler VAE'ler gibi ELBO ile eğitilir ve eğitilmiş bir kod çözücü ile örneklenir. DDPM, difüzyon modellerini eğitmek için kullanılan ELBO'nun esasen, skor-tabanlı üretici modellemede kullanılan skor-eşleştirme hedeflerinin ağırlıklı kombinasyonuna eşdeğer olduğundan bahsetmektedir.

Tablo 4.4 VP-SDE için tahmin edici-düzeltilci örneklemme algoritması

```

 $x_N \sim \mathcal{N}(0, I)$ 
for  $i = N - 1, \dots, 0$  do
     $x'_i \leftarrow (2 - \sqrt{1 - \beta_{i+1}})x_{i+1} + \beta_{i+1}s_{\theta^*}(x_{i+1}, i + 1)$ 
     $z \sim \mathcal{N}(0, I)$ 
     $x_i \leftarrow x'_i + \sqrt{\beta_{i+1}}z$ 
    for  $j = 1, \dots, M$  do
         $z \sim \mathcal{N}(0, I)$ 
         $x_i \leftarrow x_i + \epsilon_i s_{\theta^*}(x_i, i) + \sqrt{2\epsilon_i}z$ 
    end for
end for
return  $x_0$ 

```

Ayrıca, kod çözücüyü bir Gürültü şartlı skor ağları dizisi olarak parametrelendirip U-Net mimarisi kullanıldığında, ilk kez GAN'larla karşılaştırılabilir veya daha üstün kaliteli görüntüler üretilebileceğini göstermişlerdir.

Score SDE'ler [30] skor-tabanlı üretici modeller ile difüzyon modellerini aynı çerçeve içine almaktadır. Gürültü ölçeklerinin sayısını sonsuza genelleyerek, skor-tabanlı üretici modellerin ve difüzyon modellerinin her ikisinin de skor-fonksiyonları ile belirlenen stokastik diferansiyel denklemlerin farklı türleri olarak görülebileceğini kanıtlamışlardır. Huang ve ark. [44] log-olasılık tahminini açık bir şekilde izlemek için Brownian hareketini gizli bir değişken olarak ele almıştır. Bu çalışma skor eşleştirme kaybını minimize etmenin, Song ve ark. [30] tarafından önerilen ters SDE eklentisi ile ELBO'yu maksimize etmeye eşdeğer olduğunu göstererek teorik boşluğu doldurmaktadır. Bu gelişmeler doğrultusunda, skor-tabanlı üretici modeller ile difüzyon modellerinin aynı model ailesinin farklı perspektifleri olduğu anlaşılmaktadır. Bu model ailesi bu çalışmada literatürde kabul görmüş olan ismiyle, üretici difüzyon modelleri veya kısaca **Difüzyon modelleri** olarak ifade edilecektir.

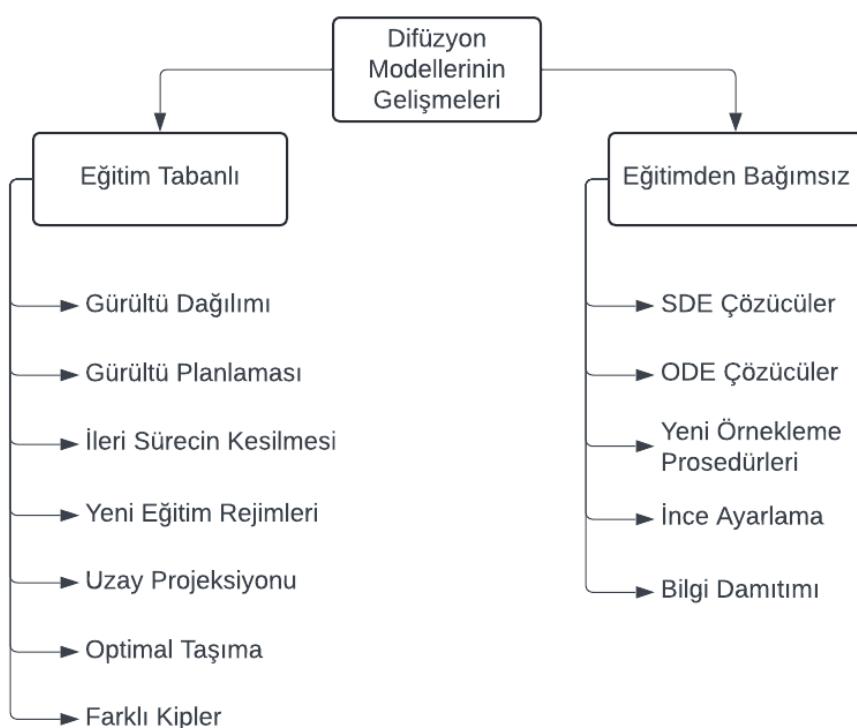
Difüzyon modelleri, yüksek üretkenlik kapasitesi ve izlenebilir(tractable) bir süreç sunar ancak, örneklemme adımlarının uzun olması ve örneklemme hızının yavaş olması bu modellerin uygulanabilirliğini sınırlamaktadır. Dahası eğitim sürecindeki bazı kilit faktörler, öğrenme kalıplarını(pattern) ve modellerin performansını etkiler. Eğitimi iyileştirmek için bu kilit faktörlerin etkisi araştırılmalıdır. Difüzyon modelleri için eğitim hedefi varyasyonel alt sınırıdır. Ancak bu sınır bir çok durumda sıkı(tight) olmadığından potansiyel olarak alt-optimal log-olasılık bulmaya neden olur. Bu nedenle difüzyon modelleri için olasılık maksimizasyonu da bir araştırma konusudur. ELBO ile log-olasılık aynı anda minimize edilmediğinden bazı

yöntemler doğrudan olasılık optimizasyonu problemine odaklanır. ELBO ile gerçek olasılık arasındaki varyasyonel boşluk(gap) optimizasyonu da bir araştırma konusudur.

Difüzyon modelleri, görüntü ve ses gibi veri alanları için büyük başarı elde etmiş olsa da, diğer modalitelere sorunsuz bir şekilde çevrilemezler. Birçok veri alanı(domain), difüzyon modellerinin etkin bir şekilde çalışması için dikkate alınması gereken özel yapılara sahiptir. Örneğin, modeller yalnızca sürekli uzayda tanımlanan skor fonksiyonlarına bağlı olduğunda veya veriler düşük boyutlu manifoldlarda bulunduğuanda, zorluklar ortaya çıkabilir. Bu zorluklarla başa çıkılmak için difüzyon modelleri çeşitli şekillerde uyarlanarak genelleştirilmelidir.

4.3 Difüzyon Modellerinin Teorik Gelişmeleri

Bu bölümde, difüzyon modellerinin algoritmasında yapılan iyileştirmeler incelenmektedir. İyileştirme çalışmaları 2 ana başlık altında kategorilendirilmiştir. Birincisi, eğitim tabanlı yaklaşımlar, ikincisi ise eğitimden bağımsız yaklaşımlardır. Bu kategoriler altındaki gelişmeler, konularına göre alt kategorilere ayrılmıştır. Tüm alt kategoriler Şekil 4.1 'de verilmektedir.



Şekil 4.1 Teorik gelişmelerin kategorileri

4.3.1 Eğitim Tabanlı Gelişmeler

Geleneksel eğitim şemasını değiştiren iyileştirmeler, eğitim sürecindeki temel faktörlerin öğrenme kalıplarını ve model performansını etkilediğini göstermiştir. Eğitim tabanlı yaklaşımalar, sırasıyla gürültü dağılımı, gürültü planlaması, ileri sürecin kesilmesi, yeni eğitim rejimleri, uzay projeksiyonu, optimal taşıma ve farklı kipler olmak üzere 7 alt başlık altında incelenmiştir.

4.3.1.1 Gürültü Dağılımı

Geleneksel difüzyon sürecinde gürültü dağılımı gaussiandır. Ancak, gürültü dağılımına daha fazla serbestlik derecesi verilirse, bu performansı iyileştirebilir. Nachmani ve diğerleri [45, 46] Gaussian olmayan gürültü dağılımlarını görüntü ve konuşma üretimi için uygulamış ve Gamma dağılımıyla daha iyi sonuçlar elde etmiştir. Bu çalışmalar ayrıca difüzyon sürecinde Gaussian gürültülerinin bir karışımının kullanılmasının tek bir dağılıma göre performansı iyileştirdiğini göstermektedir. Sonuçlardan bazıları Şekil 4.2'de gösterilmiştir. İlk satır Gauss gürültüsü, ikinci satır Gauss gürültüsünün karışımı ve üçüncü satır Gamma gürültüsüdür.



Şekil 4.2 Farklı gürültü dağılımları için sonuçlar[45].

Xiao ve diğerleri, [47] gürültü giderme sürecindeki Gauss varsayıminın örnekleme hızını yavaşlattığını ve bu varsayımin yalnızca küçük adım boyutları için işe yaradığını öne sürmüştür. Önerdikleriyle, gürültü giderme sürecindeki her adım koşullu bir GAN ile modellenerek daha büyük adım boyutuna izin verilmiştir. Bu yaklaşım örnekleme adımlarının sayısını azaltmaktadır. Ayrıca, Düzgün olmayan(non-uniform) difüzyon modelleri [48] ve İzotropik olmayan Gauss gürültü modelleri [49] standart düzgün(uniform) gürültüyü değiştirmek için farklı formülasyonlar kullanır. Her iki model de daha genel dağılımlarla karşılaştırılabilir sonuçlar elde etmişlerdir.

Yumuşak Difüzyon(Soft Diffusion) [50] gürültüleme sürecini gerçekleştirmek için Yumuşak Skor Eşleştirmeyi önermiştir. Bu model herhangi bir doğrusal gürültüleme süreci için önemli hesaplama avantajları elde etmenin yanı sıra en son

teknoloji(state-of-the-art) sonuçlara ulaşmıştır. Bulanıklaştıran Difüzyon(Blurring Diffusion)[51] ısı dağılımını veya bulanıklaştırmayı izotropik olmayan gürültülü bir Gauss difüzyon süreci olarak tanımlamıştır. Bu çalışma, standart Gauss gürültü giderme ile ters ısı dağılımı (inverse heat dissipation) [52] arasındaki boşluğu kapatmaya yönelik katkılarda bulunmuştur. Ayrıca Chen ve diğerleri [53] veri dağılıminin ve gürültü dağılıminin iki yörüngे ile düzgün bir şekilde bağlantılı olduğunu keşfederken literatüre katkıda bulunmuşlardır. Bunlardan birincisi açık(explicit), yarı doğrusal örnekleme yörüngesi ve ikincisi daha hızlı yakınsayan örtük bir gürültü giderme yörüngesidir.

4.3.1.2 Gürültü Planlaması

Gürültü ölçüğünü(scale) öğrenmek açıklanabilir bir gürültüleme süreci sağlayabilir ve bu da düzenli bir örnekleme süreci sağlar. Gürültü ölçüğünü ters süreçte ve ileri süreçte öğrenilebilir bir parametre olarak ele alan bazı yöntemler vardır. Temel yöntemler, ters Markov zincirindeki geçiş çekirdeklerinin sabit kovaryans parametrelerine sahip olduğunu varsayar ve bu kovaryans değerleri herhangi bir koşul dikkate alınmadan elle yapılır. Daha verimli örnekleme adımları elde etmek için kovaryans optimizasyonu dinamik olarak gerçekleştirilebilir ve ters süreçteki gürültü ölçüği bu şekilde öğrenilebilir.

Geliştirilmiş DDPM [54] ters kovaryansları bir doğrusal enterpolasyon(linear interpolation) olarak parametrelendirmeyi ve hibrit bir objektif fonksiyon kullanmayı önermiştir. Ters kovaryans Eşitlik 4.31'deki gibi parametrelendirilmiş ve örnek kalitesinden ödün vermeden daha yüksek log-olasılıklar ve daha iyi bir örnekleme hızı elde edilmiştir.

$$\Sigma_\theta(x_t, t) = e^{\theta \cdot \log \beta_t + (1-\theta) \cdot \log \bar{\beta}_t} \quad (4.31)$$

Geliştirilmiş DDPM, Eşitlik 4.32'de verilen kosinus gürültü planlamasının log olasılıklarını iyileştirebileceğini göstermiştir. Burada $s, t = 0$ zamanında gürültü ölçüğünü kontrol etmek için verilen bir hiperparametredir. Bu yaklaşım literatürde kabul görüp standart bir yöntem haline gelerek birçok çalışmada yaygın olarak kullanılmaktadır.

$$\bar{a}_t = \frac{h(t)}{h(0)}, \quad h(t) = \cos\left(\frac{t/t+s}{1+s} \cdot \frac{\pi}{2}\right) \quad (4.32)$$

FastDPM [55], farklı alanlar(domain), farklı veri kümeleri ve koşullu üretimde

farklı miktarda bilgi için hızlı örnekleme yöntemlerini inceler. FastDPM, gürültü tasarımını ELBO optimizasyonuyla ilişkilendirerek DDPM için varyans sabitinden, DDIM [56] içinse zaman adımından türetir. Ek olarak, bu çalışma yöntemlerin performansının alana (örneğin, görüntü veya ses), örnekleme hızı ile örnek kalitesi arasındaki dengeye ve koşullu bilgi miktarına bağlı olduğunu göstermiştir.

DiffFlow [57], her adımda bir akış fonksiyonu ile ileri ve geri işlem arasındaki KL-ıraksamasını en aza indirerek gürültü ekleme sürecini yürütür. Bu yöntem, akış fonksiyonlarının geri yayılımı(backpropagation) nedeniyle adım başına daha uzun bir zaman gerektirir, ancak daha az adım gerektirdiği için DDPM'den hala 20 kat daha hızlıdır. DiffFlow, DDPM'ye kıyasla daha az ayriklaştırma(discretization) adımıyla daha genel dağılımları öğrenerek katkılarında bulunmuştur.

Varyasyonel Difüzyon Modelleri (VDM) [58], gürültü planlamasını ve diğer difüzyon modeli parametrelerini birlikte eğiterek sürekli zamanlı difüzyon modellerinin log olasılığını iyileştirmiştir. Gürültü süreci, monoton bir sinir ağı $\gamma_\eta(t)$ kullanılarak parametrelendirilir. İleri gürültü süreci $\sigma_t^2 = \text{sigmoid}(\gamma_\eta(t))$, $q(x_t|x_0) = \mathcal{N}(\bar{a}_t x_0, \sigma_t^2 I)$ ve $\bar{a}_t = \sqrt{(1 - \sigma_t)^2}$ olarak tasarlanmıştır. Ayrıca yazarlar, herhangi bir x veri noktası için varyasyonel alt sınır VLB'nin yalnızca sinyal-gürültü oranına $R(t) := \frac{\bar{a}_t^2}{\sigma_t^2}$ bağlı bir forma basitleştirileceğini belirtmişlerdir. Son durumda L_{VLB} Eşitlik 4.33'teki gibi gösterilebilir. Burada birinci ve ikinci terimler VAE'ler gibi doğrudan optimize edilebilir. Üçüncü terim ise Eşitlik 4.34'te verilmiştir.

$$L_{VLB} = -E_{x_0} \text{KL}(q(x_T|x_0) || p(x_T)) + E_{x_0, x_1} \log p(x_0|x_1) - L_D \quad (4.33)$$

$$L_D = \frac{1}{2} E_{x_0, \epsilon \sim \mathcal{N}(0, I)} \int_{R_{min}}^{R_{max}} \|x_0 - \bar{x}_\theta(x_\nu, \nu)\|_2^2 d\nu \quad (4.34)$$

Eşitlik 4.34'teki $R_{max} = R(1)$ ve $R_{min} = R(T)$, $x_\nu = \bar{a}_\nu x_0 + \sigma_\nu \epsilon$ x_0 'ın $t = R^{-1}(\nu)$ olana kadar ileri difüzyondan elde edilen gürültülü veri noktasını ifade eder. \bar{x}_θ , difüzyon modeli tarafından tahmin edilen gürültüsüzleştirilmiş veri noktasını gösterir. Gürültü süreçleri VLB'yi etkilemez çünkü R_{min} ve R_{max} aynı değerlere sahiptir. Bu parametrelendirme yalnızca Monte Carlo tahmin edicilerinin varyansını etkiler. VDM, otoregresif modellerden önemli ölçüde daha hızlı bir optimizasyonla görüntü yoğunluğu tahmini üzerinde en son teknoloji log-olasılıklarını elde etmiştir.

İkili gürültü giderme difüzyon modelleri (BDDM) [59] bir skor ağı ve bir planlama ağını birlikte öğrenir ve çıkışım zamanındaki gürültü planlamalarını optimize

ederek örnekleme sürecini hızlandırır. Katkıları, sundukları alt sınırın standart kanıt alt sınırından daha sıkı olduğunu ve sadece 3 örnekleme adımıyla yüksek doğrulukta örnekler ürettiğini göstermiştir.

Dinamik Çift Çıktılı Difüzyon Modelleri [60] gürültü giderme işlemi için iki zit denkleme sahiptir, birincisi uygulanan gürültüyü tahmin etmek için, ikincisi ise görüntüyü tahmin etmek içindir. Kalite ve hız açısından verimlilik elde etmek için bunlar arasında dinamik olarak geçiş yapılır. Bu çalışma, en son teknoloji mimarilere uygulandığında genel bir çözüm ve geliştirilmiş üretim kalitesi sağlar.

San Roman ve diğerleri [61], de gürültü giderme sürecini ayrı bir sinir ağı ile kontrol etmiş ve şartlandırma verilerine dayalı bir gürültü planlaması elde etmiştir. Bu yaklaşım, gürültü planlamasını ayrı tahmin ederek örnekleme hızını ve performansını iyileştirmeye katkı sağlamıştır.

Lin ve diğerleri [62] gürültü planlamasını yeniden ölçeklendirdi ve son adımı sıfır sinyal-gürültü oranına (SNR) sahip olacak şekilde uyguladı. Yaptıkları değişikliklerle, çıkışım sürecinin eğitim süreciyle daha uyumlu olmasını ve modelin orijinal veri dağılımıyla benzer örnekler üretmesini sağlamıştır.

Optimal doğrusal alt uzay araması (OLSS) [63], daha hızlı bir zamanlayıcı önermiştir. Bu yöntem, optimal tahminlemeyi(approximation) arayarak üretim sürecinin hızlanmasına katkıda bulunur.

Chen ve diğerleri [64] gürültü planlamasının görüntü boyutuna bağlı olması gerektiğini öne sürmektedir. Görüntü boyutu arttıkça daha gürültülü bir planlama uygulanmalıdır. Başlıca katkıları, giriş verilerini basitçe b faktörüyle ölçeklerken gürültü planlaması işlevini sabit tutmanın ($\log b$ 'ye eşdeğer) farklı görüntü boyutları için daha iyi olduğunu göstermektedir.

Ayrıca Choi ve diğerleri [65] ağırlıklandırma şemasını yeniden tasarlamış ve bazı gürültü seviyelerine öncelik vererek daha iyi sonuçlar elde etmiştir. Bu genel tasarım, çeşitli mimariler ve örnekleme stratejileri üzerinde çalışabilmekte ve performansı önemli ölçüde iyileştirmektedir.

Gürültü giderme süreciyle ilgili başka bir bakış açısı Cold Diffusion [66] çalışmasında önerilmiştir. Cold diffusion, difüzyon modellerinin üretken davranışının gürültü seçimine güçlü bir şekilde bağlı olmadığını ancak bu seçimi değiştirek bir üretken model ailesinin tasarlanabileceğini iddia eder. Tamamen deterministik bir gürültü giderme süreciyle güncelleme kurallarını genelleştirerek farklı modeller elde ettiler. Bu katkılar, Langevin dinamikleri veya varyasyonel

çıkarmı gibı genel perspektifin ötesinde keyfi(arbitrary) süreçleri tersine çeviren genelleştirilmiş difüzyon modellerinin yolunu açar.

4.3.1.3 İleri Sürecin Kesilmesi

Kesmenin ana fikri, ileri difüzyon sürecini erken durdurmak ve gürültü giderme sürecini Gauss olmayan bir dağılımla başlatmaktadır. Kesme işlemi, örnekleme hızı ve örnek kalitesini dengeleyen bir kesme hiperparametresine sahiptir.

Kesmede, daha az difüze verilerden üretim, GAN ve VAE gibi diğer üretken modellerin yardımıyla gerçekleştirilebilir. TDPM [67] hem difüzyon hem de örnekleme sürecini kesebilmek için GAN ve koşullu taşıma (CT) [68] ile bulunan örtük bir dağılımdan örnekler alır. TDPM'nin üretim performansı ve gereken ters difüzyon adımlarının sayısı açısından katkıları vardır.

ES-DDPM [69], örneklerin yörüngelerini öğrenmek için erken durdurma fikrini önerir. ES-DDPM, örneklemeyi erken durdurmak için DiffuseVAE [70] örneğinden gerekli koşulu öğrenir. ES-DDPM, ayrıca görüntü oluştururken hem küresel hem de bölgesel olarak anlamsal kontrole olanak sağlar.

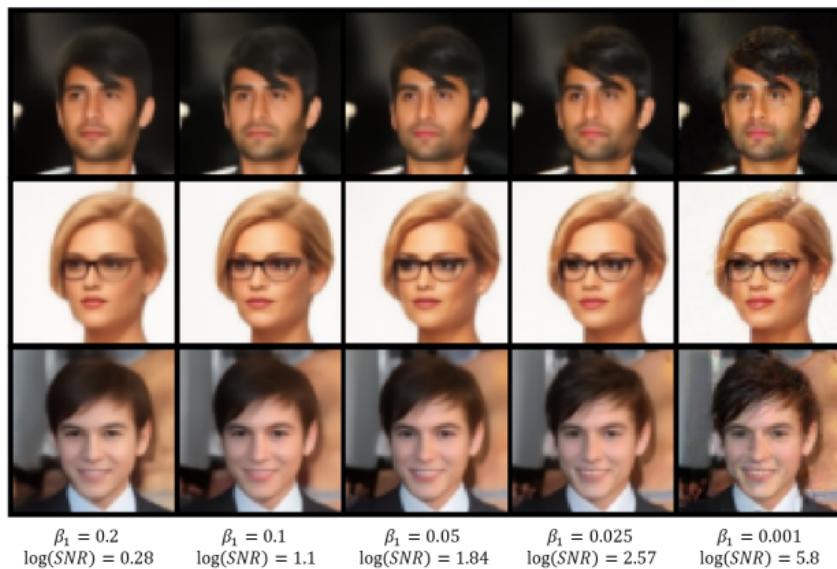
Difüzyon modelleri yoğunluk tahmini ile örnekleme performansı arasında ters bir korelasyon içerir. Küçük difüzyon süresi yoğunluk tahminine önemli ölçüde katkıda bulunurken, büyük difüzyon süresi örnek kalitesini iyileştirir. Kim ve diğerleri [71], Yumuşak Kesme(Soft Truncation)'de yeterli deneysel kanıtla böyle bir ters korelasyonu araştırır. Bu yöntem, sabit kesme hiperparametresini rastgele bir değişkene yumaşatan bir eğitim teknigi olarak önerilmiştir. Yöntemin Gürültü koşullu skor ağlarına uygulanmasıyla NCSN++ modeli, gürültü arındırın olasılıksal difüzyon modellerine uygulanmasıyla DDPM++ modeli elde edilmiştir. Her iki model de karşılaştırma veri kümelerinde son teknoloji sonuçlar elde etmiştir.

CCDF [72], gürültüden arındırma sürecini Gauss gürültüsünden başlatmaya gerek olmadığını ve bu şekilde örnekleme adımlarını azaltarak en son teknoloji sonuçların elde edilebileceğini göstermiştir. Pahali olmayan bir veri tutarlılığı adımıyla ters difüzyonu gerçekleştirerek yeni bir örnekleme stratejisi önermişlerdir. Ayrıca, Franzese ve diğerleri [73] ELBO'yu minimize etmek için difüzyon adımlarının sayısını optimize etmiştir. Burada, ideal ve simüle edilmiş ileri dinamikler arasındaki boşluğu kapatmak için yardımcı bir modelden yararlanılmıştır. Standart bir ters difüzyon sürecini takip eden bu yardımcı model, örnek kalite ölçümüleri ve log-olasılık bağlamında rekabetçi sonuçlar elde etmiştir.

4.3.1.4 Yeni Eğitim Rejimleri

Bazı yaklaşımlar hesaplama maliyetini düşürmek ve daha iyi üretim kalitesi elde etmek için farklı eğitim stratejileri önermektedir. Difüzyon modellerini iyileştirmek için diğer mevcut üretken modellerin avantajlarını kullanan bazı yöntemler vardır. Jolicoeur-Martineau ve diğerleri [74], hem Gürültü-Arındıran Skor Eşleştirmesi (DSM) hem de çekişmeli hedeflerden oluşan Çekişmeli Skor Eşleştirmesi (ASM) adı verilen bir hibrit eğitim formülasyonu önermişlerdir. Tavlanmış Langevin Örneklemesine (ALS) daha kararlı bir alternatif olarak Tutarlı Tavlanılmış Örneklemme (CAS) ile katkıda bulunmuşlardır. Ayrıca görsel kalite ile yüksek FID arasındaki sorunun nasıl çözüleceğini de göstermişlerdir.

Deja ve diğerleri, [75] difüzyon sürecinin adımlarının yeniden yapılandırma hatasıyla nasıl ilişkilendirildiğini analiz eder. Analizlerine dayanarak, bir Gürültü arındıran otokodlayıcı(Denoising Auto-Encoder) ve Difüzyon tabanlı bir üreticiden oluşan yeni bir model sınıfı olan Difüzyonlu Gürültü arındıran otokodlayıcı(DAED)'yı önerirler. Gürültüden arındırma sürecini iki parçaya bölen bir anahtarlama(switching) noktası bulunur. Bu parçalardan birincisi gürültüden arındırma parçası ve diğeri üretici parçasıdır ve bu parçalarda farklı prosedürler kullanılır. Şekil 4.3, aynı gürültüye ve farklı anahtarlama noktalarına sahip DAED'den örnekler gösterir. Burada daha yüksek β , uzun vadeli gürültü anlamına gelmektedir.



Şekil 4.3 Farklı anahtarlama noktaları(β) için DAED'den örnekler[75].

Çoklu mimari Çoklu Uzman (MEME) [76] difüzyon sürecinin her zaman adımında alt-optimal performanslar üreten çok farklı işlevselliklere sahip olduğunu göstermiştir. Bu nedenle, her zaman adımı için en uygun işlemleri yapmak üzere gürültü aralıklarında farklı uzmanları dikkate alırlar. Ayrıca, Cho ve diğerleri[77],

biri mekansal içerik kodu ve diğerisi stil kodu olmak üzere iki gizli kod ile bir difüzyon modeli eğitmişlerdir. Ayrıca, kontrol edilebilir örneklemeyi iyileştirmek için iki yeni yöntem önermişlerdir.

Yi ve diğerleri [78] difüzyon modellerinin genelleme yeteneğine vurgu yapmış ve genelleme sorunu olmayan yeni bir hedef önermiştir. Önerilen hedef orijinaline benzer bir model döndürür ve ayrıca genelleme yeteneğine sahiptir. Hızlı Difüzyon Modeli (FDM) [79] hem eğitimi hem de örneklemeyi hızlandırmak için stokastik optimizasyon perspektifini kullanır. SGD'den daha hızlı ve daha kararlı yakınsama elde etmek için hem gradyan hem de ekstra bir momentum kullanan momentum SGD'den esinlenmişlerdir ve momentumu difüzyon sürecine entegre etmişlerdir. Bu şekilde daha hızlı bir yakınsama elde etmiş ve eğitim maliyetini düşürmüştür.

Xu ve diğerleri, [80] eğitim hedeflerinin varyansını azaltmayı önermiştir. Kararlı Hedef Alanı (STF) adı verilen genelleştirilmiş skor eşleştirme hedefinde ağırlıklı koşullu skorları hesaplamak için bir referans batch kullanılmıştır. Bu şekilde, önerilen STF hedefi skor tabanlı yöntemlerin performansını, kararlılığını ve eğitim hızını iyileştirir. Ayrıca, Ning ve diğerleri [81] maruz kalma önyargısı (exposure bias) sorununu hafifletmek için gerçek örnekleri bozarak eğitimi düzenli hale getirdi.

Phoenix [82] koşulsuz bir difüzyon modelini eğitmek ve istatistiksel heterojenliğe sahip veya Non-IID (Bağımsız ve Özdeş Dağıtılmış) olan verilerde bile veri çeşitliliğini(diversity) iyileştirmek için federasyonlu öğrenmeyi (federated learning) kullandı. Bu çalışma ayrıca veri gizliliğini koruma ve veri kaynakları arasındaki iletişimini azaltma konusunda da katkılarında bulunmuştur.

Gürültü arındıran öğrenci (Denoising Student) [83] çok adımlı gürültü arındırma sürecini tek bir adıma indirgedi ve diğer tek adımlı üretken modellere (GAN, VAE) benzer bir örneklemeye hızı elde etti. Ayrıca, bu model gürültü planı ve adım boyutu gibi hiperparametrelere sahip olmadığından bunların optimizasyonu ile uğraşmaya da gerek yoktur. ProDiff [84]'te, metinden sese üretim için bilgi damıtımı(knowledge distillation) [85] kullanılmıştır. Burada, öğrenciler iki kategorik dağılım arasındaki KL-ıraksamasını en aza indirerek bilgiyi doğrudan sıfırdan damıtırlar. ProDiff, yalnızca 2 adımla yüksek doğrulukta mel-spektrogramları sentezler ve uzun vadeli yinelemelerle diğer son teknoloji çalışmalarına karşı rekabetçi kalite ve çeşitlilik performansına sahiptir.

Piramidal difüzyon modeli [86], konumsal bir yerleştirme ile bir puan fonksiyonunu eğitti ve çok ölçekli süper çözümürlük problemini tek bir skor fonksiyonu kullanarak çözdü. Performanstan ödün vermeden zaman açısından verimli bir

çerçevede düşük çözünürlüklü girdilerden yüksek çözünürlüklü görüntüler ürettiler. Ayrıca Basamaklı(Cascaded) Difüzyon Modeli (CDM) [87], artan çözünürlükte görüntüler üreten bir difüzyon modelleri zinciridir ve koşullandırmayı artırmanın (augmentation) koşullu üretim görevlerinde en son teknoloji FID puanlarına ulaşmıştır.

Sınıflandırma ve Regresyon Yayılma Modelleri (CARD) [88] hem ileri hem de geri difüzyon süreçlerine kovaryat(covariate) bağımlılığını ve öneğitimli bir koşullu ortalama tahmin edicisi enjekte eder. Bu şekilde, verilen koşullar altında veri dağılımını tahmin etmek için $p(y|x, D)$ 'nin tam ve kesin bir tahminini sağlayan bir model elde edilmiştir.

Sınıflandırıcıdan bağımsız rehberlik (Classifier-free guidance) [89] koşullu ve koşulsuz bir difüzyon modelini birlikte eğitir ve tahmin ettikleri skorları birleştirerek örnek kalitesi ve çeşitliliğini dengeler. Sınıflandırıcıdan bağımsız rehberlik, koşullu görüntü oluşturma görevlerinde yaygın olarak kullanılmıştır. Örneğin, Kendini yönlendiren(Self-guided) difüzyon modeli [90] rehberlik için görüntü-açıklama çiftleri elde etmenin maliyetine dikkat eder ve kendi kendini denetleme sinyallerini kullanarak açıklama ihtiyacını ortadan kaldırır.

Blattmann ve diğerleri [91, 92] harici bir veritabanından en yakın komşuların bir kümesini alır ve bu örnekleri modeli koşullandırmak için kullanır. Ayrıca, belirli görevlerde üretim performansını iyileştirmek için Almayla artırılmış (Retrieval augmented) yöntemleri kullanan bazı çalışmalar da vardır[93–95].

4.3.1.5 Uzay Projeksiyonu

Normalleşiren akışlara veri dönüşümü uygulayan uzay projeksiyonu yöntemlerinden Parametrelî Difüzyon Modeli (PDM) [96] akış fonksiyonu ile gizli değişkenleri bularak daha hızlı bir hesaplama elde etmiştir. Ek olarak, ELBO ve log-olasılık optimizasyonları arasındaki boşluğa karşılık gelen varyasyonel boşluk ifadesini tanımlamışlardır. Ayrıca, kolektif öğrenme ile boşluğu ortadan kaldırmak için bir çözüm önermişlerdir. Ayrıca, Kapalı Doğrusal Olmayan Difüzyon Modeli (INDM) [97] akış modelini varyasyonel boşluğu ifade etmek için kullanır ve çift yönlü akış modeli ve doğrusal difüzyon modelini gizli uzayda birlikte eğiterek boşluğu en aza indirir. INDM [97] ve PDM[96]'nin daha küçük bir alanda modeller oluşturması, daha az değerlendirme adımı ve daha hızlı örneklemeyi sağlar.

p_{θ}^{ODE} , \cdot yi doğrudan maksimize etmenin maliyetini azaltmak için Song ve diğerleri [98], p_{θ}^{SDE} , \cdot nin varyasyonel alt sınırını maksimize etmeyi önermiş ve ScoreFlows adı verilen bir difüzyon modelleri ailesi tanıtmıştır. ScoreFlows, veri dağılımını

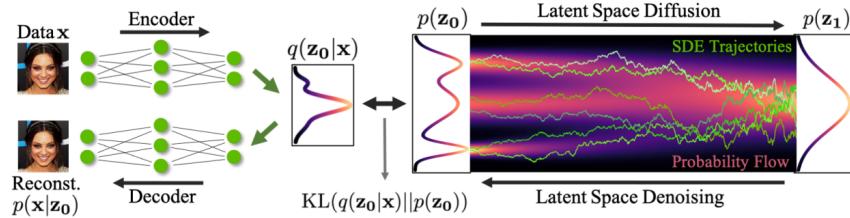
bir dekuantizasyon uzayına dönüştürmek ve dekuantize edilmiş örnekler üretmek için normalleştiren akışları kullanarak difüzyon sürecini çalıştırır. Süreci dekuantizasyon uzayında çalıştmak, sürekli ve ayrık veri dağılımları arasındaki uyumsuzluğu çözer.

Lu ve diğerleri, [99], yalnızca skor eşleştirme kaybını değil, aynı zamanda daha yüksek dereceli genellemeleri de en aza indirmeyi önererek ScoreFlows'u daha da geliştirir. $\log p_\theta^{\text{ODE}}$ 'nin birinci, ikinci ve üçüncü derece skor eşleştirme kayıplarıyla sınırlandırılabilceğini kanıtlamışlardır. Bunun da ötesinde, yüksek dereceli skor eşleştirme kayıplarını en aza indiren verimli eğitim algoritmalarıyla p_θ^{ODE} , yi iyileştirmişlerdir. ODE olasılığı ile skor eşleştirme hedefleri arasında bir boşluk olduğunu öne sürerek, birinci derece skor eşlestirmenin ODE olasılığını en üst düzeye çıkarmak için yeterli olmadığını kanıtladılar. Skor tabanlı difüzyon ODE'lerinin maksimum olasılık eğitimini sağlamak için yeni bir yüksek dereceli gürültü arındıran skor eşleştirme yöntemi önerdiler. Birinci, ikinci ve üçüncü derece skor eşleştirme hatalarını takip ederek ODE'nin negatif olasılığını sınırlamışlardır. Daha sonra bu yöntemle yüksek olasılık ve örnek kalitesi elde etmişlerdir.

Kritik söyümlenmiş Langevin Difüzyonunda (CLD)[100], süreç değişkenlerin verilere bağlı olan "hızlar(velocities)" olarak kabul edildiği genişletilmiş bir uzaydadır. Bu yöntemle ters SDE'yi elde etmek için, veri skorlarını doğrudan öğrenmek yerine "hızların" koşullu dağılımının skor fonksiyonunu öğrenmek yeterlidir. Bu yöntemin örnekleme hızını ve kalitesini iyileştirme açısından katkıları vardır.

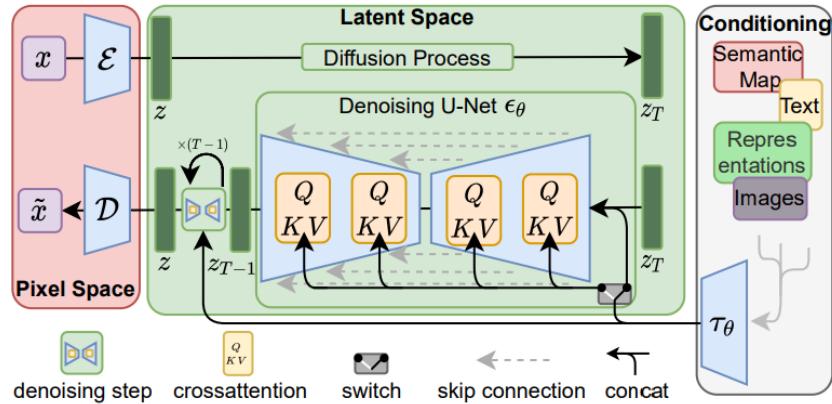
Ters süreçte her adımda yüksek boyutlu bir girdiyi yüksek boyutlu bir çıktıya eşlemek hesaplama maliyetini artırır. Verileri düşük boyutlu bir uzaya yansıtma, modeli hızlandırmak için akla gelen ilk çözümlerden biridir. Ancak, difüzyon modelleri eşdeğer uzaylarda olması gereken geçiş çekirdeklerine dayanır ve bu boyut azalmasını önler. Bu nedenle, difüzyon modelleri bir otokodlayıcı kullanılarak gizli uzayda eğitilebilir. Şekil 4.4'teki Gizli Skor Tabanlı Üretken Model (LSGM) [101], verileri bir kodlayıcı ile gizli uzaya eşler ve difüzyon işlemi gizli uzayda uygulanır. Daha sonra gizli uzayda üretilen örnekler bir kod çözümü kullanılarak veri uzayına eşlenir. LSGM ayrıca ELBO'nun gizli uzay difüzyonu için belirli bir skor eşleştirme hedefi olduğunu göstermiştir. Bu nedenle ELBO'daki çözülemeyen çapraz entropi terimi, skor eşleştirme hedefine dönüştürüllererek hesaplanabilir. VAE'nin kanıt alt sınırını (ELBO) ve difüzyon modelinin skor eşleştirme hedefini birleştirerek log-olasılık için yeni bir alt sınır optimize etmişlerdir. LSGM, örnekleri daha düşük bir boyutla gizli uzaya yansittığı için daha hızlı bir üretim performansına sahiptir. Ayrıca, LSGM, gizli uzaya

yansıtarak ve bunları sürekli uzayda kullanarak ayrı verilerle çalışma konusunda katkılarında bulunmuştur.



Şekil 4.4 Gizli Skor Tabanlı Üretken Model (LSGM) ’in mekanizması[101].

Öte yandan, Şekil 4.5’teki Gizli Difüzyon Modeli (LDM) [102] öncelikle otokodlayıcıyı eğiterek düşük boyutlu bir gizli uzay elde etmiştir. Daha sonra difüzyon modeli bu gizli uzayda, gizli kodları üretmektedir. LDM, UNet’i 2B parametreli evrişimsel katmanlardan inşa etmekte, böylece yeniden ağırlıklandırılmış hedef en alaklı bitlere odaklanmaktadır. LDM’nin temel katkılarından biri, UNet’i esnek koşullu üretimdeki etkinlikle birlikte gelen çapraz dikkat(cross-attention) mekanizmasıyla güçlendirmektir. Stable Diffusion, bir gizli difüzyon modelidir ve en sık kullanılan metinden görüntü üretime modellerinden biridir.



Şekil 4.5 Gizli Difüzyon Modeli (LDM)’in mekanizması [102].

Alt Uzay Difüzyonu [103], difüzyon modellerinde boyut indirmeyi araştırmıştır. İleri difüzyon sürecinde girdiyi daha küçük bir alt uzaya yansıtarak bu alt uzay modellerini eğitmiş ve değerlendirmiştirlerdir. Bu sayede örnek kalitesinde ve çalışma zamanında iyileştirmeler elde etmişlerdir.

Pandey ve diğerleri [104], ileri sürecin istenen bir ön dağılıma asimptotik olarak yakınsamayı garanti eden bir parametrelendirmesini önermiştir. İleri süreci yardımcı değişkenlerle zenginleştirilmiş bir alanda oluşturulmuş ve Faz Uzayı Langevin Difüzyonunu (PSLD) önermişlerdir. Üstün örnek ve hız kalitesi sunan bir reçeteyle literatüre katkı sağlamışlardır.

DPM-Encoder tabanlı CycleDiffusion [105], ilgili alanlarda bağımsız olarak eğitilen iki difüzyon modelinden ortaya çıkan gizli bir alana sahiptir. CycleDiffusion görüntüleri kodlamak ve kodlarını çözmek için ortak gizli alanı kullanır. Ayrıca, yöntemlerini büyük ölçekli metinden görüntü üreten difüzyon modellerine uygulayarak sıfır-atımlı(zero-shot) bir görüntü düzenleyici sunarak katkıda bulunurlar.

Çok-modlu Gizli Difüzyon [106], bağımsız olarak eğitilmiş tek modlu otokodlayıcılarından oluşan bir dizi kullanır. Bireysel gizli değişkenleri daha sonra ortak bir gizli uzayda birleştirir. Ortak ve koşullu üretme yetenekleri, ortak gizli değişkenle ilişkili bir olasılık yoğunluğunu öğrenen bir difüzyon modeli tarafından sağlanır. Bu çalışma, hem ortak hem de koşullu üretim için kullanılabilen çok modlu skor ağını eğitmek için yeni bir yöntem olarak katkılarda bulunmuştur.

Boyut değiştiren difüzyon süreci (DVDP) [107] evrim sürecinde, özellikle erken nesil aşamasında, yüksek boyutluluğun korunmasına gerek olmadığını savunmuştur. Kaynak görüntüyü, çok az bilgi kaybı olan düşük boyutlu bir sinyale dönüştürmüştür. Bu teknik hesaplama maliyetini düşürerek optimizasyon zorluğunu azaltmıştır. Ayrıca, Zıplayan difüzyon süreci [108] yeni bir ELBO öğrenmesiyle ileri süreçte boyutu yok ederek ve ters süreçte yeniden oluşturarak boyutsal uzaylar arasında atlar.

Xu ve diğerleri, Poisson-akışı üretken modelleri (PFGM) [109]’da veri noktalarını artırılmış bir uzaya elektrik yükleri olarak yansımış ve çalışmalarını difüzyon modelleriyle birleştirerek PFGM++ [110]’daki büyük ölçekli veri kümelerine genişletmişlerdir. Katkıları, artırılmış değişkenlerin skaler normuna göre uzayda yer alan simetrilerinin, üretici yolları daha basitçe tanımlamayı mümkün kıldığını gösterir. Bu şekilde, üstün performansla sağlamlığı(robustness) iyileştirmiştir. Ayrıca, WaveDiff [111] her girişi dört frekans alt bandı (LL, LH, HL, HH) olarak almış ve hem eğitim hem de çıkarım sürelerini azaltmıştır.

DiffuseVAE [70] koşullu üretim için VAE’leri kullanmış ve düşük boyutlu gizli bir alanda kontrol edilebilir sentezde başarılı olmuştur. Kullandıkları üretici-iyileştirici çerçevede bir koşullandırma sinyali (y) önce standart bir VAE ile modellenir. Daha sonra veriler koşullandırılmış bir DDPM kullanarak y ve y ’nin düşük boyutlu gizli kod gösterimiyle modellenmiştir.

4.3.1.6 Optimal Taşıma

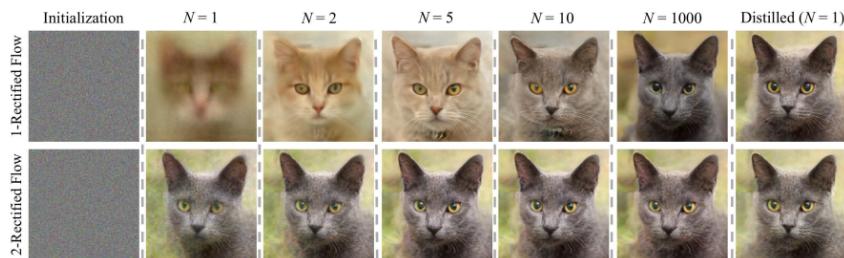
De Bortoli ve diğerleri [112], tarafından başlangıç ve bitiş durumları koşullu difüzyon süreçleri olan difüzyon köprüleri simüle edilmiştir. Ters dönüşüm,

varyasyonel bir skor eşleme formülasyonu kullanarak fonksiyonlarla öğrenilir. Yakın zamanda yapılan bir çalışmada Khrulkov ve diğerleri [113] DDPM kodlayıcı haritasının çok değişkenli normal dağılımlar için Monge optimal taşıma haritasıyla örtüştüğünü göstermiştir.

Su ve diğerleri [114], Çift Difüzyon Kapalı Köprülerini (DDIB), kaynak ile gizli temsil ve gizli temsil ile hedef arasındaki birleşim olarak tanımladılar. Bu, Schrodinger köprüleri entropi düzenlemeli optimal taşimanın bir biçimidir. Lee ve diğerleri [115] tarafından önerilen bir diğer yöntemde, öğrenilen üretken yörüngelerin yüksek eğriliğinin yavaş örneklemeye hızıyla ilişkili olduğunu iddia edilmektedir. Üretken yörüngelerin eğriliğini en aza indirmek ve rekabetçi performansla örneklemeye maliyetini azaltmak için ileri süreci eğitmeyi önermişlerdir.

α -karıştırma(blending) [116], doğrusal olarak interpolate edilen (karıştırın) ve karıştırmayı kaldırın örneklerin yinelemeli olarak iki yoğunluk arasında kesin bir eşleme ürettiğini ve bu eşlemenin, örneklerdeki karıştırmayı kaldırın minimalist bir difüzyon modelini eğitmek için kullanılabileceğini gösterdi.

Düzeltme akışı(Rectified flow) [117], Şekil 4.6'da gösterildiği gibi en küçük kareler optimizasyonyla düz giden en kısa yolları bulup bunları izleyerek gürültü dağılımını veri dağılımına taşıyan bir ODE modelidir. Yalnızca tek bir Euler ayrıştırma adımıyla yüksek kaliteli sonuçlar elde ederler.



Şekil 4.6 Düzeltme akısı [117] ($N \geq 2$) için görüntü oluşturma görevinde çok az sayıda adımla iyi örnekler üretir.

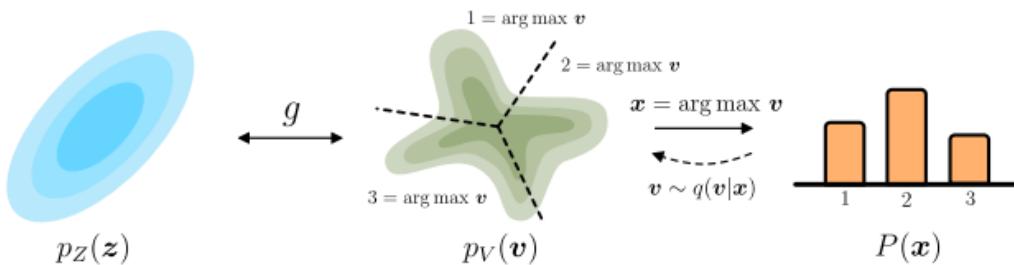
Stokastik İnterpolant [118], temel ve hedef dağılım çiftleri arasında dönüşüm yapan bir normalleştiren akış modelidir. Bu modelin hızı(velocity), veri skorunu ve ondan örnek tahmin eden bir difüzyon modelini oluşturmak için kullanılabilir. Ayrıca Akış Eşleştirme [119] de Gauss olasılık yollarında gürültüyü veri örneklerine dönüştürmek için uygundur.

4.3.1.7 Farklı Kipler

Difüzyon modelleri görüntü ve ses gibi veri alanlarında büyük başarılar elde etmiştir. Ancak, diğer kiplerle kullanıldığında dikkate alınması gereken belirli yapıları vardır. Örneğin, skor fonksiyonları sürekli alanda olduğunda veya veriler düşük boyutlu manifoldlarda olduğunda zorluklar çıkabilir. Bu bölümde bu tür zorlukların üstesinden gelmek için geliştirilen bazı yaklaşımalar tartışılacaktır. Bu yaklaşımalar 3 başlık altında incelenecaktır: Ayrık veri, değişmez-varyans yapıları ve manifold yapıları.

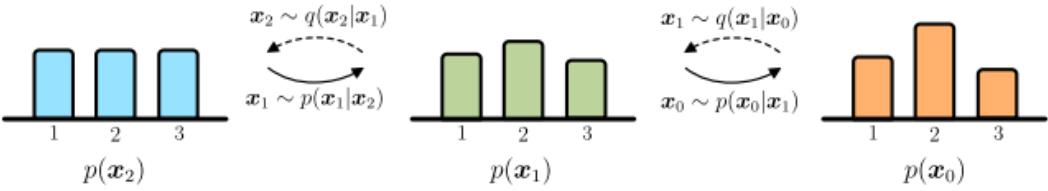
Ayrık veriler: Gauss gürültüsü ekleme ayrık veriler için uyumlu olmadığından, çoğu difüzyon modeli yalnızca sürekli veriler için uygulanabilir. Ayrıca, skor fonksiyonları yalnızca sürekli veriler için tanımlanmıştır. Ancak, makine öğrenimi problemlerinde cümleler, atomlar/moleküller ve vektörleştirilmiş veriler gibi ayrık verilerle çalışmak da gerekmektedir. Bu nedenle, difüzyon modelleriyle yüksek boyutlu ayrık veriler üreten bazı yöntemler geliştirilmiştir.

Hoogeboom ve diğerleri [120], Şekil 4.7 ve 4.8'de gösterildiği gibi kategorik veriler için akış tabanlı modellere ve difüzyon modellerine iki farklı eklenti sunar. Akışlardaki eklenti, sürekli bir dağılımin (normalleştirilen akış gibi) ve bir Argmax fonksiyonunun birleşimidir. Bu modeli optimize etmek için, kategorik verileri sürekli bir uzaya taşıyan Argmax'ın olasılıksal tersi öğrenilir. Öte yandan, difüzyon modellerine eklenti olarak difüzyon süreci boyunca kategorik gürültüyü kademeli olarak eklemişlerdir ve bu yönteme çok terimli(multinomial) difüzyon adını vermişlerdir. Önerilen yöntem, log-olasılık optimizasyonu açısından metin modelleme ve görüntü segmentasyon haritalarındaki mevcut yaklaşımardan daha iyi performans göstermiştir.



Şekil 4.7 Hoogeboom ve diğerleri[120] tarafından önerilen Argmax akışları

Ayrık gürültü arındıran olasılıksal difüzyon modelleri (D3PM) [121] çok terimli difüzyon modelini [120] genelleştirmiştir. D3PM, difüzyon modellerini ayrık verilere uyarlamak için ileri gürültü işlemeyi emici durum çekirdekleri veya ayıraltılmış Gauss çekirdekleri ile gerçekleştirmiştir.



Şekil 4.8 Hoogeboom ve diğerleri[120] tarafından önerilen Çok terimli(multinomial) difüzyon

Bu çalışma, difüzyon modelleri ile otoregresif/maske tabanlı üretken modeller arasında bir bağlantı kurmaktadır. Geçiş matrisinin seçiminin görüntü ve metin uzaylarındaki sonuçları iyileştiren önemli bir tasarım kararı olduğunu göstermektedir. Ayrıca varyasyonel alt sınırı çapraz entropi kaybıyla birleştiren yeni bir kayıp fonksiyonu sunmaktadır.

Otoregresif modellerin örnekleme süreci, özellikle yüksek boyutlu veriler için, ardışık olması gerekiğinden zaman alıcıdır. Otoregresif Difüzyon Modelleri (ARDM) [122], temsillerde nedensel maskeleme kullanmak yerine olasılıksal difüzyon modelleri kullanan bir hedefle eğitilmiştir. ARDM, test aşamasında paralel olarak veri üretebilir ve bu da koşullu üretim görevlerini gerçekleştirebilmesini sağlamaktadır.

Gu ve diğerleri [123] vektör nicelemeli(quantized) (VQ) veriler için ilk kez difüzyon modelleri kullandı. Bu çalışmada, VQ-VAE [124]'deki tek yönlü önyargı ve birikim hatası sorunları çözülmüştür. VQ-Diffusion [123], Gauss gürültüsü yerine ayrik veri uzayında rastgele yürüyüş veya maskeleme işlemi kullanır.

Analog Bitler [125] ayrik verileri ikili bitler olarak temsil eder ve bu bitleri sürekli bir difüzyon modeliyle modellemiştir. Üretim kalitesini önemli ölçüde iyileştirmek için iki yeni teknik (Kendi Kendini Koşullandırma ve Asimetrik Zaman Aralıkları) bildirmiştirlerdir.

Karartma(Blackout) Difüzyonu [126] Gauss ileri işlemlerini ayrik durum işlemlerine genelleştirmiştir ve gürültü yerine boş bir görüntüden örnekler üretmiştir.

Campbell ve diğerleri, [127] ayrik difüzyon modelleri için ilk sürekli zaman çerçevesini tanıtmıştır. Sürekli zamanlı Markov zincirlerini kullanarak ayrik emsallerinden daha iyi performans gösteren verimli örnekleyicilere sahiptirler.

DeğişmezInvariant Yapılar: Birçok alanda değişmez veri yapıları mevcuttur. Örneğin, graflar permütasyona izin vermez, nokta bulutları hem

öteleme(translation) hem de dönme(rotation) için uygun değildir. Difüzyon modellerinin gürültü ekleme süreci değişmez yapıları bozar. Bu nedenle, bu koşulları göz ardı etmek düşük performansla sonuçlanacaktır. Difüzyon modellerinde değişmez verilerle çalışan birkaç çalışma vardır.

Niu ve diğerleri [128] difüzyon modelleriyle değişmez graflar üretmektedir. Gürültü koşullu skor ağlarını parametrelendirmek için EDP-GNN adı verilen bir permütasyon eşdeğer graf sinir ağı [129] kullanırlar. Graf verileri, geleneksel ayrik skor eşleştirme boru hattına uygulanmadan önce bir bitişiklik(adjacency) matrisiyle işlenir. GDSS [130] graflar için sürekli zamanlı bir difüzyon süreci önererek bu fikri geliştirmiştir. Her iki düğümün ve kenarların ortak dağılımını SDE'lerle modellemişler ve permütasyon değişmezliğini garantilemek için mesaj ileme işlemlerini kullanmışlardır.

Xu ve diğerleri, [131], eşdeğişken(equivariant) Markov çekirdekleriyle eğitilen Markov zincirleriyle değişmez bir marginal dağılım elde etmiştir. Önceki dağılım ve geçiş çekirdekleri aynı değişmezliğe sahip olduğu sürece, difüzyon modellerinin dönme ve öteleme açısından değişmez moleküller yapılar üretebileceğini göstermişlerdir. Benzer şekilde, Shi ve diğerleri, [132] difüzyon modelleriyle hem öteleme hem de dönme açısından değişmez olan moleküller yapılar üretti.

Graf-GDP [133], SDE ileri işleminde bilinen bir kenarın olasılığını takip ederek, grafların karmaşık dağılımını rastgele graflara dönüştürür. NVDiff [134] yalnızca gizli uzaydaki düğüm vektörlerini modeller ve örneklemeye hızını önemli ölçüde iyileştirir. Graf Spektral Difüzyon Modeli (GSDM) [135], daha az hesaplama maliyetiyle kaliteli graf verileri üretmek için graf spektrum uzayında düşük-rütbeli(low-rank) difüzyon SDE'lerini kullanır.

Manifold Yapılar: Manifold yapılar, bakış açısını değiştirerek veri kaybetmeden boyut indirmeyi mümkün kılar. Manifold hipotezi [136], doğal verilerin azaltılmış boyutlardaki manifoldlarda olduğunu ileri sürer. Bu nedenle, bu manifoldları öğrenmek ve doğrudan bunlar üzerinde difüzyon modelleri eğitmek performansı artırabilir. Manifold yapılara sahip difüzyon modelleri geliştirmeye odaklanan bazı çalışmalar vardır.

Riemann Skor Tabanlı Üretken Model (RSGM) [137] difüzyon modellerinin kompakt Riemann manifoldlarına genişletilebileceğini gösterir. RSGM eğer hafif koşullar sağlanırsa küreler ve toruslar da dahil olmak üzere çok çeşitli manifoldları kapsayabilir. Ayrıca bir manifold üzerindeki difüzyon sürecini tersine çevirmek için bir Jeodezik Rastgele Yürüyüş kullanır.

Riemann Difüzyon Modeli (RDM) [138] sürekli zamanlı difüzyon modellerini Riemann manifoldlarına genelleştirir. RDM, log-olasılığın varyasyonel alt sınırını optimize eder. Bu ayrıca bir Riemann skor eşleştirme kaybını en aza indirmeye eşdeğerdir. İlgili Riemann manifoldunun yüksek boyutlu bir Öklid uzayına yerleştirildiğini varsayan RDM, RSGM'den daha geniş bir görüş aralığı sağlar.

Boomerang [139], bir giriş görüntüsünü bozarak gizli manifold uzayına taşır, ardından görüntü manifoldları üzerinde yerel örneklemeye yoluyla görüntü uzayına geri döner. Ayrıca, Cheng ve diğerleri [140], Riemann manifoldları üzerinde geometrik SDE'leri araştırmış ve ayrik Gauss olmayan stokastik süreçleri modellemek için geometrik SDE'leri kullanmayı önermiştir.

Difüzyon Olasılıksal Alanları (DPF) [141] difüzyon modellerinin formülasyonunu alanlara(fields) genişletir. Böylece aynı modelle farklı modaliteler, 2-boyutlu görüntüler ve 3-boyutlu nesnelerle çalışabilirler. Riemann Difüzyon Schrödinger Köprüsü [142] Difüzyon Schrödinger Köprüsü [112]'nü Öklid dışı ortama genelleştirmiştir. Ayrıca Park ve diğerleri [143], metin girdileri ve görüntü özellik haritalarını Riemann geometrisine uyarlayarak aralarındaki gizli semantik yönleri keşfetmiştir.

4.3.2 Eğitimden Bağımsız Gelişmeler

Bu bölümde öneğitimli modeller kullanarak doğrudan örneklemme algoritmasını iyileştiren yaklaşımları incelenecaktır. Bunlar sırasıyla SDE çözüçüler, ODE çözüçüler, yeni örneklemme prosedürleri, ince ayarlama ve bilgi damıtımı ile ilgili gelişmelerdir.

Genel olarak iki temel diferansiyel denklem formülasyonu vardır: SDE formülasyonu gürültü alanında rastgele bir doğrultuda yürürlükte, ODE formülasyonu ise deterministik olduğundan yüksek hızda sahiptir. Yüksek mertebeden çözüçüler daha yüksek derecede yakınsama ve daha düşük tahmin hatalarına sahiptir ancak daha fazla hesaplama maliyeti ve istikrarsızlık sorunları getirir.

4.3.2.1 SDE Çözüçüler

Song ve diğerleri, [98], Score-SDE [30]'leri eğitmek için kullanılan hedefin, özel bir ağırlıklandırma fonksiyonu (olasılık ağırlıklandırması) ile üretilen dağılımin beklenen değerini maksimize ettiğini kanıtlamıştır. SDE tarafından üretilen dağılım p_{θ}^{SDE} ile gösterilirse, hedef fonksiyonu Eşitlik 4.35'te olduğu gibi verilebilir.

Burada $\mathcal{L}(\theta; g(\cdot)^2)$, Eşitlik 4.29'da verilen hedef fonksiyonudur ve $\lambda(t) = g(t)^2$.

$$D_{KL}(q_0||p_\theta^{\text{SDE}}) \leq \mathcal{L}(\theta; g(.)^2) + D_{KL}(q_t||\pi) \quad (4.35)$$

$D_{KL}(q_0||p_\theta^{\text{SDE}}) = -E_{q_0} \log(p_\theta^{\text{SDE}}) + \text{const}$ ve $D_{KL}(q_t||\pi)$ bir sabittir. $\mathcal{L}(\theta; g(.)^2)$ eğitimi, $-E_{q_0} \log(p_\theta^{\text{SDE}})$ verilerindeki negatif log-olasılığı en aza indirir. Eşitlik 4.36, $p_\theta^{\text{SDE}}(x)$ için uygulanan sınırı gösterir[44, 98].

$$-\log(p_\theta^{\text{SDE}}(x)) \leq \mathcal{L}'(x) \quad (4.36)$$

Burada $\mathcal{L}'(x)$ Eşitlik 4.37'de olduğu gibi hesaplanır. Bu kaybin ilk kısmı örtük skor eşleştirme [144]'ı hatırlatır. Eşitlik 4.37 Monte Carlo yöntemleriyle verimli bir şekilde tahmin edilebilir.

$$\begin{aligned} \mathcal{L}'(x) = & \int_0^T E \left[\frac{1}{2} \|g(t)s_\theta(x_t, t)\|^2 + \nabla \cdot (g(t)^2 s_\theta(x_t, t) - f(x_t, t)) \middle| x_0 = x \right] dt \\ & - E_{x_T} [\log p_\theta^{\text{SDE}}(x_T) | x_0 = x] \end{aligned} \quad (4.37)$$

Tutarlı Tavlanmış Örneklemme (CAS) yöntemi [74], Tavlanmış Langevin Örneklemesine daha kararlı bir alternatif olarak önerildi. Gürültü arındırma adımlarını ölçekleyen bir skor tabanlı MCMC yaklaşımıdır. Bu yöntem, yüksek kaliteli görüntüler üretirken FID puanını iyileştirmeye katkıda bulunur.

Bazı yöntemler, hem doğrusal çözüçüleri hem de yüksek dereceli çözüçüleri kullanarak örneklemme sürecini iyileştirir. Gotta-go-fast [145], uyarlanabilir adım boyutlarına sahip bir SDE çözümüne sahiptir. Örneklemme sürecini hızlandırmak için doğrusal bir çözümü (Euler-Maruyama Yöntemi) yüksek dereceli bir çözümü (Geliştirilmiş Euler Yöntemi) ile birleştirilir. Yüksek ve düşük dereceli çözüçüler, her zaman adımda önceki x_{prev} örneğinden yeni x_{high} ve x_{low} örnekleri üretir. Daha sonra adım boyutu, bu iki örnek arasındaki fark karşılaştırılarak ayarlanır. x_{high} ve x_{low} benzerse, algoritma x_{high} değerini döndürür ve adım boyutunu artırır. x_{high} ve x_{low} arasındaki fark Eşitlik 4.38' de verilmiştir. Burada $\delta(x, x_{prev}) := \max(\epsilon_{abs} - \epsilon_{rel} \max(|x|, |x_{prev}|))$, ϵ_{abs} ve ϵ_{rel} mutlak ve göreli toleranslardır. Gotta-go-fast, modeli ayarlamaya gerek kalmadan daha iyi veya eşit örnek kalitesi elde ederken verileri 2 ila 10 kat daha hızlı üretir.

$$E_q = \left\| \frac{x_{high} - x_{low}}{\delta(x, x_{prev})} \right\|^2 \quad (4.38)$$

Yüksek mertebeden türevler, veri dağılımı hakkında ek yerel bilgi sağlar ve yeni uygulamalara olanak tanır. Meng ve diğerleri, [146] gürültü arındırma skoru eşleştirmesini yüksek mertebeden türevlere uygulayarak genelleştirmiştir. Veri skorlarını doğrudan tahmin etmek için yüksek mertebeden momentlerde Tweedie formülünden yararlanarak bir yöntem önermişlerdir. Önerilen yöntem kullanılarak ikinci mertebeden skorların yaklaşık olarak tahmin edilmesinin otomatik türev uygulamaktan daha iyi sonuçlar verdiği gösterilmiştir.

Genişletilmiş bir uzayda ters-SDE’yi elde eden Kritik sökümlenmiş Langevin Difüzyonu (CLD)[100] Bölüm 4.3.1.5’té incelenmiştir. Bu yöntem ayrıca Euler-Maruyama örnekleyicisini önemli ölçüde geride bırakan Simetrik Bölmeli CLD Örnekleyici (SSCS) adlı yeni bir üretken SDE geliştirmiştir.

Verma ve diğerleri [147], sürekli üstel SDE’leri kullanarak Varyasyonel Gauss süreçleri için alternatif bir parametrelendirme sağlamış ve dışbükey optimizasyon için sabit nokta yinelemelerine sahip hızlı bir algoritma elde etmiştir.

Taylor genişlemesine dayalı daha yüksek dereceli algoritmalarla, skor fonksiyonunun türevini tahmin etmek, büyük ölçekli, iyi eğitilmiş sinir ağlarının karmaşıklığı nedeniyle çözülemez hale gelmektedir. Skor-integrand Çözücü (SciRE-Solver) [148] skor fonksiyonunun türevini hesaplamak için özyinelemeli fark (RD) yöntemini tanıtmıştır. Önerilen yöntem, mevcut eğitimsiz örneklemeye algoritmalarına karşı en son FID’leri elde etmiştir.

4.3.2.2 ODE Çözüçüler

Olasılıksal akışlı ODE, sınırsız ODE’lerin veya sürekli normalleştirici akışların özel bir durumu olduğundan, ODE p_θ^{ODE} tarafından üretilen dağılım, Eşitlik 4.39’daki yaklaşımalar üzerinden hesaplanır.

$$\log p_\theta^{\text{ODE}}(x_0) = \log p_T(x_T) + \int_0^T \nabla \cdot \bar{f}_\theta(x_t, t) dt \quad (4.39)$$

Buradaki tek integral, sayısal ODE çözüçüler ve Skilling-Hutchinson iz(trace) tahmincisi [149, 150] ile hesaplanabilir. Ancak, bu formül p_θ^{ODE} ,yi maksimize etmek için doğrudan optimize edilemez çünkü her veri noktası x_0 için pahalı ODE çözüçüleri çağırmayı gerektirir.

Kapalı(implicit) örnekleyiciler, difüzyon modelinin yeniden eğitimesini gerektirmeyen adım atlamlı bir örnekleyici sınıfıdır. Difüzyon sürecindeki adım sayısı genellikle örneklemeye sürecindeki adım sayısına eşittir. Ancak difüzyon

ve örnekleme süreçleri ayrı olduğundan bu bir gereklilik değildir. Song ve diğerleri DDIM’lerde [56] deterministik ileri süreç ve adım atlamalı örnekleme kullanmışlardır. Markovian olmayan difüzyon süreçleri tasarlayarak, deterministik süreçlerle çok daha hızlı yüksek kaliteli örnekler üretebilmesini sağlamışlardır.

x_t örneğinden x_{t-1} örneğinin elde edilmesi Eşitlik 4.40’ta verilmiştir. Burada $\epsilon_t \sim \mathcal{N}(0, I)$, x_t ’den bağımsız standart Gauss gürültüsüdür ve $a_0 := 1$ olarak tanımlanır. σ değerlerinin farklı seçimleri farklı üretken süreçlerle sonuçlanır, bu üretken süreçlerin hepsi aynı ϵ_θ desenini kullanır, bu nedenle yeniden eğitim gerekli değildir. $\sigma_t = \sqrt{\frac{1-a_{t-1}}{1-a_t}} \sqrt{1 - \frac{a_t}{a_{t-1}}}$ ise ileri süreç Markovian olur ve model DDPM olur. Tüm t için $\sigma_t = 0$ durumunda, x_{t-1} ve x_0 bilindiğinden, süreç $t = 1$ hariç deterministik hale gelir. Üretken süreçteki rastgele gürültü ϵ_t önündeki katsayı sıfır olur ve örnekler örtük olasılıksal bir modelde sabit bir prosedürle gizli değişkenlerden üretilir.

$$x_{t-1} = \sqrt{a_{t-1}} \left(\frac{x_t - \sqrt{1-a_t} \epsilon_\theta^{(t)}(x_t)}{\sqrt{a_t}} \right) + \sqrt{1-a_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(2)}(x_t) + \sigma_t \epsilon_t \quad (4.40)$$

DDIM yinelemesi Eşitlik 4.41’de olduğu gibi yeniden yazıldığında, ODE’leri çözmek için Euler integraline benzerdir. Burada, $\frac{\sqrt{1-a}}{\sqrt{a}}$, σ_t ve $\frac{x}{\sqrt{a}}$ \bar{x} ile parametrelendirilerek karşılık gelen ODE türetilir. Sürekli durumda σ ve \bar{x} , t ve $\sigma(0) = 0$ ’ın fonksiyonlarıdır.

$$\frac{x_{t-\Delta t}}{\sqrt{a_{t-\Delta t}}} = \frac{x_t}{\sqrt{a_t}} + \left(\sqrt{\frac{1-a_{t-\Delta t}}{a_{t-\Delta t}}} - \sqrt{\frac{1-a_t}{a_t}} \right) \epsilon_\theta^{(t)}(x_t) \quad (4.41)$$

Eşitlik 4.41, Eşitlik 4.42’de verilen ODE’nin Euler metodudur. Başlangıç koşulları $x(T) = \mathcal{N}(0, \sigma(T))$ çok büyük σ_T için $a \approx 0$ ’a karşılık gelir. Bu, üretken sürecin yeterli ayrıklaştırma adımlarıyla tersine çevrilebileceğini gösterir. x_0 ’ı x_T ’ye kodlamak, Eşitlik 4.42’deki ters-ODE’yi simüle eder.

$$d\bar{x}(t) = \epsilon_\theta^{(t)} \left(\frac{\bar{x}(t)}{\sqrt{\sigma^2 + 1}} \right) d\sigma_t \quad (4.42)$$

Yazarlar ayrıca, DDIM örnekleme sürecinin olasılıksal akışlı ODE’nin özel bir ayrıklaştırması olan VE-SDE’ye karşılık geldiğini ima eden eş zamanlı çalışma [30]’a da atıfta bulunmaktadır.

gDDIM [151], tüm örtük örneklemeye difüzyon modellerini çeşitli çekirdek tiplerine sahip bir DDIM ailesine genelleştirmiştir. Ayrıca, Difüzyon Üstel Entegrasyon Örnekleyicisi (DEIS) [152], daha verimli özelleştirilmiş ODE çözümleri geliştirmek için olasılıksal akışlı ODE'nin yarı doğrusal yapısından yararlanır. Olasılıksal akışlı ODE'nin doğrusal kısmı analitik olarak hesaplanır ve doğrusal olmayan kısmı üstel entegratörlerle çözülür. DEIS, birinci dereceden tahmin edici olarak DDIM[56]'i kullanmış, ardından yüksek kaliteli örnekler yüksek dereceli integraller kullanılarak yalnızca 10-20 yinelemede üretilmiştir.

EDM [153] Tahmin edici-düzeltici (Predictor-Corrector) örnekleyicinin [30] düzelticisini geliştirir ve örneklemeye sürecinde bir "çalkalama(churn)" adımı önerir. EDM, ikinci dereceden Heun çözümüyle [154] deterministik difüzyon ODE'nin zaman adımlarını kullanır. Heun'un yöntemi, örnek kalitesi ve örneklemeye hızı arasında mükemmel bir denge sağlar ve daha az örneklemeye adımıyla rekabetçi örnekler üretir.

Ayrıca, DPM-Solver [155] farklı derecelerde çözümleri kullanır. DPM Solver-Fast, alternatif çapraz sıralarda birleştirilmiş çözümleri daha iyi performans elde etmiştir. Ayrıca Pang ve diğerleri [156] öneğitimli bir DPM-çözücü modelini kalibre etmek için yeni bir yöntem önermiş ve ayrık zamanlı modellerde daha iyi sonuçlar elde etmiştir.

PNDM [157], manifold uzayının yardımıyla örneklemeye sırasında diferansiyel denklemi çözer. Diferansiyel denklem örnekleyicisi için genelleştirilmiş bir sürüm sağlar. PNDM, farklı sayısal çözümlerin aynı eğimi paylaşabileceğini varsayıarak R^N 'deki belirli bir manifolddan örneklemeye yapmak için sözde sayısal bir yöntem sahiptir. Üç adımlı yüksek seviyeli bir çözümü (Runge-Kutta yöntemi) ve ardından örneklemeye için çok adımlı doğrusal bir yöntem kullanılır.

ODE tabanlı örnekleyiciler hızlıdır, SDE tabanlı örnekleyiciler ise yüksek kaliteli örnekler üretir. Yeniden başlatma(Restart) [158] ek ileri adımlarda tamamıyla gürültü ekleme ve geri ODE'yi sıkı bir şekilde takip etme arasında geçiş yapar. Yeniden başlatma, hem hesaplama süresi hem de kalite açısından önceki örnekleyicilerden daha iyi performans gösterir.

Cao ve diğerleri [159], SDE veya ODE tabanlı difüzyon modellerini seçmek için araştırma yapmış ve veri dağılımını en sonuna kadar bozuklarında ODE modelinin SDE modelinden daha iyi performans gösterdiğini, ancak bozmayı daha erken durduruklarında SDE modelinin ODE modelinden daha iyi performans gösterdiğini bulmuşlardır.

4.3.2.3 Yeni Örnekleme Prosedürleri

Geleneksel örnekleme prosedürüni değiştiren ve modeli yeniden eğitmeden daha iyi sonuçlar elde eden birçok yaklaşım vardır.

Verimli Örnekleme [160], öneğitimli bir difüzyon modeli verildiğinde hedefi dinamik programlama için yeniden düzenler. Bu yöntemde örnekleme sürecinde uygun zaman planlamalarını seçmek bir optimizasyon problemdir. ELBO’yu ayrı ayrı KL terimlerine ayırmış ve ELBO’yu maksimize eden örnekleme yörungesini bulmuşlardır. Yazarlar ELBO optimizasyonunun FID puanlarıyla eşleşmediğini doğrulamış ve örnekleme yörungesini optimize etmek için başka bir yol keşfetmeyi önermişlerdir. Aynı yazarların daha sonraki bir çalışmasında, Genelleştirilmiş Gauss Difüzyon Modelleri (GGDM) [161], Diferansiyel Difüzyon Örnekleyici Araması (DDSS) kullanarak doğrudan Çekirdek Başlangıç Mesafesini (KID) [162] optimize eder. Bu çalışma, Markov olmayan örnekleyiciler ve geniş bir marginal varyans aralığına sahip difüzyon modellerini genelleştirir. Bu yöntem, örnekleme sürecinde geri difüzyon için yeniden parametrelendirme(reparametrization) ve gradyanın yeniden materyalizasyonu (rematerialization) hilelerini kullanır ve hesaplama süresi yerine bellek maliyeti getirir.

Analitik-DDPM [163] her adımda ters ortalamaya dayanarak ters kovaryansı bulur. Eşitlik 4.43 analitik formda öneğitimli bir skor modelinden optimum ters kovaryansı elde etmeyi gösterir. Öneğitimli bir skor modeli verildiğinde, optimum ters kovaryansları elde etmek için birinci ve ikinci dereceden momentleri tahmin etmek ve bunları hedef fonksiyonunda kullanmak daha sıkı varyasyonel alt sınır ve daha yüksek olasılık değerleriyle sonuçlanır. Aynı yazarların daha sonraki bir çalışması [164] öneğitimli DDPM modelleri üzerinde başka bir ağ eğıterek kovaryansı tahmin etmeyi önermektedir. Bu çalışmaların her ikisinde de uygulamalar hem DDPM hem de DDIM modelleri için gerçekleştirilmiştir.

$$\Sigma_{\theta}(x_t, t) = \sigma_t^2 + \left(\sqrt{\frac{\bar{\beta}_t}{a_t}} - \sqrt{\bar{\beta}_{t-1} - \sigma_t^2} \right)^2 \cdot \left(1 - \bar{\beta}_t E_{q_t(x_t)} \frac{\|\nabla_{x_t} \log q_t(x_t)\|^2}{d} \right) \quad (4.43)$$

Eğitim süreci her zaman gerçek örnekleri kullanırken, çıkarım süreci daha önce oluşturulmuş gürültülü örneği kullanır. Gerçek ve gürültülü örnekler gürültü tahmin ağına verildiğinde tutarsızlığa neden olur, bu da hata birikimine ve örnekleme kaymasına yol açar. Maruz kalma önyargısı(Exposure bias) sorunu, eğitim ve çıkarım süreçlerindeki girdilerin uyumsuzluğudur. Ning ve diğerleri [165], pratik örnekleme dağılımının her bir adımda temel gerçek dağılımindan daha

büyük bir varyansa sahip olduğunu göstermiştir. İki dağılım arasındaki varyans farkını değerlendirmek için bir metrik kullanmış ve Epsilon Ölçeklemesini (ES) önermişlerdir.

Difüzyon modellerinin yaptığı iş, yüksek boyutlu bir gizli uzaydaki noktaları, bilinen düşük boyutlu bir manifolda, tipik olarak bir görüntü manifolduna eşlemek olarak görülebilir. Boomerang [139] bir giriş görüntüsüne bozucu etki yaparak gizli manifold uzayına taşıır, ardından görüntü manifoldları üzerinde yerel örneklemeye yoluyla görüntü uzayına geri döner. Boomerang, difüzyon modellerinin eğitiminde herhangi bir değişiklik gerektirmez ve pahalı olmayan tek bir GPU'da öneğitimli modellerle kullanılabilir.

Zheng ve diğerleri [166], Gauss gürültü dağılımını sürekli zamanlı çözümdeki ters difüzyon sürecinin yönigesine eşleyen sınır operatörü (DSNO) yöntemi ile difüzyon modeli örneklemesini önermiştir. Olasılıksal akışlı ODE'yi çözmek için sınır operatörlerini kullanmış ve en son teknoloji FID puanlarına ulaşmışlardır.

Zaman Kaydırma Örnekleyicisi [167], bir sonraki adımı mevcut örneklerin varyansına göre ayarlayarak maruz kalma önyargısı sorununu hafifletir. Ayrıca Diff-Pruning [168], öneğitimli modelleri yeniden eğitmeden hafif difüzyon modelleri elde etmek için etkili bir sıkıştırma yöntemidir.

Ayırt edici Rehberlik [169], üretilen örneğin gerçekçi olup olmadığını tahmin eden açık bir ayırt edici(discriminator) ağından yararlanarak öneğitimli modellerin üretim performansını geliştirmiştir. Ayrıca, Kawar ve diğerleri [170], difüzyon tabanlı görüntü sentezini yönlendirmek için dayanıklı çekişmeli sınıflandırıcıların gradyanlarını kullanmıştır.

FSDM [171] koşullu difüzyon modelleri için az atışlı(few-shot) örneklemeye çerçevesidir. FSDM farklı örneklemeye süreçlerine uyum sağlayıp görüntü dönüştürücülerinin gizli uzayını kullanarak birkaç adımda iyi bir örneklemeye performansı elde etmiştir.

Sehwag ve diğerleri [172], örneklemeye sürecini düşük yoğunluklu bölgelere doğru yönlendirir ve oradan yeni yüksek doğruluklu örnekler üretir. Ayrıca, Öz Dikkat Rehberliği [173] modelin görüntüdeki dikkat verdiği bölgeleri bulanıklaştırır ve buna göre kalan bilgiyle modeli yönlendirir.

Gürültü Giderme Difüzyon Geri Yükleme Modelleri (DDRM) [174] süper çözünürlük, bulanıklık giderme, boyama ve renklendirme görevlerinde kaliteyi ve çalışma süresini iyileştiren bir örneklemeye yöntemi önermektedir.

4.3.2.4 İnce Ayar

ParaDIGMS [175] örnekleme adımlarının çözümünü tahmin ederek ve yakınsayana kadar yinelemeli olarak ayarlayarak örnekleme sürecini paralel hale getirir. ParaDIGMS, hesaplama maliyetini hız için takas eden ilk örnekleme yöntemidir ve literatürde mevcut olan diğer hızlı örnekleme tekniklerine de uyarlanabilir.

Parametre Verimli Ayarlama [176], kritik faktörün adaptörlerin giriş konumu olduğunu ve bunun aşağı akış(downstream) görevlerinin performansını etkilediğini ileri sürmektedir. En iyi sonuçlar, giriş bloğunu çapraz dikkat bloğundan sonra koyarak elde edilmiştir. Aiello ve diğerleri [177], öğrenilen dağılımı belirli bir zaman adımı bütçesiyle ince ayarlamak için Maksimum Ortalama Sapmayı (MMD) en aza indirmiştir. Bu, hız-kalite dengesinde önemli bir iyileştirme sağlamıştır. Yinelemeli Gizli Değişken İyileştirme (ILVR) [178], üretken süreci belirli bir referans görüntüyle yönlendirir. Ayrıca, metinlere düşük çözünürlüklü bir görüntü sağlayarak anlamsal uyum sorununa bir çözüm bulmuşlardır.

Koşullu üretim bağlamında, Dhariwal & Nichol [179] örnek kalitesini artırmak için eğitim sonrası bir yöntem olarak sınıflandırıcı rehberliğini önerdi. Öneğitimli difüzyon modelinin skorunu, koşulsuz bir modeli koşullandırmak için başka bir görüntü sınıflandırıcısının gradyanıyla birleştirmiştir.

Graikos ve diğerleri [180], öneğitimli koşulsuz bir modeli, gürültü arındırma ağının gizli temsillerini kullanarak koşullara uyarlar. Difüzyon Gizli Temsillerinin Doğrudan Optimizasyonu (DOODL) [181] geri difüzyonu bellek açısından verimli hale getirmek için öneğitimli bir sınıflandırıcının gradyanlarını, gerçek üretilen piksellerde optimize etmeyi önerir. Shen ve diğerleri [182] koşulsuz difüzyon modeli ve sınırlı açık(explicit) rehberlik kullanarak koşullu örnekleme sorununu ele almıştır. Önermiş oldukları Model Tahmini Kontrolü (MPC), ek zaman adımlarında açık rehberliği geri yayarak difüzyon sürecini yönlendirir.

4.3.2.5 Bilgi Damıtımı

Bilgi damıtımı(Knowledge distillation) [85], yüksek öğrenme kapasitesine sahip karmaşık öğretmen modellerinden basit öğrenci modellerine "bilgi" aktararak verimli küçük ölçekli ağlar elde etmek için ortaya çıkan bir yöntemdir [183]. Bu sayede, öğrenci modelleri sıkıştırılmış ve hızlandırılmış modeller olarak avantajlara sahiptir. Difüzyon modelleri bağlamında, bilgi damıtımı, küçük ölçekli bir öğrenci modeli kullanarak örnekleme sürecini hızlandırmak için kullanılmıştır.

Salimans ve diğerleri, Progressive Distillation [184]'da bir örnekleme modelinden

diğerine kademeli olarak bilgi damıtmayı uyguladı. Her damıtma adımında, öğrenci modelleri eğitimden önce öğretmen modelleri tarafından yeniden ağırlıklandırılır, böylece tek adımda öğretmen modelleri kadar yakın örnekler üretebilirler. Sonuç olarak, öğrenci modelleri her damıtma sürecinde örnekleme adımlarını yarıya indirir. DDPM’lerle aynı eğitim hedefi ile sadece dört adımda rekabetçi örnekleme başarıları elde etmişlerdir.

Meng ve diğerleri [185], Stable Diffusion [102], DALL-E 2[186] ve Imagen [187] gibi sınıflandırıcı içermeyen rehberli difüzyon modellerinin hesaplama maliyetine odaklanmıştır. Çünkü bunlar sınıf koşullu bir model ve koşulsuz bir modeli tekrar tekrar değerlendirmeyi gerektirmektedir. Öneğitimli sınıflandırıcı içermeyen rehberli bir model verildiğinde, önce koşullu ve koşulsuz modellerin çıktısını eşleştirip ardından giderek daha az örnekleme adımı gerektiren başka bir modele bilgi damıtmışlardır.

Berthelot ve diğerleri, ikili zaman damıtımını genişleten ve FID hesaplama hızını 2.4 kata kadar artıran Geçişli Kapanış Zamanı Damıtma (TRACT) [188]’yı tanıttı. TRACT, bir öğrenci modelini, bir öğretmen modelinin çıktısını damıtmak üzere eğitir ve damıtmanın sonunda, öğrenci modeliyle tek adımlı çıkarım yapılabilir.

Sun ve diğerleri, Sınıflandırıcı Tabanlı Özellik Damıtma (CFD) [189]’yı önermiştir. Çıktı görüntülerini veri kümesinden bağımsız bir sınıflandırıcı ile hizalamak yerine öğretmenin keskinleştirilmiş özellik dağılımı ile damıtırlar. Bu, öğrencinin önemli özelliklere odaklanması ve yüksek kaliteli ve hızlı örnekleme elde etmesini sağlar.

Poole ve diğerleri Dreamfusion makalesinde skor damıtma örneklemesini (SDS) önermiştir [190]. Bu yöntemde metinden 3-boyutlu sentezlemeyi gerçekleştirmek için öneğitimli 2-boyutlu metinden görüntüye difüzyon modelini damıtmışlardır. Bu şekilde, 3-boyutlu eğitim verisi ve değişiklik gerektirmeden öneğitimli görüntü difüzyon modellerinin 3-boyutlu üreticiler olabildiğini göstermişlerdir.

Mevcut birçok damıtma tekniği, öğretmen modelinden sentetik veri üretirken hesaplama maliyetine neden olur veya pahalı çevrimiçi öğrenmeye ihtiyaç duyar. BOOT [191], verilen zaman adımlarında bir öğretmen modelinin çıktısını tahmin eden verimli bir damıtma algoritmasıdır. Temel fikir, verilen herhangi bir zaman adımdında öneğitimli bir difüzyon modeli öğretmeninin çıktısını tahmin eden zaman koşullu bir model öğrenmektir. Eğitim setlerinin genellikle büyük ve erişilmesi zor olduğu gerçeğine sahip alanlarda katkıları vardır.

Tutarlılık(Consistency) modeli (CM) [192] öneğitimli difüzyon modellerini damıtarak tek adımlı üretimle gürültüyü doğrudan veriye eşlemeyi sağlar. Ayrıca,

daha yüksek örnek kalitesi elde etmek için hesaplama maliyeti karşılanabiliyorsa örneklemme adımlarının uzatılmasına da olanak tanır. Ayrıca Tutarlılık Yörunge Modelleri (CTM) [193] Tutarlılık modellerini ve Skor tabanlı modelleri genelleştirmiştir ve ODE çözümünde uzun atlamalar içeren yeni bir örneklemme şeması önermiştir. CTM'de CM'den farklı olarak, yerleşik koşullu üretim yöntemlerinin benimsenmesi için skor fonksiyonuna erişim bulunmaktadır.

4.3.3 Gelişmelerin Değerlendirilmesi

4.3.3.1 Değerlendirme Ölçütleri

Bu bölümde, üretilen örneklerin kalitesini ve çeşitliliğini değerlendirmek ve modelleri birbirleriyle karşılaştırmak amacıyla en sık kullanılan değerlendirme ölçütleri ele alınacaktır.

4.3.3.2 Başlangıç Puanı (IS)

Başlangıç(Inception) Puanı [194], ImageNet veri kümesi [195] ile öneğitimli bir Inception v3 ağının [196] kullanılarak hesaplanır. Bu hesaplama çeşitlilik ölçümü ve kalite ölçümü olarak iki bölüme ayrılabilir. Üretilen bir örnek, ImageNet veri kümesindeki karşılık gelen sınıfına yakın olmasına bakılarak yüksek çözünürlüğe sahip olduğu kabul edilir. Bu nedenle, kalite ölçümü için sınıf görüntüleri ile örnek arasındaki benzerlik hesaplanır. Diğer yandan çeşitlilik ölçümü, üretilen örneklerin sınıf entropisine göre hesaplanır. Daha büyük entropi, örneklerin daha çeşitli olduğu anlamına gelir. Daha yüksek kalite ve daha çeşitli örnek üretimini gösteren Başlangıç Puanının hesaplanması için KL-ıraksaması, Eşitlik 4.44'te olduğu gibi uygulanır. Burada p_{gen} üretilen dağılım, p_{dis} ise modelin bildiği dağılımdır.

$$IS := \exp\left(E_{x \sim p_{gen}} \left[D_{KL}\left(p_{dis}(\cdot|x) \parallel \int p_{dis}(\cdot|x)p_{gen}(x)dx\right)\right]\right) \quad (4.44)$$

4.3.3.3 Frechet Başlangıç Mesafesi (FID)

Başlangıç Puanı mantıklı bir değerlendirme teknigi olmasına rağmen, 1000 sınıfı sahip bir veri kümesine ve bu verilerle eğitilmiş rastgelelik içeren bir ağa dayanmaktadır. Frechet Başlangıç Mesafesi (FID) [197] önceden belirlenmiş referans sınıflarının önyargısını çözmek için önerilmiştir. FID'de, gerçek veri dağılımı ile oluşturulan örnekler arasındaki mesafe ortalama ve kovaryans kullanılarak hesaplanır. Başka bir deyişle, Inception v3 ağının en son sınıflandırma katmanı kullanılmaz. Frechet Inception Mesafesi Eşitlik 4.45'te gösterilir. Burada,

μ_g ve Σ_g üretilen dağılımin ortalamasını ve kovaryansını gösterir, μ_r ve Σ_r modellenen bilinen dağılımin ortalamasını ve kovaryansını gösterir.

$$FID := \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}\right) \quad (4.45)$$

4.3.3.4 Negatif Log-olasılık (NLL)

Negatif log-olasılık (NLL), her türlü veri dağılımı için kullanılabilen yaygın bir değerlendirme ölçütüdür. Tahminlerimizin ne kadar iyi olduğunu belirlemek için doğru etiketlerin sınıf olasılıkları dikkate alınır. Logaritma hesaplamasının amacı, çarpımları toplamaya dönüştürerek hesaplama maliyetini azaltmaktadır. VAE ile yapılan birçok çalışma, NLL’yi bir değerlendirme ölçütü olarak kullanır[198]. Difüzyon modelleri de verilerin dağılımını açıkça modelleyebildiğinden, NLL hesaplaması mümkündür. Hatta Geliştirilmiş DDPM [54] gibi bazı difüzyon modelleri NLL’yi bir eğitim hedefi olarak kabul eder. Model p_θ ile gösterilirse, negatif log-olasılık hesaplaması Eşitlik 4.46’daki gibi ifade edilebilir.

$$NLL := E[-\log p_\theta(x)] \quad (4.46)$$

4.3.3.5 Kıyaslama Veri Kümeleri

Bu bölümde en sık kullanılan veri kümelerinde kalite testlerinin (benchmark) sonuçları verilmektedir. Sonuçlar FID(daha düşük daha iyidir), IS(daha yüksek daha iyidir) ve NLL(daha düşük daha iyidir) şeklinde verilmiştir.

CIFAR-10. CIFAR-10 veri kümesi (Kanada İleri Araştırma Enstitüsü, 10 sınıfı) Tiny Images [199] veri kümelerinin bir alt kümesidir ve 60000 adet 32x32 piksel renkli görüntüden oluşur. Her sınıfta 5000 eğitim ve 1000 test görüntüsü olmak üzere toplam 6000 görüntü vardır. Tablo 4.5, CIFAR-10’un koşulsuz görüntü oluşturma görevinde FID’ye göre sıralanmış sonuçlarını göstermektedir. EDM [153] CIFAR-10 görevinde üstün performans göstermiştir ve bu nedenle en son teknoloji yöntemlerin çoğu bu modele dayanmaktadır. Ayrıca bu tablo bize eğitimden bağımsız yaklaşımın daha güçlü olduğunu gösterir. Çünkü hem en iyi eğitim yöntemini hem de örneklem stratejilerini kullanabilmektedirler. Eğitim tabanlı yaklaşımalar arasında ise Uzay projeksiyon yöntemleri performanslarıyla öne çıkmaktadır.

Tablo 4.5 CIFAR-10 görüntüyü üretimi

Model	FID	IS	NLL
SciRE-Çözücü-EDM [148]	1.76	-	-
EDM-G++ [169]	1.77	-	2.55
EDM-ES [165]	1.8	-	-
CTM [193]	1.87	-	2.43
STF [80]	1.90	-	-
LSGM-G++ (FID) [169]	1.94	-	3.42
EDM [153]	1.97	-	-
PSLD (ODE) [104]	2.10	9.93	-
LSGM (FID) [101]	2.10	-	3.43
ADM-ES [165]	2.17	-	-
Alt uzay Difüzyonu (NSCN++) [103]	2.17	9.94	-
LSGM (dengeli) [101]	2.17	-	2.95
NCSN++ [30]	2.20	9.89	3.45
PSLD (SDE) [104]	2.21	-	-
CLD-SGM (EM-QS) [100]	2.23	-	-
CLD-SGM (Olas. akışlı) [100]	2.25	-	3.31
gDDIM [151]	2.28	-	-
INDM (FID) [97]	2.28	-	3.09
Yumuşak Kesme UNCSN++ (RVE) [71]	2.33	10.11	3.04
PFGM++ [110]	2.35	9.68	3.19
BDDM [59]	2.38	-	-
Alt uzay Difüzyonu (DDPM++) [103]	2.40	9.66	-
SciRE-Çözücü-VP(sürekli) [148]	2.40	-	-
DDPM++ [30]	2.41	9.68	-
Alt-VP-SDE [30]	2.41	9.83	2.99
Gotta Go Fast VP-deep [145]	2.44	9.61	-
Yumuşak Kesme DDPM++ (VP, FID) [71]	2.47	9.78	2.91
DEIS [152]	2.55	-	-
Progressive Distillation [184]	2.57	-	-
DPM-Çözücü [155]	2.59	-	-
DiffuseVAE-72M [70]	2.62	9.75	-
TDPM [67]	2.83	-	-
FastDPM [55]	2.86	-	-
Gotta Go Fast VE [145]	2.87	9.57	-
iDDPM (FID) [54]	2.90	-	3.37
Tutarlılık Modeli [192]	2.93	9.75	-
Verimli Örnekleme [160]	2.94	-	-
SB-FBSDE [200]	3.01	-	2.96
PDM (VE, FID) [96]	3.04	-	3.36
ES-DDPM [69]	3.11	-	-
SciRE-Çözücü-VP(discrete) [148]	3.15	-	-
DDPM [28]	3.17	9.46	3.72
SN-DDIM [164]	3.22	-	-
INDM (ST) [97]	3.25	-	3.01
PNMD [157]	3.26	-	-

Tablo 4.5 CIFAR-10 görüntüyü üretimi (devamı)

Model	FID	IS	NLL
SN-DDPM [164]	3.31	-	-
Kalibre DPM-Çözücü Ayırık [156]	3.31	-	-
Analitik DDIM [163]	3.39	-	-
NPR-DDIM [164]	3.42	-	-
DPM-Çözücü Ayırık [155]	3.45	-	-
Yumuşak Kesme DDPM++ (VP, NLL) [71]	3.45	9.19	2.88
Analog Bitler [125]	3.48	-	-
NPR-DDPM [164]	3.57	-	3.79
GENIE [201]	3.64	-	-
DSM-EDS [74]	3.65	-	-
Difüzyon Adımlarının Opt. [73]	3.72	-	3.07
Gürültü Arındırınan Difüzyonlu GAN [47]	3.75	9.63	-
TS-DDIM (kuadratik) [167]	3.81	-	-
Yumuşak Difüzyon [50]	3.86	-	-
ScoreFlow (VP, FID) [98]	3.98	-	3.04
WaveDiff [111]	4.01	-	-
DDIM [56]	4.04	-	-
GGDM [161]	4.25	9.19	-
Analitik DDPM [163]	4.31	-	3.42
Karartma Difüzyonu [126]	4.58	9.01	-
INDM (NLL) [97]	4.79	-	2.97
Dinamik Çift Çıkış [60]	5.10	-	-
Diff-Pruning [168]	5.29	-	-
ScoreFlow (deep, alt-VP, NLL) [98]	5.40	-	2.81
PDM (VP, NLL) [96]	6.84	-	2.94
LSGM (NLL) [101]	6.89	-	2.87
D3PM [121]	7.34	8.56	3.44
Gürültü arındırınan öğrenci [83]	9.36	8.36	-
EBM-DRL [202]	9.58	8.30	-
NCSNv2 [203]	10.87	8.40	-
iDDPM (NLL) [54]	11.47	-	2.94
DiffFlow [57]	14.14	-	3.04
DAED [75]	14.2	8.6	-
NCSN [29]	25.32	8.87	-
VDM [58]	-	-	2.49

CelebA. CelebFaces Attributes dataset[204] 10.177 ünlünin 178×218 piksel olan 202.599 adet yüz görüntüsünü içerir. Görüntülerin her biri saç rengi, cinsiyet ve yaş gibi yüz özelliklerini gösteren 40 adet ikili etikete sahiptir. CelebA veri setinin 64x64 piksel sürümü difüzyon modelleriyle görüntü oluşturmada daha sık bir kıyaslama olarak kullanılmıştır. Tablo 4.6'da sonuçlar FID'ye göre sıralanmış

olarak gösterilmektedir.

Tablo 4.6 CelebA(64X64) görüntü üretimi

Model	FID	NLL
DDPM-IP [81]	1.27	-
STDDPM-G++ [169]	1.34	-
INDM (VP, FID) [97]	1.75	2.27
Yumuşak Difüzyon(VE-SDE + Blur) [50]	1.85	-
Yumuşak Kesme DDPM++ (VP, FID) [71]	1.90	2.10
Yumuşak Kesme UNCSN++ (RVE) [71]	1.92	1.97
SciRE-Çözücü [148]	2.02	-
PDM (VP, FID) [96]	2.04	2.23
Kalibre DPM-Çözücü [156]	2.33	-
PDM (VE, FID) [96]	2.50	2.00
INDM (VE, FID) [97]	2.54	2.31
ES-DDPM [69]	2.55	-
DPM-Çözücü Ayrık [155]	2.71	-
PNDM [157]	2.71	
SN-DDIM [164]	2.85	-
Yumuşak Kesme DDPM++ (VP, NLL) [71]	2.90	1.96
Gamma Dağılımlı DDIM [45]	2.92	
INDM (VP, NLL) [97]	3.06	2.05
Analitik DDIM [163]	3.13	-
NPR-DDIM [164]	3.15	-
Karartma Difüzyonu [126]	3.22	-
DDPM [28] ¹	3.26	-
DDIM [56]	3.51	
Gaussian Karışıklı DDIM [45]	3.71	
NCSN++ [30] ²	3.95	2.39
DiffuseVAE [70]	3.97	-
Dinamik Çift Çıkış [60]	4.07	-
Gamma Dağılımlı DDPM [45]	4.09	
TS-DDIM (kuadratik) [167]	4.18	-
SN-DDPM [164]	4.42	-
Analitik DDPM [163]	5.21	2.66
NPR-DDPM [164]	5.33	2.65
Gaussian Karışıklı DDPM [45]	5.57	
EBM-DRL [202]	5.98	-
Diff-Pruning [168]	6.24	-
FastDPM [55]	7.85	-
NCSNv2 [203]	10.23	-
DAED [75]	15.1	-
NCSN [29] ³	25.30	-

¹ Song ve diğ. [56] tarafından raporlanmıştır.

² Kim ve diğ. [71] tarafından raporlanmıştır.

³ Song ve diğ. [203] tarafından raporlanmıştır.

ImageNet. ImageNet veri kümesi [195], WordNet hiyerarşisine göre 14.197.122

açıklamalı görüntü içerişer. 2010'dan beri, görüntü sınıflandırması ve nesne tanıma için bir ölçüt olarak "ImageNet Büyük Ölçekli Görsel Tanıma Yarışması'nda (ILSVRC)" kullanılmaktadır. Görüntü düzeyinde etiketleme görevi için görüntüdeki bir nesne sınıfının varlığını veya yokluğunu gösteren ikili etiketler içerir. Ayrıca, nesne düzeyinde etiketleme görevi için görüntüdeki bir nesne örneğinin etrafında sıkı bir sınırlayıcı kutu ve bir sınıf etiketi içerir. Tablo 4.7, 64x64 piksel ImageNet veri kümelerinin görüntü oluşturma görevinde FID'ye göre sıralanan kıyaslamalarını göstermektedir. Modeller, koşullu ve koşulsuz üretim görevleri olarak iki gruba ayrılmıştır. Eğitimden bağımsız yaklaşım, koşulsuz oluşturma görevinde daha iyi performans göstermektedir ve ayrıca koşullu oluşturma görevinde rekabetçidir.

Tablo 4.7 ImageNet(64x64) görüntü üretimi

Görev	Model	FID	IS	NLL
Koşullu	VDM++ [205]	1.43	64.6	-
	CDM [87]	1.48	67.95	-
	CTM [193]	1.90	-	63.90
	ES-DDPM [69]	2.07	55.29	-
	ADM [179]	2.07	-	-
	iDDPM [54]	2.92	-	-
	Tutarlılık Modeli [192]	4.70	-	-
	Analog Bitler [125]	4.84	-	-
	Kendini Yönlendiren DM [90]	12.1	23.1	-
	BOOT [191]	16.3	-	-
Koşulsuz	Sınıflandırıcıdan Bağımsız Rehberlik [89]	26.22	260.2	-
	Analitik-DDPM [163]	16.14	-	3.61
	SN-DDPM [164]	16.22	-	-
	NPR-DDPM [164]	16.32	-	3.71
	TS-DDIM (kuadratik) [167]	17.20	-	-
	SN-DDIM [164]	17.23	-	-
	NPR-DDIM [164]	17.30	-	-
	Analitik-DDIM [163]	17.44	-	-
	DPM-Çözücü Ayrık [155]	17.47	-	-
	DDIM [56]	17.73	-	-
	GGDM [161]	18.4	18.12	-
	iDDPM [54]	19.2	-	3.57
	VDM [58]	-	-	3.40
	Verimli Örnekleme [160]	-	-	3.55

4.4 Sonuç

Bu bölümde, üretken difüzyon modellerinin teorik perspektifinin kapsamlı ve güncel bir incelemesi sunulmuştur. Üç ana çalışma ile difüzyon modellerine bir giriş yapılmış, sonrasında ise mevcut algoritmayı iyileştirmeye yönelik çalışmalar

incelenmiştir. Teorik gelişmeler, eğitim tabanlı ve eğitimden bağımsız geliştirmeler olarak iki kategoriye ayrılmıştır. Sonunda ise literatürde geçen yöntemlerin karşılaştırılması için mevcut değerlendirme kriterleri ve veri kümelerinden bahsedilmiş ve sonuçlar tablolar halinde raporlanmıştır.

Difüzyon modelleri güçlü bir araştırma altyapısına sahiptir. İleri sürecin veri hakkındaki tüm bilgileri tamamen sildiği ve rastgele bir dağılımla sonuçlandığı varsayılar, ancak durum her zaman böyle olmayı bilir. Gerçekte, bilginin tamamen gürültüye dönüştürülmesi sonlu bir sürede gerçekleştirilemez. Gürültü Arındırın Difüzyon Örnекleyicileri (DDS) [206], karşılık gelen zaman tersine çevirmeyi(time reversal) Monte Carlo örneklemesiyle tahmin ederek teorik garanti sağlar. Ayrıca bazı çalışmalar [137, 207, 208] iyi bilinen fonksiyon uzayları difüzyon modellemesinde teorik garanti sağlar. McAllester [209] ise difüzyon modellerinin matematiksel anlayışını sunarak difüzyon modellerinin daha iyi anlaşılmasına katkıda bulunmuştur.

Difüzyon modelleri hem teorik hem de pratik açıdan hızla gelişmektedir. Aşağıda gelecekteki araştırmalar için bazı yollar sunulmuştur:

- Hedef fonksiyonu: Çoğu difüzyon modeli eğitim hedefi olarak ELBO'yu alır. Ancak, ELBO ve NLL'yi aynı anda optimize etme konusu tartışımalıdır. Log-olasılık optimizasyonunu mevcut değişkenlere bağlayan veya ELBO yerine olasılık açısından tutarlı yeni eğitim hedefleri kullanan daha gelişmiş analitik yaklaşım performansı iyileştirebilir.
- Ağ tasarımı: Ağ tasarımı hakkında, görsel dönüştürücüler (ViT) içeren hibrit modeller vardır, U-ViT [210], GenViT ve HybViT [211], difüzyon modelleriyle büyük ölçekli çok kipli veri kümelerini öğrenmek için önemli olan uzun atlama bağlantıları sağlar. Farklı ağ yapıları kullanılarak da iyileştirmeler yapılabilir.
- Yüksek derecede bozulmuş örnekler: Bazı uygulama alanlarında araştırmacıların bozulmamış örnekler erişimi yoktur veya sahip olmak pahalıdır. Daras ve diğerleri [212] difüzyon sürecinde ek ölçüm bozulmasını tanıtarak modelin yalnızca yüksek derecede bozulmuş örnekler kullanarak örnek üretebildiğini göstermiştir. Bu alanlardaki gelişmelerle difüzyon modelleri ilgili araştırmacılara daha çok yardımcı olabilir.
- Ölçeklenebilirlik ve verimlilik: Difüzyon modelleri ile daha büyük veri kümelerini ve daha yüksek çözünürlüklü girdileri işlemek önemli bir zorluktur. Gelecekteki araştırmalar, modern donanım mimarilerinde büyük

ölçekli üretken difüzyon modellerinin eğitimini mümkün kılmak için daha verimli eğitim prosedürleri, paralellik teknikleri ve dağıtık(distributed) işleme stratejileri tasarlamaya odaklanabilir.

- Çok kipli üretim: Modelin aynı anda birden fazla kipte çeşitli çıktılar (örneğin, görüntüler ve karşılık gelen metinsel açıklamalar) üretebildiği çok kipli üretimi destekleyecek şekilde üretken difüzyon modellerini genişletmek heyecan verici bir yöndür. Bu konuda, farklı kipler arasındaki karmaşık ilişkileri ve bağımlılıkları yakalamak için yeni mimariler ve eğitim hedefleri tasarlanabilir.
- Yapılandırılmış üretim: Sekanslar, graflar veya 3 boyutlu yapılar gibi yapılandırılmış çıktılar üretmek için üretken difüzyon modellerini etkinleştirmek, bir diğer araştırma yönüdür. Bu konuda, yapılandırılmış verilerde bulunan karmaşık bağımlılıkları modelleme yeteneğine sahip uzmanlaşmış mimariler ve çıkarım algoritmalarının geliştirilmesi gereklidir.
- Önceki bilgiyi dahil etme: Önceki bilgiyi veya alana özgü kısıtlamaları üretken difüzyon modellerine entegre etmek, performanslarını ve yorumlanabilirliklerini iyileştirebilir. Gelecekte, daha gerçekçi ve anlamsal olarak tutarlı çıktıların üretilmesine rehberlik etmek için yapılandırılmış bilgi grafiklerini, ontolojileri veya fizik yasalarını modelleme sürecine dahil etme teknikleri araştırılabilir.
- Sağlamlık(robustness) ve genelleme: Üretken difüzyon modellerinin sağlamlık ve genelleme yeteneklerini geliştirmek, özellikle sınırlı veya gürültülü eğitim verilerine sahip senaryolarda önemli bir zorluktur. Bu araştırma konusu, düşmanca saldırılara, alan kaymasına(domain shift) ve dağılım dışı girdilere karşı sağlam eğitim için teknikler geliştirmeyi ve model kararsızlığını(uncertainty) değerlendirme ve ölçme tekniklerini içerir.
- Değerlendirme kriterleri: Veri dağılımları olasılık eşleşmesiyle kovaryant değildir, bu nedenle örnek çeşitliliğini ve modelin üretim kapasitesini daha doğru ve kapsamlı bir şekilde belirleyen bazı değerlendirme kriterlerine ihtiyaç vardır. Stein ve diğerleri [213] mevcut değerlendirme kriterlerini tartışmış ve bu kriterlerin ezberlemeyi(memorization) tespit etmediğini ileri sürmüştür.

Bilimsel topluluk içinde devam eden araştırma çabaları ve işbirliklerinin yönlendirmesiyle, üretken difüzyon modellerinin geleceğinde model mimarisindeki gelişmelerin, eğitim algoritmalarının ve çeşitli alanlardaki uygulamaların bir kombinasyonunun yer alması muhtemeldir.

5

DİFÜZYON MODELLERİ İLE METİNDEN GÖRÜNTÜ ÜRETİMİ

Bu bölümde öncelikle difüzyon modelleri ile metinden görüntü üretimi konusunda tez kapsamında ilgili olan bazı çalışmalara, ardından sonuçlanmayan bazı yöntemlerimize kısaca değinilmektedir.

5.1 İlgili Çalışmalar

Metinden görüntü üretimi, difüzyon modellerinin bulunması ile son yıllarda oldukça gelişen bir araştırma konusudur. Metinden görüntü üreten difüzyon modelleri [102, 186, 187], istikrarlı öğrenme hedefleri ve büyük ölçekli eşleştirilmiş veri kümeleri ile eğitilmeleri nedeniyle yüksek görüntü kalitesi ve anlamsal tutarlılık göstermektedir. Bunlar arasından Stable Diffusion [102] açık kaynaklı bir gizli difüzyon modeli olarak ölçeklenebilir yapısıyla bir çok araştırmada referans model olarak kullanılmıştır.

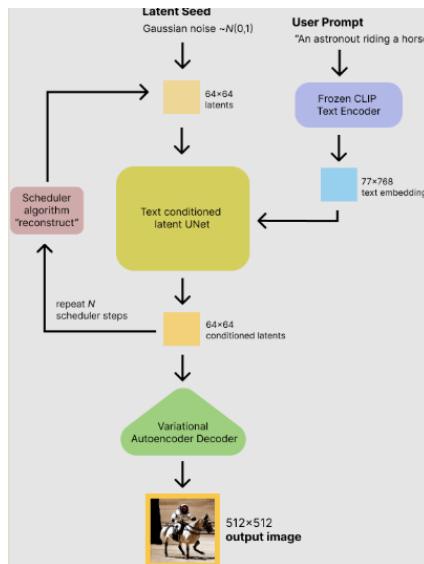
Tüm başarılarına rağmen, difüzyon modellerinde metin istemi ile üretilen görüntünün anlamsal olarak uyumu noktasında eksikliklerle karşılaşmaktadır. Metin isteminde birden çok nesne olduğu durumlarda nesnelerin özniteliklerini ve aralarındaki ilişkileri yakalamak oldukça zordur. Model büyülüüğünü artırmak her ne kadar bu zorluğun üstesinden gelmeyi sağlasa da maliyet açısından etkin bir yöntem değildir. Eğitim tabanlı yaklaşım uzun bir eğitim süreci gerektirmekte olduğundan yine hesaplama maliyeti getirmektedir. Diğer yandan eğitimden bağımsız olarak çıkarım zamanında yapılan iyileştirmelerle anlamsal uyumun daha iyi yakalanması da mümkündür. Bu tez kapsamında odaklanılan konu, metin / görüntü uyumunun çıkarım zamanında iyileştirilmesidir.

Bu bölümde öncelikle metinden görüntü üreten Stable Diffusion modeli tanıtılacak, ardından metinden görüntü üretiminde anlamsal uyumu çıkarım zamanında iyileştirmeye çalışan bazı yöntemlerden bahsedilecektir.

5.1.1 Stable Diffusion

Stable Diffusion(SD), Gizli Difüzyon Modeli(LDM) mimarisine sahiptir. Gizli difüzyon modellerinde metinler ve onlara karşılık gelen görüntüler ortak bir gizli uzayda temsil edilir. Verilen metin girdisinin bu uzaydaki temsili bulunur ve görüntü üreten model bu temsil ile koşullandırılarak metinden görüntü üretimi gerçekleştirilir. Bölüm 4.3.1.5'te de debynildiği gibi LDM'de öncelikle otokodlayıcı eğitilerek düşük boyutlu bir gizli uzay elde edilir. Daha sonra, difüzyon modeli gizli kodları bu gizli uzayda üretir. LDM'de, UNet 2 boyutlu evrişimsel katmanlardan inşa edildiğinde yeniden ağırlıklandırılmış hedef en alaklı bitlere odaklanır. UNet'in çapraz dikkat mekanizmasıyla zenginleştirilmesi ile de esnek koşullu üretim etkinleştirilmiş olur.

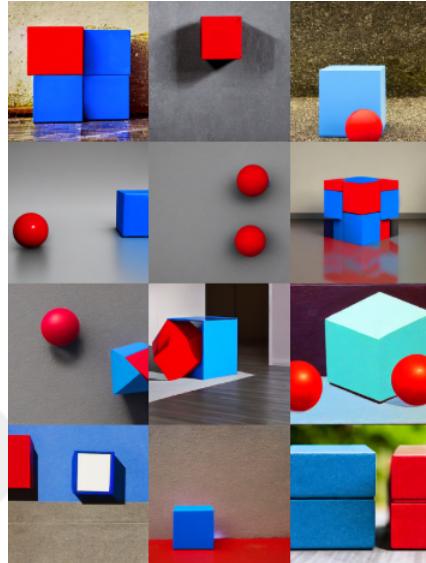
Stable Diffusion, gizli difüzyon modeli mimarisi kullanılarak internetteki çeşitli kaynaklardan elde edilen metin-görüntü çiftleri ile önceden eğitilmiş bir modeldir. Stable Diffusion'da, metinler ve görüntüler CLIP gizli temsilleriyle eğitilir. Şekil 5.1 Stable Diffusion'ın çıkışım zamanını göstermektedir. Çıkarım zamanında rastgele bir gürültüyle beraber metinlerin gizli temsilleri metin şartlı Unet'e verilerek gizli uzayda bir görüntü temsili elde edilir. Daha sonra rastgele gürültü yerine bu temsil kullanılarak belli bir adım boyunca bu işlem tekrarlanır ve temsil iyileştirilmeye çalışılır. Görüntü temsili yeterince iyi hale geldiğinde VAE kodçözücü ile gerçek görüntüye çevrilir.



Şekil 5.1 Stable Diffusion'ın çıkışım zamanı [214]

Stable Diffusion, görüntü üretmede oldukça başarılı olsa da, birden fazla nesne, öznitelik ve ilişki içeren metin koşullarını üretmekte zorlanır. Özellikle de metin isteminde yer alan bir nesnenin görüntüde yer almaması yani nesne ihmali(neglect) ve nesne ile sahip olduğu özniteligin eşleşmemesi yani yanlış

öznitelik bağlanması(binding) problemleriyle oldukça sık karşılaşılmaktadır. Şekil 5.2'de SD v1.4'ün "a blue cube and a red ball" istemi için ürettiği görüntülerde bu problemler açıkça görülmektedir. İlk versiyonlar örnekteki gibi basit ilişkileri yakalamakta zorlanırken, sonraki versiyonlarda daha karmaşık sahnelerde yine aynı problemlerle karşılaşılmaktadır. Bu bağlamda metinden görüntü üretiminde görüntü-metin hizalamasını sağlamak önemli bir araştırma alanıdır.



Sekil 5.2 Stable Diffusion'ın zorlanması.

5.1.2 Anlamsal Uyumun Çıkarım Zamanında İyileştirilmesi

Görüntü-metin hizalamasını yeniden eğitim gerektirmeden çıkışım zamanında iyileştirmek mümkündür. Literatürde çıkışım zamanında anlamsal uyumu iyileştirmeyi amaçlayan bir kaç çalışma mevcuttur.

Çapraz-dikkat haritalarını kullanmak, çıkışım zamanında nesnelerin konumlarını belirleyip diğer nesneleri ona göre yerleştirmemize yardımcı olabilir. Ancak, nesneler ve ilişkiler alışılmadık, çok sayıda ve karmaşık olabileceğinden çapraz dikkat haritaları ile bunları yakalamak zor olabilir.

Metin girdisi ile beraber yerleşim(layout) bilgisini de alan yerleşimli metinden görüntü üretimi [215–217] belirli sayıda nesne ve öznitelik söz konusu olduğunda bu sorunu kısmen aşmaktadır. Ancak yerleşim bilgisini girdi olarak vermek için kullanıcıların bu bilgiyi modele sağlaması gerekmektedir ve bu bilginin her seferinde elle sağlanması ugraşılıcı bir durumdur.

Yerleşim bilgisini elde etmek Büyük Dil Modellerinin (LLM) doğal dilleri analiz etme ve anlama yetenekleriyle mümkün olabilir. Bazı çalışmalar [218–

220] ilişkileri yakalamak için yerleşim bilgisi elde etmek amacıyla LLM'leri kullanmıştır. Ancak LLM'lerden elde edilen yerleşim bilgisi ile nesneler arasındaki ilişkileri gerçekçi bir şekilde yakalamak çoğu zaman mümkün olmamaktadır.

Bu bölümde anlamsal uyumu çıkarım zamanında iyileştirmeye çalışan araştırmalar, ilgili konu başlıklarını altında ele alınacaktır.

5.1.2.1 Çapraz Dikkat Haritalarını Kullanan Çalışmalar

Yapıldırılmış(Structured) Difüzyon [221], CLIP temsillerinin nedensel dikkat maskeleri nedeniyle, bir dizinin sonraki bölümündeki birimlerin(token) önceki birimlerin anamlarıyla harmanlandığını ve bu nedenle yanlış öznitelik bağlanması neden olabileceğini ileri sürmüştür. Bu durumu çözmek için, metin istemi birkaç isim öbegine bölünmüştür. Her isim öbegi için bir dikkat haritası hesaplanır ve çapraz dikkat biriminin çıktısı, tüm dikkat işlemlerinin ortalamasıdır.

İkincisi, önceden eğitilmiş bir difüzyon modelinin birden fazla çıktısını oluşturan Şekillendirilebilir(Composable) Difüzyondur [222]. Karmaşık bir görüntüyü elde etmek için görüntünün farklı bileşenleri ayrı ayrı üretilir ve daha sonra elde edilen çıktılar operatörlerle birleştirilir.

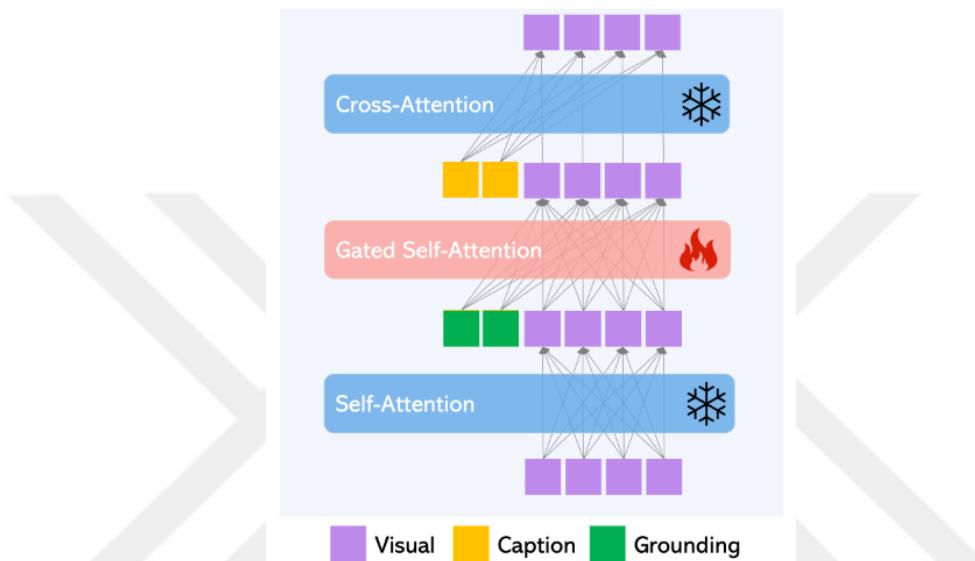
Üçüncü araştırma Attend and Excite [223]'dır. Bu çalışmada, her bir nesnenin görüntünün bir parçasında dikkate alınmasını sağlayacak bir mekanizma geliştirilmiştir. Bir nesnenin oluşturulan görüntüde mevcut olması için, model nesneye en az bir görüntü parçası atamalıdır ya da başka bir deyişle her bir nesnenin görüntünün bir parçasında baskın olması gereklidir.

Takip eden bir araştırma Divide and Bind [224], metinlerin karmaşıklığı arttıkça, nesnelerin rekabetinin yoğunlaşlığı gerçekini gün yüzüne çıkartır. Bu rekabetle başa çıkabilmek için dikkat haritasını birden fazla bölgeye bölmeyi önermişlerdir. Bu sayede her nesne için dikkati mekansal olarak dağıtarak, belirtilen tüm nesnelerin yüksek rekabet altında bile oluşturulması için olanak tanımaktadır.

Tüm bu çalışmalar etkileyici olmasına rağmen, metin istemi birçok nesneye sahip olduğunda ve bu nesnelerin öznitelikleri ve hatta aralarında ilişkileri de olduğunda modeller bu karmaşık metin koşullarını gerçekleştirme zorlanmaktadır. Bu gibi durumlarda, nesnelerin kesin konumu gibi metin girişini destekleyen diğer girdilere sahip olmak verimli olabilir.

5.1.2.2 Yerleşimli Metinden Görüntü Üretime

Literatürde yerleşimli metinden görüntü üretimi başlığıyla adlandırılabilenek bazı çalışmalar vardır [215–217, 225–229]. Bunlardan en popüler olan çalışma GLIGEN: Açık Kümeli Yerleşimli Metinden Görüntü Üretimi [215], önceden eğitilmiş Stable Diffusion üzerine kuruludur ve yerleşim bilgisini Şekil 5.3'te gösterildiği gibi kapılı(gated) bir mekanizma aracılığıyla yeni eğitilebilir katmanlarda enjekte eder. GLIGEN'deki Kapılı Öz-Dikkat katmanı, metni yerleşim bilgisıyla yerleştirmek için kullanılmaktadır.



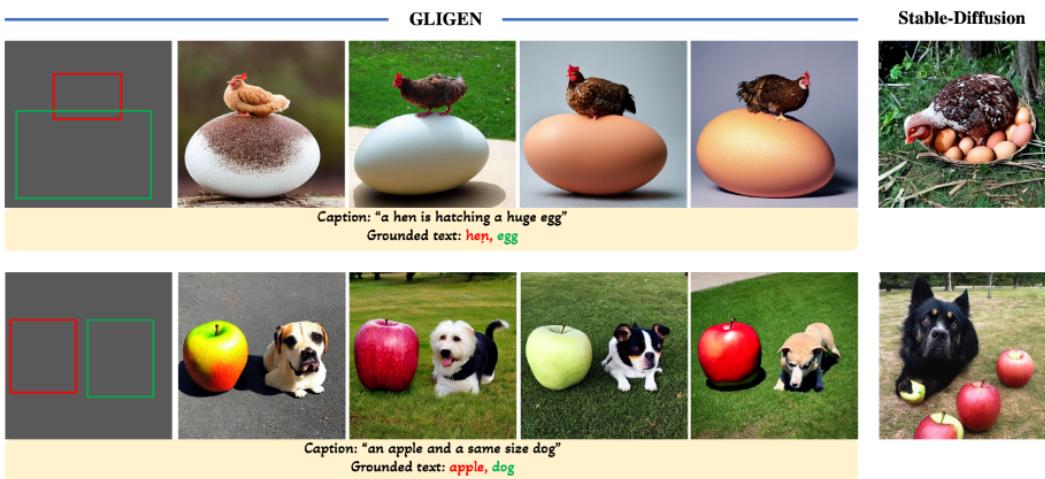
Şekil 5.3 GLIGEN'in kapılı öz dikkat katmanı[215]

Yerleşimli metinden görüntü üretiminde, metin isteminin yanı sıra yerleşim bilgisi de koşul girişi olarak kabul edilir. Şekil 5.4'te gösterildiği gibi GLIGEN, her biri iki tür bilgiden oluşan yerleştirme birimleri kullanır: yerleştirilen nesnenin anlamı (kodlanmış metin veya görüntü) ve bölgesel konum (kodlanmış sınırlayıcı kutu veya anahtar noktalar).

GLIGEN ince ayrıntılı bir rehberlik sağlayarak başarılı olsa da, kullanıcıların bu rehberlik bilgilerini elle sağlaması gerekmektedir ve bu bilginin her seferinde elde edilmesi zor olabilir. Bu noktada doğal dilde verilen bir metni, yerleşim bilgisine dönüştürmek için büyük dil modellerinden yardım alınabilir.

5.1.2.3 Büyük Dil Modeli Rehberliği

LLM'ler, az atımlı öğrenme [230, 231], düşünce zinciri muhakemesi [232] ve kendi kendini düzeltme [233] yetenekleriyle birçok kullanım durumunda güçlerini kanıtlamışlardır.



Şekil 5.4 GLIGEN ile yerleştirme[215]

Ayrıca, çok-kipli bir uzayda farklı kipleri birleştirebilmeleri görsel yorumlama yeteneğine sahip olmalarını sağlamaktadır. Bu yetenekten faydalananarak çeşitli görevlerde daha iyi başarılar elde eden çalışmalar mevcuttur [234–238].

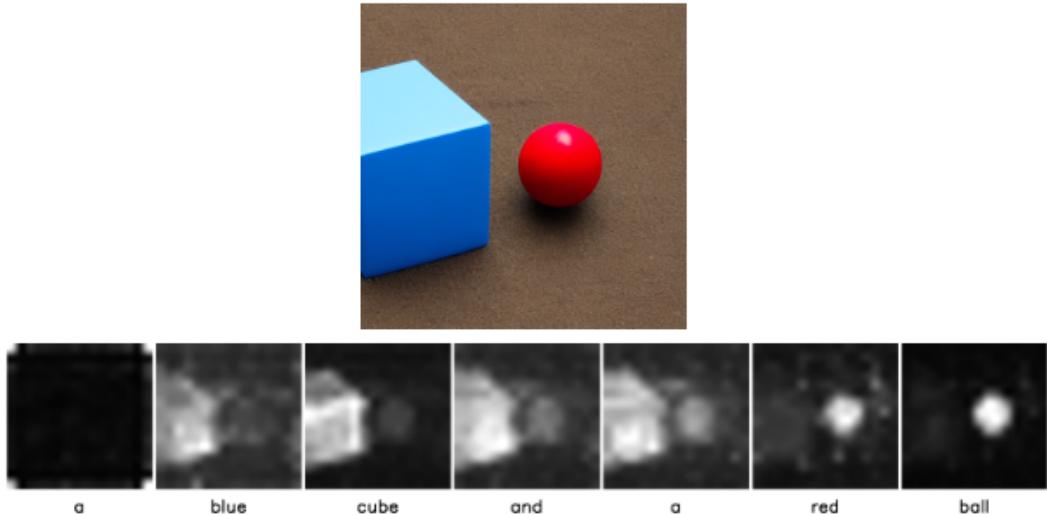
Verilen bir metin koşulu C için, yerleşim bilgisi oluşturmanın amacı $O = \{o_j; j = 1, 2, \dots, n\}$ yerleştirme birimlerinin bir kümesini tahmin etmektir; burada her o_j yerleştirme birimi, j nesnesinin bilgisini belirtir. o_j , bir d_j tanımı ve sınırlayıcı kutu bilgisi b_j 'den oluşur. Yani $o_j = \{d_j, b_j\}$. $b_j = \{x_j, y_j, w_j, h_j\}$ şeklinde gösterebiliriz; burada x_j, y_j sol üst noktayı ve w_j, h_j sırasıyla genişliği ve yüksekliği belirtir. Bu şekilde, yerleşim bilgisini metinsel olarak ifade edilebilir. Dolayısıyla, herhangi bir LLM'den C metin koşulunu girdi istemi olarak vererek bir yerleşim bilgisi istenebilir. LLM'lerden yararlanarak yerleşim bilgisi üreten bazı araştırmalar [218–220] vardır. Bu fikri, [239], kendi kendini düzeltten [240] ve üretilen görüntüyü puanlayarak [241] daha da geliştiren bazı çalışmalar da mevcuttur.

5.2 Sonuçlanmayan Yöntemler

Bu bölümde çapraz dikkat haritaları ve büyük dil modelleri ile denemiş olduğumuz ancak sonuçlanmayan bazı çalışmalara kısaca degeinilecektir.

5.2.1 Çapraz Dikkat Haritaları ile Nesneleri Yerleştirme

Çapraz dikkat haritaları Şekil 5.5 'te gösterildiği gibi nesne ile görüntüdeki hangi piksellerin ilişkili olduğunu ortaya çıkartabilir. Nesneler arasındaki ilişkileri çözebilmek için önce bir nesne üretilir, daha sonra çapraz dikkat haritası ile bu nesnenin yerlesiği bölge belirlenir. Sonrasında ise aralarındaki ilişkiye göre diğer nesneye pikseller atanır.

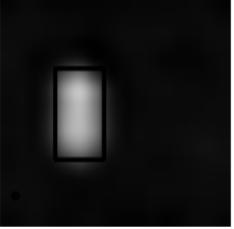


Şekil 5.5 Çapraz dikkat haritaları

Çapraz dikkat haritaları ile yerleştirme işlemi aşağıdaki maddelerde açıklanmaktadır:

1. Şekil 5.6'da gösterildiği gibi, ilk nesnenin çapraz dikkat haritalarıyla sınırlayıcı kutusu belirlenir. Burada en az 5 adımlık bir çıkarım süreci gerekmektedir.
2. Daha sonra ilişkide olduğu ikinci nesnenin merkezi hesaplanır:
 - önünde, altında, aşağısında → yukarıdaki pikseller
 - arkasında, üzerinde, yukarısında → aşağıdaki pikseller
 - yakınında, yanında → yandaki pikseller
 - içinde, arasında → etrafındaki pikseller
3. Daha sonra ikinci nesne için bir sınırlayıcı kutu oluşturulur.
4. Son olarak, Şekil 5.7'deki gibi iç boyama(inpainting) ile ikinci nesne üretilir.

Burada gösterilen sonuçlar umut verici birlikte çapraz dikkat haritalarının her durumda yeterli olabileceği bir tartışma konusudur. Kedi-köpek, anne-baba, domates-elma gibi şekli benzeyen nesnelerin çapraz dikkat haritaları aynı pikselleri işaretlemektedir. Dahası nesnelerin büyülüğu ile ilgili bir ölçekleme problemi de oluşabilmektedir. Örneğin, önce kedi üretip sonra ev yerleştirmek isteyince, evi siğdıracak yer bulunamadığından istenen metin ile görüntü karşılığı uyumlu olmamaktadır. Karmaşık metin istemleri daha fazla nesne ve ilişkiler içerdiginden çapraz dikkat haritaları bunları yorumlamakta çoğu zaman zorlanmaktadır.

a white dog				
3. adım	4. adım	5. adım	6. adım	
				

Şekil 5.6 Çapraz dikkat haritaları ile ilk nesnenin yerini bulma

a white dog behind a blue ball				
Stable Diffusion	Attend-and-Excite	Diffusion Layout + GLIGEN	Referans Görüntü	Çapraz Dikkat Haritası
				
				

Şekil 5.7 Çapraz dikkat haritaları ile yerleştirme

5.2.2 Yerleşim Bilgisi için Büyük Dil Modeli İnce Ayarlama

Bir büyük dil modeli, metin-sınırlayıcı kutu şeklinde veri çiftlerine sahip bir veri kümesi ile eğitilirse bu modeli kullanarak metin isteminden yerleşim bilgisi elde edilebilir. Bunun için öncelikle uygun veri kümesini bulmak gereklidir. COCO veri kümesindeki sınırlayıcı kutular ve etiketler metinlerle eşleştirerek gereken veriler şu şekilde elde edilmiştir:

*'text': 'A clock with the appearance of the wheel of a bicycle. ',
'bbox': 'bicycle [5.16, 54.19, 390.54, 293.34] clock [8.99, 171.75, 167.19, 172.58]'*

Eğitim veri kümesinde 118287 örnek, doğrulama veri kümesinde ise 5000 örnek bulunmaktadır. Bu verilere ön işlemler uygulandıktan sonra ve T5-base [242] (sekanstan sekans üretimi için 220M parametreli bir kodlayıcı-kod çözümü modeli) 4 çevrim(epoch) (5 saat) eğitilmiştir. Üretilen bazı örnekler aşağıda verilmektedir:

*Kitten looking puzzled sitting in a bathroom sink.
cat [0.0, 0.0, 638.0, 418.45] sink [0.0, 0.0, 640.0, 418.45]*

*A small plane taking off from a runway in mid air.
airplane [0.0, 0.0, 640.0, 418.45]*

*A young man takes a picture of himself in the mirror.
person [0.0, 0.0, 640.0, 478.9]*

Burada görüldüğü gibi model nesneleri çıkartmakta iyi olsa da yerleşim bilgileri aynı koordinatları işaret etmektedir. Yerleşim bilgilerini öğrenen bir model tasarlamak için bu bilgiyi optimize edecek bir formül bulunmalıdır.

Nesne tespiti(object detection) görevlerinde kayıp fonksiyonlarında bir sınırlayıcı kutu regresyonu metriği yer almaktadır. IoU (Intersection over Union) adı verilen bu metrikte tahmin edilen kutu ile gerçek kutu arasındaki kesişimin kutuların birleşimine oranı bulunmaktadır. Bunu geliştirerek GIoU (Generalized IoU)[243], DIoU (Distance IoU) ve CIoU (Complete IoU)[244] metrikleri önerilmiştir. Bu metrikler nesnelerin sınırlayıcı kutularının konumlarını optimize etmek için kullanılabilir. Ancak büyük dil modellerinin kayıp fonksiyonunu doğrudan bu metrikle değiştirmek işe yaramayacaktır.

5.2.2.1 Dönüştürüçülerin Pekiştirmeli Öğrenmesi

Dönüştürüçü modellerin doğası gereği üretim birim(token) bazında gerçekleştiriliyor ve gradyanlar eğitimin sonunda güncellenir. Bu nedenle eğitim sırasında güncel

modelden çıktı üretmek mümkün değildir. Sekans seviyesindeki kayıpları hesaplamanın yolu, bu işi yapmak üzere bir işlevi eğitmektir. Örnekleme yapmak için bir Pekiştirmeli(Reinforcement) Öğrenme aracı kullanıldığında, bu aracının politikası ile sekans seviyesindeki kayıplar hesaplanabilir.

Bu bağlamda bir Proksimal Politika Optimizasyonu (PPO) tasarlanmış, model çıktılarının IoU metriklerini hesaplayıp gerçek sınırlayıcı kutu etiketlerine yakınlığını ölçen bir fonksiyon politika olarak belirlenmiştir. Model çıktıları gerçeklerine yakın olduğunda model ödüllendirilecek ve böylelikle daha iyi sınırlayıcı kutular üretmek için model optimize edilmiş olacaktır. İşlem adımları aşağıda verilmektedir:

- COCO veri kümesi yüklenir.
- İnce ayarlanmış T5-base yüklenir.
- Her bir örnek için modelin cevabı alınır.
- Eğitim kümesinden gerçek sınırlayıcı kutular alınır ve ödül hesaplanır.
- Hesaplanan ödül ile T5-base pekiştirmeli olarak eğitilir.

Görevi basitleştirmek için denemeler öncelikle metin-nesneler veri kümesi ile gerçekleştirilmiştir. 64 doğrulama örneğinin F1-score ortalamaları referans model için 0.273, ince ayarlanmış model için 0.316 olarak bulunmuştur. Yani metin-nesneler veri kümesi için PPO iyileşme sağlamaktadır.

Başka bir kolay görev tek nesneli verilerden oluşmaktadır. Bu görev için öncelikle bir Regex tanımlanarak tahmin formatının geçerli olup olmadığı kontrol edilecektir. Daha sonra tahmin edilen nesnenin doğru olup olmadığını kontrol edilecek ve son olarak da tahmin edilen sınırlayıcı kutular çıkartılarak ve regresyon puanı hesaplanacaktır.

Pekiştirmeli öğrenme, model tahminlerinin bir bütün halinde ele alındığı durumlarda efektif bir yöntemdir. Ancak sınırlayıcı kutu tahminine(IoU) gelince, modelin çıktısı parçalara ayrılarak değerlendirmeye tabi tutulduğundan model politikayı öğrenmemiştir. Bu nedenle başka bir dil modelini ince ayarlayarak denemelere devam edilmiştir.

5.2.2.2 Stable Beluga-13B İnce Ayarlama

Stable Beluga [245], Llama2 [246] mimarisini kullanarak önceden eğitilmiş açık kaynaklı bir büyük dil modelidir. 21391 örnek içeren 2 nesneli COCO veri

kümlesi Stable Beluga'nın 13 milyar parametreli versiyonunu ince ayarlamak için kullanılmıştır. Sonuçta bağlam-içi öğrenen ve ince ayarlanmış modeller baz modelden daha iyi yerleşim planı üretmeyi başarmıştır. Ancak yerleşim planının daha mantıklı olması üretilen görüntünün metin istemiyle uyumlu olacağını garanti etmemektedir.

Bu noktada bir metin istemi için modele belli bir yerleşim bilgisini öğretmeye çalışmanın ne kadar verimli olacağı sorusu akla gelmektedir. Bu durumun model çıktılarında çeşitliliği azaltması mümkündür. Ancak üretken modellerden beklenen istenen şartı sağlamak koşuluyla olabildiğince fazla çeşitlilikte örnek üretmektir. Dahası, mantıklı gibi görünen bir yerleşim planı kullanılarak elde edilen görüntü her zaman metin istemini karşılamamaktadır. Bu noktada görüntünün metinle uyumlu olduğundan emin olmak için gözle görülmesi gerekiği gerçeği ortaya çıkmıştır.

5.2.3 Sahneyi Planlama

Sahneyi planlamak için LLM kullandığımızda aynı sahne için çeşitli yerleşim bilgileri üretilebilecektir. Aynı zamanda LLM'lerin görüntü yorumlama yeteneği kullanılarak görüntü-metin uyumunun sağlandığından emin olunabilir.

Öncelikle LLM'den sahneyi planlaması istenir. Daha sonra verilen plana göre yerleşim bilgileri sorgulanır. Yerleşim bilgileri kullanılarak arka plan görüntüsünün üzerine iç boyama yapılır ve son görüntü elde edilir. Kullanılan komutlar aşağıda verilmektedir:

Komut 1: Sahneleri görselleştirmeme yardımcı olan akıllı asistanımsın. Bir fotoğraf, resim veya çizim için bir başlık sağlayacağım. Görevin, başlıklı tüm öğeleri doğru şekilde görselleştirmek için adım adım bir kılavuz oluşturmaktır. Uygun bir arka planla başlamalı ve nesneler arasındaki eylemlere ve ilişkilere dikkat etmelisin. Lütfen adımlarda seçenekler sunma ve sahneyi en fazla 5 adımda tamamla.

Komut 2: Rehberliğine dayanarak ilk adımı görselleştirdim ve görseli sağladım. Görsele ve rehberliğine göre, sonraki adımları nereye yerlestirebilirim? Lütfen bana her bir adının sınırlayıcı kutu koordinatlarını şu biçimde ver: [sol üst x koordinatı, sol üst y koordinatı, sağ alt x koordinatı, sağ alt y koordinatı]

Bu sorgular için öncelikle Stable Beluga modeli kullanılmış, daha sonra doğruluk ve çeşitlilik bakımından daha iyi olan ve ayrıca görsel yorumlama yeteneği olan GPT-4o modeline geçiş yapılmıştır. Bu şekilde elde edilen yerleşim bilgilerine göre "A dog is barking at a squirrel in a tree." istemi için tek adımda ve adım adım

üretilen görüntüler sırasıyla Şekil 5.8 ve 5.9'da verilmiştir.



Şekil 5.8 Tek adımda tüm nesneleri yerleştirme



Şekil 5.9 Nesneleri adım adım yerleştirme

Bu bakış açısı mantıklı gibi görünse de pratikte arka plan bilgisi ile ilgili sorunlar yaşanmaktadır. LLM farklı nesnelerin bulunduğu bir arka plan ürettiğinde, bu arka plan üzerine başka nesnelerin yerleştirilmesi uygun olmamaktadır(çamaşır odası-sepet). Ayrıca arka plan çok geniş kapsamlı olduğunda küçük nesneler görülmeyecek kadar küçük olmaktadır(göl-harita).

5.3 Çok Kipli Büyük Dil Modeli Rehberliği

Çok Kipli Büyük Dil Modeli Rehberliği Şekil 5.10'da verilmiştir.

5.3.1 Rehberlik Adımları

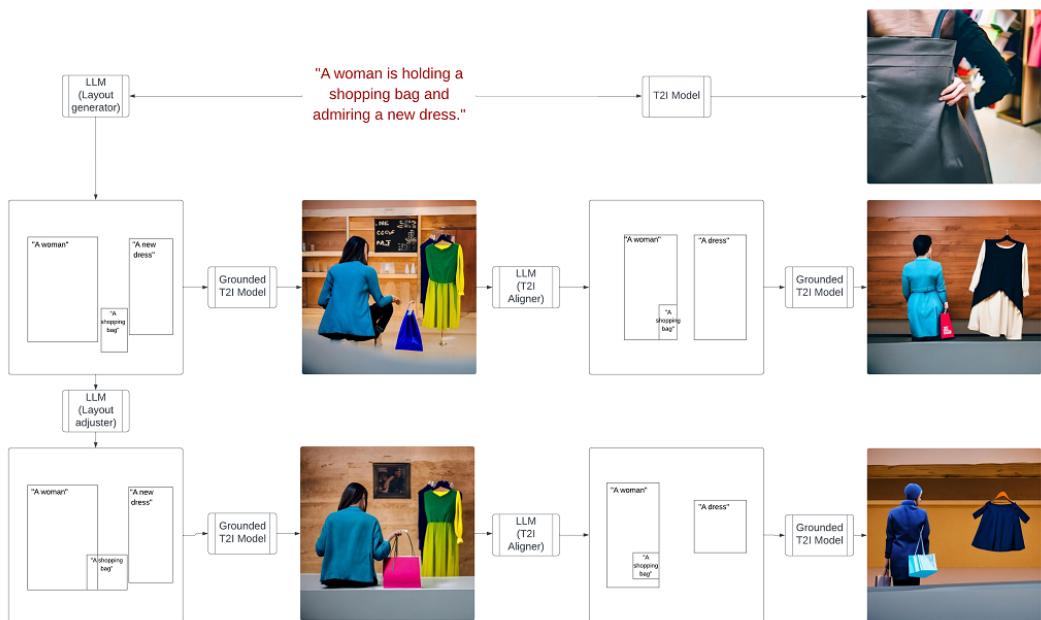
Rehberlik LLM-yerleştirme, LLM-kendini-düzelme, LLM-hızalama ve LLM-akıl-yürütmeye olmak üzere 4 adımdan oluşur. Bu adımlar sırayla açıklanacaktır.

5.3.1.1 LLM-yerleştirme

LLM-yerleştirme Şekil 5.11'de ve Tablo 5.1'de verilmiştir.

5.3.1.2 LLM-kendini-düzelme

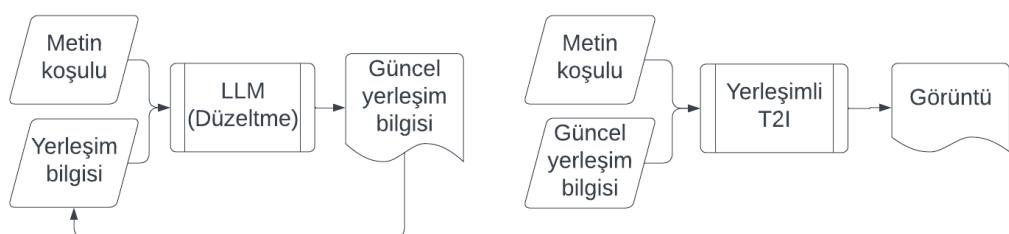
LLM-kendini-düzelme Şekil 5.12'de ve Tablo 5.2'de verilmiştir.



Şekil 5.10 Çok Kipli Büyük Dil Modeli Rehberliğine Genel Bakış



Şekil 5.11 LLM-yerleştirme ile görüntü üretimi



Şekil 5.12 LLM-kendini-düzelme ile görüntü üretimi

Tablo 5.1 LLM-yerleştirme ile görüntü üretimi algoritması

Girdiler: Metin koşulu C , Destekleyici örnekler S_g

$O \leftarrow \text{LLM-yerleştirme}(C, S_g)$

$I \leftarrow \text{Yerleşimli-T2I}(C, O)$

Çıktı: Görüntü I .

Tablo 5.2 LLM-kendini-düzelte ile görüntü üretimi algoritması

Girdiler: Metin koşulu C , Nesnelerin önceden üretilen yerleşim bilgisi O_0 , Destekleyici örnekler S_s , Kendini düzeltme için en fazla tur sayısı K .

for $k = 1$ **to** K **do**

$O_k \leftarrow \text{LLM-kendini-düzelte}(C, O_{k-1}, S_s)$

if $O_k == O_{k-1}$ **then**

break

end if

end for

$I \leftarrow \text{Yerleşimli-T2I}(C, O_k)$

Çıktı: Görüntü I .

5.3.1.3 LLM-hızalama

LLM-hızalama Şekil 5.13'te ve Tablo 5.3'te verilmiştir.



Şekil 5.13 LLM-hızalama ile görüntü üretimi

5.3.1.4 LLM-akıl yürütme

LLM-akıl-yürütme Şekil 5.14'te ve Tablo 5.4'te verilmiştir.

5.3.2 Nitel Deneyler

5.3.2.1 Yerleştirme & Kendini Düzeltme

Önerilen yöntemlerin nitel performansı, Stable Diffusion v2.1, LMD+ [220] ve LayoutGPT [219] ile karşılaştırılmıştır.

Tablo 5.3 LLM-hizalama ile görüntü üretimi algoritması

Girdiler: Metin koşulu C , Önceden üretilen görüntü I_0 , Destekleyici örnekler S_a , Hizalama için en fazla tur sayısı K .

for $k = 1$ to K **do**

$O_k \leftarrow \text{LLM-hizalama}(C, I_{k-1}, S_a)$

if BoşMu(O_k) **then**

break

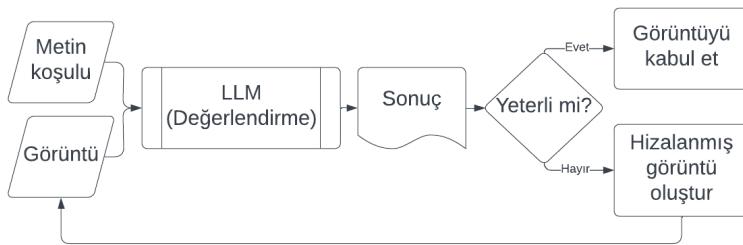
end if

$I_k \leftarrow \text{Yerleşimli-T2I}(C, O_k)$

end for

$I = I_{k-1}$

Çıktı: Görüntü I .



Şekil 5.14 LLM-akıl-yürütmeye ile görüntüyü hizalama

Tablo 5.4 LLM-akıl-yürütmeye ile görüntüyü hizalama algoritması

Girdiler: Metin koşulu C , Önceden üretilen görüntü I_0 , Hizalama için destekleyici örnekler S_a , Akıl yürütme için destekleyici örnekler S_e , Hizalama için en fazla tur sayısı K .

for $k = 1$ to K **do**

$R_{k-1} \leftarrow \text{LLM-akıl-yürütmeye}(C, I_{k-1}, S_e)$

if İyiMi(R_k) **then**

break

end if

$O_k \leftarrow \text{LLM-hizalama}(C, I_{k-1}, S_a)$

$I_k \leftarrow \text{Yerleşimli-T2I}(C, O_k)$

end for

$I = I_{k-1}$

Çıktı: Görüntü I .

LMD+ ve LayoutGPT için GLIGEN v2.1 kullanılmıştır. Şekil 5.15 nitel sonuçları göstermektedir. Nesnelerin etkileşimlerine dikkat çekildiğinde, etkileşimleri daha iyi yansıtan yerleşim bilgileri elde edilmektedir. Dahası, yerleşim bilgisini düzeltmek, nesne ilişkilerine daha fazla dikkat ederek daha iyi bir anlamsal

tutarlılık sağlamaktadır. Mevcut yöntemler etkileşimde olan nesneleri tek bir sınırlayıcı kutuya yerleştirdiği için eksik nesne problemi ile karşılaşmak mümkündür. Örneğin (c)'de LMD+ tek bir kutuya "muz ağacı" yerleştirmiştir ve ağacı oluşturamamıştır. Önerilen LLM komutlarında de her nesne için ayrı bir sınırlayıcı kutu oluşturmanın gerekliliği açık bir şekilde ifade edildiğinden, eksik nesne sorunu ortaya çıkmamıştır.

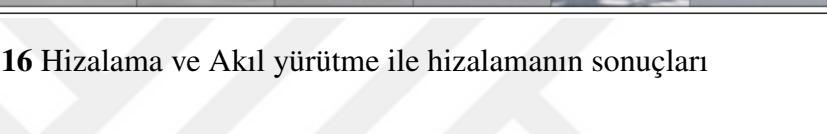
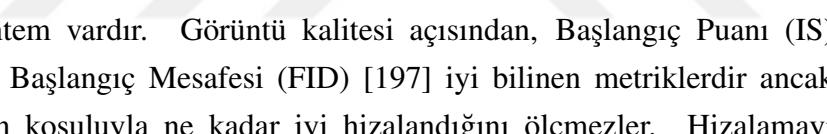
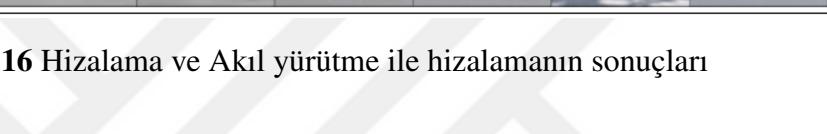
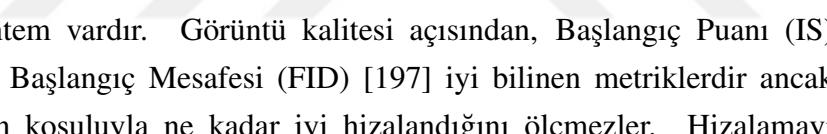
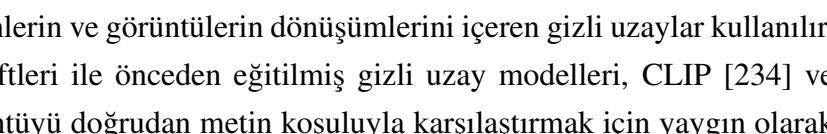
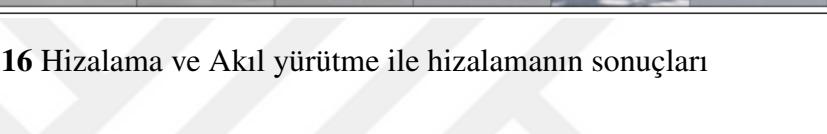
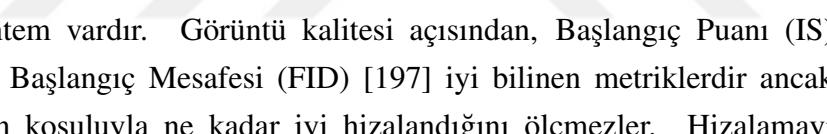
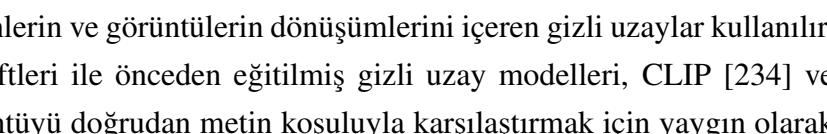
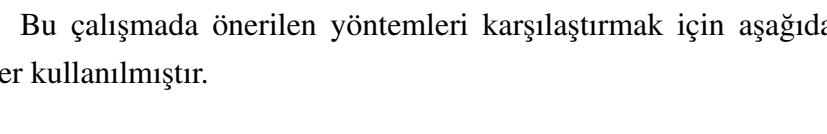
	Stable Diffusion	LMD+	LayoutGPT	Yerleştirme (bizimki)	Kendini Düzelten Yer. (bizimki)
(a) A man is standing on a street corner and waiting for a bus.					
(b) A man is sitting under a tree while a bird is in the tree.					
(c) A person is picking bananas from a tree.					
(d) A boy is shooting a soccer ball to goal.					

Şekil 5.15 Mevcut yöntemlerle karşılaştırma

5.3.2.2 Hizalama & Akıl yürütme

Şekil 5.16 görüntünün hizalanmasının nitel sonuçlarını göstermektedir. Yerleştirme ve düzeltme yöntemleri her ne kadar nesneler ve ilişkiler açısından tutarlı olsa da, tüm sahne dikkate alındığında uyumsuz görünmektedirler. örn. (a)'da öğrencinin defter ve tahtaya ilişkisi tamamlanmamıştır, (b)'de baba ve kız yürüyormuş gibi değil, oynuyormuş gibi görülmektedir. Bu nedenle görüntüleri hizalayarak daha uyumlu sahneler üretilebilir. Ayrıca, düşünce zinciri ile akıl yürütme olmadan hizalama yapıldığında bazı ayrıntıların göz ardı edildiği görülmektedir. örn. (a)'da defter kapalıdır, (b)'de baba ve kız gereksiz bir şekilde uçurtmayla ilgilenmektedir. Görüntü değerlendirilerek hizalama yapıldığında anlamsal hizalamanın daha da iyileştirildiği görülebilir. Hatta (c)'de neredeyse gerçek bir at üretilmiş ve en iyi

olarak değerlendirilmiştir.

	Temel		Hizalama		Düşünce Zinciri ile Hizalama	
	Yerleştirme	Kendini Düzelten Yer.	Yerleştirme	Kendini Düzelten Yer.	Yerleştirme	Kendini Düzelten Yer.
(a) A student is writing notes to their notebook from the board.						
(b) A father is walking with his daughter while a kite flies in the sky.						
(c) A child is on a rocking horse and a cat is walking by.						

Şekil 5.16 Hizalama ve Akıl yürütme ile hizalamanın sonuçları

5.3.3 Nicel Deneyler

Değerlendirme metrikleri. Metinden görüntü üreten modellerin performansını ölçen birkaç yöntem vardır. Görüntü kalitesi açısından, Başlangıç Puanı (IS) [194] ve Frechet Başlangıç Mesafesi (FID) [197] iyi bilinen metriklerdir ancak görüntünün metin koşuluyla ne kadar iyi hizalandığını ölçmezler. Hizalamayı ölçmek için metinlerin ve görüntülerin dönüşümlerini içeren gizli uzaylar kullanılır. Metin-görüntü çiftleri ile önceden eğitilmiş gizli uzay modelleri, CLIP [234] ve BLIP [247] görüntüyü doğrudan metin koşuluyla karşılaştırmak için yaygın olarak kullanılmaktadır. Bu çalışmada önerilen yöntemleri karşılaştırmak için aşağıda listelenen metrikler kullanılmıştır.

1. CLIPScore [248] : Oluşturulan görüntünün ve verilen metin koşulunun gizli dönüşümleri CLIP gizli uzayından çıkartılır ve bunların kosinüs benzerliği hesaplanır.
2. BLIP-CLIP [223] : Öncelikle önceden eğitilmiş BLIPv2 [249]'den oluşturulan görüntü için bir açıklama alınır ve ardından CLIP gizli uzayında bu açıklama metninin verilen metin koşuluyla kosinüs benzerliği hesaplanır.
3. BLIP-VQA [250] : Tüm soruyu tek bir soru olarak sorup "evet" yanıtını alma olasılığını bulmak yerine, karmaşık metin koşulunu açık bağımsız

sorulara ayıran ve her soru için "evet" yanıtını alma olasılığını çarpan, çözülmüş(disentangled) BLIP-VQA kullanılmıştır.

4. Çok kipli LLM: Değerlendirme için GPT-4o [251]'nun görüntü-dil capraz-kip anlama yeteneği [250]'deki yöntemle kullanılmıştır.

Tablo 5.5'te, "Bir kadın, bir kurabiye tepsisi tutuyor ve misafirlere uzatıyor." metin istemi için nicel değerlendirme metriklerinin puanları verilmiştir. 3 nesne (kadın, kurabiye tepsisi, misafirler) ve 2 etkileşim (kadın tepsiyi tutar, kadın misafirlere uzatır) içeren bir metin koşuluyla görüntüleri nasıl puanladığı incelenmektedir. Burada aynı metin şartının farklı oranlarda karşılandığı görüntüler üzerinde değerlendirme metriklerinin puanlamaları görülmektedir. CLIPScore, nesneler açıkça görülebildiğinde daha yüksek puanlar verir, ancak ilişkilerin içerilmediğini hesaba katmaz (bkz. görüntü 3&5). BLIP, karmaşık sahneler için uzun ve ayrıntılı açıklamalar sağladığından, bu açıklamalar genellikle kısa ve net olan metin koşulundan uzak olabilir. Bu nedenle BLIP-CLIP, tüm nesneleri içeren sahnelerle daha düşük puanlar vermektedir (bkz. görüntü 2&4). BLIP-VQA'ya gelince, CLIPScore'da gözlemlenen sorun burada da görülmektedir. Hatta BLIP-VQA, tüm nesneleri içeren sahnelerle daha düşük puanlar verirken, eksik nesnelerin olduğu sahnelerle daha yüksek puanlar vermektedir (bkz. görüntü 3&4). Çok kipli LLM değerlendirmelerine gelindiğinde, değerlendirmelerin çoğunlukla makul olduğu görülmektedir. Hatta baştan 2. görüntüde, LLM pembe şapkalı kişinin bir hizmetçi olduğunu çıkarsayarak sahnede misafir olmadığı için 4. ve 6. görüntülere göre daha düşük bir puan vermiştir.

Tablo 5.5 Farklı metriklerin nicel değerlendirme puanları

						
CLIPScore	0.2629	0.3381	0.3227	0.3178	0.2912	0.3093
BLIP-CLIP	0.873	0.784	0.846	0.761	0.901	0.672
BLIP-VQA	0.3268	0.3676	0.3908	0.1293	0.061	0.4311
GPT-4o [250]	0.64	0.76	0.76	0.8	0.72	0.8

Deney Kurulumu. Her bir metin istemi için 2 görüntü örneği üretilerek toplamda 50 görüntü çıktıSİ elde edilmiştir. Önerilen yöntemlerle birlikte Stable Diffusion v2.1, GLIGEN v2.1 kullanan LMD+ [220] ve LayoutGPT [219] yöntemleri karşılaştırmaya alınmıştır.

Sonuçlar. Tablo 5.6'da, önerilen yerleştirme ve düzeltme yöntemleri mevcut yöntemlerle karşılaştırılmaktadır. Önerilen yöntemlerin sonuçlarının tüm metriklerde mevcut yöntemlerden daha yüksek olduğu gözlemlenmiştir. CLIPScore

ve BLIP-VQA metrikleri yalnızca nesneleri içermeye odaklanır. Bu nedenle, oluşturulan görüntüler ilişkileri içeriyor olmasa bile, nesneleri açıkça içeriyor olmaları ile daha yüksek puanlar almıştır. Bu nedenle düzeltme yönteminin görüntüleri bu metriklerde daha düşük puanlar almıştır. Bunlardan farklı olarak BLIP-CLIP ve Çok kipli LLM, oluşturulan görüntüyü yorumlar ve değerlendirmede tüm yorumu kullanır. Bu, ilişkilerin var olup olmadığını daha iyi belirlemeye yardımcı olur. Böylece, ilişkilerin oluşturulduğu kendini düzeltten yerleştirme yönteminin görüntülerinde daha yüksek sonuçlar elde etmek mümkün hale gelmiştir.

Tablo 5.6 Yerleştirme ve düzeltme yöntemlerinin kıyaslamaları

	Stable Diffusion	LMD+	LayoutGPT	Yerleştirme	Düzeltme
CLIPScore	0.2878	0.3205	0.3123	0.3214	0.3151
BLIP-CLIP	0.8028	0.8045	0.8057	0.8177	0.8281
BLIP-VQA	0.4269	0.6816	0.6242	0.6869	0.6591
GPT-4o [250]	0.4216	0.5992	0.6392	0.6520	0.6816

Tablo 5.7 Hizalama ve akıl yürüterek hizalama yöntemlerinin kıyaslamaları

	Hizalama		Akıl yürütme ile Hizalama	
	Yerleştirme	Kendini düzeltme	Yerleştirme	Kendini düzeltme
CLIPScore	0.3132	0.3184	0.3132	0.3179
BLIP-CLIP	0.8203	0.8386	0.8180	0.8294
BLIP-VQA	0.6751	0.6558	0.6492	0.6678
GPT-4o [250]	0.7	0.7544	0.7664	0.7504

Tablo 5.7’de görüntünün hizalanmasının etkisi gösterilmiştir. CLIPScore ve BLIP-VQA metrikleri, nesnelerin tek tek görülebildiği görüntülere daha yüksek puanlar vermektedir. İlişkili nesneler Hizalama & Akıl yürüterek hizalama yöntemlerinin görüntülerinde iç içe geçmiş olduğundan, bu metriklerden daha düşük puanlar almışlardır. Bunun yanı sıra, BLIP-CLIP ve Çok kipli LLM tabanlı metrikler, üretilen görüntünün hizalanmasıyla daha iyi sonuçların elde edildiğini göstermektedir.

Akıl yürüterek hizalama yönteminin ürettiği görüntülerin daha doğru olduğunu ve bunun insan bakış açısından kolayca anlaşılabilceğini nitel sonuçlarımızda göstermiştık. Burada Çok kipli LLM’in insan bakış açısına benzer olduğu yani, yerleşimi düzeltilmiş görüntülerin yerleştirilmiş görüntülerden daha yüksek puan aldığı, hizalanmış görüntülerin yerleşimi düzeltilmiş görüntülerden daha yüksek puan aldığı ve akıl yürüterek hizalanmış görüntülerin hizalanmış görüntülerden daha yüksek puan aldığı görülmektedir.

6 ÇOK KİPLİ BÜYÜK DİL MODELİ REHBERLİĞİNDE METİNDEN GÖRÜNTÜ ÜRETİMİ

Metinden görüntü üreten difüzyon modelleri hem araştırma hem de uygulama alanlarında büyük başarılar elde etmiştir. Avantajlarına rağmen gerçek yaşam sahnelerinde karmaşık ilişkiler oluştururken sıkılıkla zorlanırlar. Öte yandan, büyük dil modellerinin metin istemleri ve görsel girdileri anlama konusundaki yetenekleri kanıtlanmıştır. Ancak, metinden görüntü üretimindeki potansiyelleri henüz tam olarak keşfedilmemiştir. Bu bölümde yapılan çalışmada, dikkatlice tasarlanmış adımlarda Çok kipli LLM'lerin(Multimodal LLM-MLLM) çeşitli yeteneklerinden yararlanan ve metinden görüntü üretiminde anlamsal uyumluluğu iyileştiren yeni bir eğitimsiz işlem hattı sunulmaktadır. Önerilen yöntemin mevcut çalışmalardan önemli ölçüde daha iyi performans gösterdiği ve hatta son teknoloji devleriyle rekabet ettiği Şekil 6.1'de gösterilmiştir. Ayrıca, gerçek yaşam sahnelerinden çoklu nesne ilişkileri içeren yeni bir kıyaslama veri kümesi tanıtılmıştır.



Şekil 6.1 MLLM Rehberliğinin son teknoloji devlerini geride bırakması

6.1 Giriş

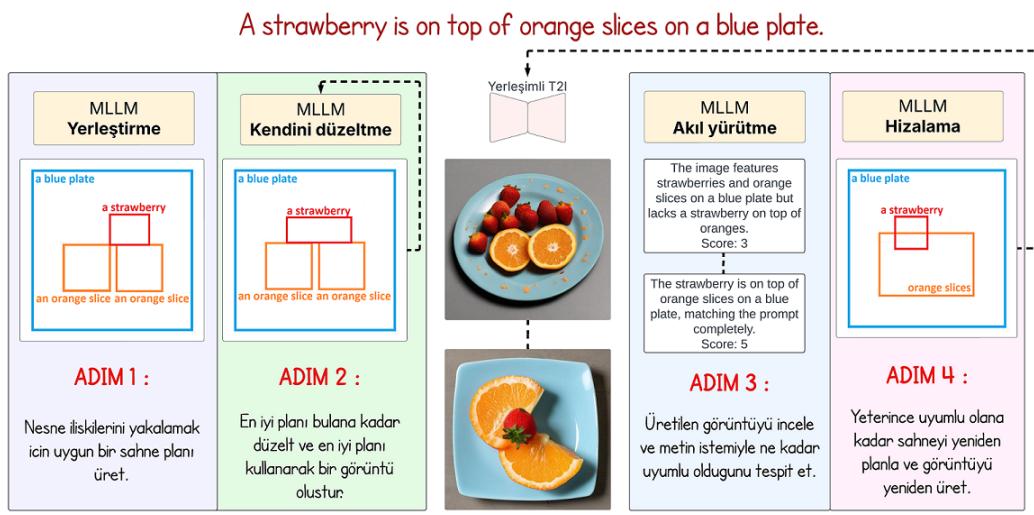
Metinden görüntüyü üretimi, son yıllarda difüzyon modellerinin gelişmesiyle daha da iyileşen heyecan verici bir araştırma konusudur. Metinden görüntüyü üreten difüzyon modelleri [102, 186, 187], istikrarlı öğrenme hedefleri ve büyük ölçekli eşleştirilmiş veri kümeleri üzerinde eğitilmeleri nedeniyle yüksek görüntü kalitesi ve anlamsal tutarlılık göstermektedir. Başarılarına rağmen, nesne ilişkilerini yakalarken eksiklikleri de vardır. Yerleşim girdisini metin girdisi ile beraber alan yerleşimli metinden görüntüyü üretimi [215–217] bu sorunu kısmen aşabilir. Bu gelişmeler, belirli sayıda nesne ve öznitelik söz konusu olduğunda yeteneklerini göstermiştir. Ancak, bu ilişkiler alışılmadık, çok sayıda ve karmaşık olabileceğinden nesne ilişkilerini yakalamak başlı başına bir araştırma konusudur.

Konuyu daha iyi anlayabilmek için nesne ilişkilerini iki başlık altında inceleyebiliriz; konumsal(spatial) ve konumsal olmayan(non-spatial) ilişkiler. Konumsal ilişkiler nesnelerin konumları hakkında bilgi içerir; içinde, üstünde, altında, yanında, yakınında, solunda, sağında, üstünde, altında, önünde, arkasında vb. Nesneler yalnızca bölgesel bir ilişkiye sahip olduğundan, yerleşim bilgisine güvenilebilir. Konumsal olmayan ilişkilerde bir nesnenin başka bir nesneyle etkileşimi vardır (örneğin bir kadın şemsiye tutuyor, bir çocuk topla oynuyor). Bu tür bilindik sahneler ilk modeller tarafından üretilebilirken, basit ama alışılmadık ilişkileri üretmek son çıkan modeller için bile zor olmaktadır. Temel yerleşim bilgileri yalnızca nesnelerin isimlerini ve sınırlayıcı kutu koordinatlarını içerdiginden bu bilgi çoğu zaman içeriği tam olarak tasvir edememektedir. Bu nedenle yerleşim kipini eklemek de bu görev için muhtemelen yeterli olmayacağındır. Bu noktada, olası bir yol, ilişkinin var olup olmadığını kontrol etmek için oluşturulan görüntüyü incelemektir.

Son yıllarda, Büyük Dil Modelleri (LLM'ler) çeşitli akıl yürütme görevlerinde önemli gelişmeler göstermiştir. Bazı çalışmalar [218–220] konumsal ilişkileri yakalamak için yerleşim bilgisi elde etmek amacıyla LLM'lerin doğal dilleri analiz etme ve anlama yeteneklerini kullanmıştır. Ancak, bu görevdeki tam performansları henüz ortaya çıkmamıştır.

Bu çalışmada, metinden görüntüyü üretimi için yeni bir eğitsiz işlem hattı(pipeline) önerilmektedir. Önerilen işlem hattı, görüntü oluşturulmadan önce 2 adım ve görüntü oluşturulduğundan sonra 2 adımdan oluşmaktadır. Görüntü oluşturulmadan önce, ilk adımda istenen görüntü için bir sahne düzeni oluşturmak üzere MLLM'den yararlanılır. İkinci olarak, MLLM'nin kendi kendini düzeltme yeteneğiyle nesneler arasındaki etkileşimler güçlü bir şekilde yakalanır. Ardından, donmuş(frozen) bir yerleşimli metinden görüntüyü üretimi (Yerleşimli-T2I) modeli kullanılarak bir

görüntü oluşturulur. Sonrasında, üçüncü adımda görüntünün istenen metinle uyumlu MLLM ile düşünce zinciriyle akıl yürütme tekniği kullanılarak ortaya koyulur. Son olarak, akıl yürütme sonucuna göre, MLLM görüntüyü inceler ve görüntü metin başlığıyla iyi bir şekilde hizalanana kadar düzeni yeniden tasarlar. Bu adımların, oluşturulan görüntünün anlamsal hizalamasını kademeli olarak iyileştirdiği gözlemlenmiş ve bir MLLM'in çeşitli yetenekleriyle metinden görüntü üretimini yönlendirdiği *MLLM Rehberliği* tanıtılmıştır. MLLM Rehberliği Şekil 6.2'de sunulmuştur.



Şekil 6.2 MLLM Rehberliğinin 4 adımı

Nitel sonuçlara ek olarak, T2I-CompBench[250] görevlerinde nicek kıyaslama(marks) gerçekleştirılmıştır. Dahası, gerçek yaşam sahnelerinden birden fazla eylem ve etkileşim içeren yeni bir kıyaslama veri kümesi önerilmiştir. Önerilen yöntem, daha az kaynak gerektirmesine rağmen mevcut çalışmalarдан önemli ölçüde daha iyi performans göstermiştir ve hatta Stable Diffusion 3.5-large ve FLUX.1 gibi son teknoloji devleriyle rekabet etmektedir. Çalışmanın katkıları aşağıda listelenmiştir:

- Önerilen iyi tasarlanmış MLLM rehberliğinden yararlanarak, herhangi bir ek eğitim gerektirmeden görüntü-metin hizalaması iyileştirilmiştir.
- Önerilen işlem hattı, çok daha küçük bir modelden yararlanması rağmen, en son teknoloji Stable Diffusion 3.5-large ve FLUX.1'e meydan okuyarak parametre verimliliği vaat etmektedir.
- T2I-CompBench görevlerinin yanı sıra, gerçek yaşam sahnelerinden birden fazla eylem ve etkileşim içeren yeni bir kıyaslama veri seti önerilmiş

ve MLLM rehberliğinin mevcut çalışmalara göre üstün performansı gösterilmiştir.

- Rehberliğin her bir adımında MLLM’nin çeşitli yeteneklerinin etkisini anlamak için MLLM rehberliğinin ablasyon çalışması gerçekleştirılmıştır.
- Önerilen işlem hattı, çeşitli MLLM’lere ve metinden görüntüyü üreten modellere kolayca uyarlanarak, genellemeye yeteneği gösterilmiştir.

6.2 İlgili Çalışmalar

Metin koşullu görüntüler ilk olarak GAN’lar [2] tarafından üretilmiş ve büyük miktarda eşleştirilmiş veriyi güçlü hedef fonksiyonları ile öğrenebilen difüzyon modelleri [28] ile üretim kalitesi daha da gelişmiştir. Metinden görüntüyü üreten difüzyon modelleri [102, 179, 186, 187, 252] kısa sürede büyük bir yükseliş kazanmış ve artık her ölçekte araştırma ve endüstride kullanılabilir hale gelmiştir. Difüzyon modelleri hem görsel kalite hem de anlamsal hizalamada önemli ilerlemeler sağlasa da hesaplama maliyetleri uygulanabilirliği sınırlamaktadır. Bu bağlamda önerilen Gizli Difüzyon Modeli(LDM) [102], sadece model uzayını küçültmekle kalmamış, aynı zamanda farklı kipler arasında dönüşümün mümkün olabileceğini de göstermiştir. Son zamanlarda ise, çok kipli difüzyon dönüştürücüler [253, 254], dönüştürücü tabanlı mimariden yararlanarak metinden görüntüyü üretimini hem görüntü kalitesi hem de anlamsal tutarlılık açısından geliştirmiştir.

Metinden görüntüyü üreten difüzyon modelleri, etkileyici başarılarına rağmen, çok nesneli bir istemle karşılaşlıklarında hatta bu nesnelerin özellikleri ve ilişkileri var olduğunda bu karmaşık metin koşullarını oluşturmakta zorlanırlar. Bu gibi durumlarda, nesnelerin kesin konumu gibi metin girdisini destekleyen diğer girdilere sahip olmak elverişli olabilir. Yerleşimli metinden görüntü üretimi [215–217, 225–229] birçok durumda görüntü-metin hizalamasını iyileştirmiştir ancak bu modellere yerleşim bilgilerinin de girdi olarak verilmesi gereklidir ve bu bilginin her seferinde elle sağlanması ugraştırıcı bir durumdur.

Öte yandan, LLM’ler az atımlı(few-shot) öğrenme [230, 231], düşünce zinciriyle akıl yürütme [232] ve kendi kendini düzeltme [233] yetenekleriyle birçok alanda güçlerini kanıtlamışlardır. Yerleşim bilgisini üretmek için de LLM’lerden faydalanan bazı araştırmalar vardır [218–220, 255]. Üretilen görüntüler her ne kadar yerleşim bilgisine sadık kalsa da metin ve görüntü arasındaki anlamsal tutarlılığı tespit etmek zor bir problemdir. Son yıllarda ortaya çıkan LLM’lerin çok kipli bir uzayda farklı kipleri birleştirebilmeleri, görsel yorumlama yeteneğine

sahip olmalarını sağlamaktadır [234–238]. Bu bağlamda, T2I-CompBench [250] ve LLM-Score [241] oluşturulan görüntüleri değerlendirmek için görsel yorumlama yeteneğinden yararlanmıştır.

Konumsal kontrol yeteneğinin eklenmesi görüntülerin anlamsal tutarlığını artırırken, LLM’lerin yerleşim bilgisini elde etmek için kullanılması birçok görevde bu yaklaşımı kolaylaştırmıştır. Ödül modelinden gelen geri bildirimle görüntü tutarlığını artıran [239] ve görsel soru cevaplama yoluyla sahneyi kendi kendine düzeltten [240] bazı çalışmalar da mevcuttur. Ancak, LLM’lerin kendi çıktılarını düzeltmesi bu görev için daha önce araştırılmamıştır. Buna ek olarak, LLM’lerin görsel yorumlama yeteneğinin daha tutarlı bir görüntü elde etmek amacıyla kullanılması daha önce denenmemiştir. Son gelişmelere dayanarak, akıl yürütme yeteneklerinin eklenmesi görüntü-metin hizalamasını daha iyiye götürebilir. Bu çalışmada, Çok kipli LLM’lerin çeşitli yeteneklerinden yararlanarak metinden görüntü üretimine rehberlik yapması önerilmiştir. Metinden görüntü üretiminde herhangi bir ekstra modele ihtiyaç duymadan, aynı MLLM ile yerleştirme, kendi kendini düzeltme, akıl yürütme, hizalama ve değerlendirme yapılması önerilmiştir.

6.3 Yöntem

Verilen bir metin koşulu C için, yerleşim bilgisi üretmek $O = \{o_j; j = 1, 2, \dots, n\}$ yerleştirme birimlerinin bir kümesini tahmin etmektir; burada her o_j yerleştirme birimi, j nesnesinin bilgisini belirtir. o_j , bir d_j tanımı ve b_j sınırlayıcı kutu bilgisinden oluşur. Yani $o_j = \{d_j, b_j\}$. $b_j = \{x_j, y_j, w_j, h_j\}$ şeklinde gösterilebilir; burada x_j , y_j sol üst noktayı ve w_j , h_j sırasıyla genişliği ve yüksekliği belirtir. Bu şekilde, yerleşim bilgisi metinsel olarak ifade edilebilir. Dolayısıyla, herhangi bir MLLM ile C metin koşulu komutta verilerek bir yerleşim bilgisi üretilebilir. Metin koşuluna ek olarak, önceden belirlenmiş sabit bağlam içi(in-context) örnekler verilerek yerleşim bilgisi desteklenir.

MLLM rehberliği 4 ardışık bileşenden oluşur: İlk bileşen, MLLM’nin istenen sahne için uygun bir yerleşim planı oluşturduğu yerleştirme, ikinci bileşen, MLLM’nin oluşturulan yerleşim planını ayarladığı kendini düzeltme, üçüncü bileşen, MLLM’nin oluşturulan görüntünün uyumluluğunu bulduğu akıl yürütme ve son bileşen, MLLM’nin oluşturulan görüntüyü inceleyip akıl yürütme sonucuna göre sahneyi yeniden tasarladığı hizalamadır. Her adım için komutların tamamı Ek A’da mevcuttur.

6.3.1 Yerleştirme

Sahne yerleşimi metinsel bir çıktı olarak gösterilebildikten sonra, verilen metin koşulunu analiz edip ona göre bir sahne yerleşimi oluşturmak amacıyla MLLM'lerin dil yeteneklerinden yararlanılabilir. Nesne ilişkilerini daha iyi yakalamak adına yerleşim planı oluştururken için nesneler arasındaki etkileşimlere vurgu yapılmıştır. MLLM-yerleştirme Eşitlik 6.1'de verilmiştir.

$$O \leftarrow \text{MLLM}-\text{yerleştirme}(C) \quad (6.1)$$

6.3.2 Kendini Düzeltme

Kendi düzeltme admında, komut oluşturulan yerleşim planı ile yeniden yapılandırılır ve gerekirse ayarlanması istenir. Kendini düzeltme admında da metin istemindeki etkileşimlere dikkat çekilmiştir. MLLM-kendini-düzeltme, Eşitlik 6.2'de verilmiştir.

$$O_{yeni} \leftarrow \text{MLLM}-\text{kendini}-\text{düzeltme}(C, O_{eski}) \quad (6.2)$$

6.3.3 Akıl Yürütmeye

Görüntüde nesne ilişkilerinin var olup olmadığı ancak oluşturulan görüntüyü inceleyerek belirlenebilir. Burada, nesne etkileşimlerini tespit etmek için MLLM'nin düşünce zinciri ile akıl yürütme yeteneğinden yararlanılmıştır. MLLM akıl-yürütmeye Eşitlik 6.3 'te verilmiştir. Burada I görüntüyü ve R akıl yürütmenin sonucunu ifade etmektedir.

$$R \leftarrow \text{MLLM}-\text{akıl}-\text{yürütmeye}(C, I) \quad (6.3)$$

6.3.4 Hızalama

MLLM-hızalama bileşeninde oluşturulan görüntü akıl yürütme sonucuna göre incelenir. Görüntü iyi hizalanmamışsa, MLLM metin koşulunu sağlamak için sahneyi yeniden tasrarlar. MLLM-hızalama Eşitlik 6.4'te gösterilmiştir.

$$O_{yeni} \leftarrow \text{MLLM}-\text{hızalama}(C, I_{eski}) \quad (6.4)$$

Tablo 6.1, MLLM rehberliği için işlem hattını göstermektedir. Kendini düzeltme

aşamasında, doğru yerleşim planının elde edildiğinden emin olmak için en fazla K tur sayısı talep edilmiştir. Yerleşim planı ayarlandıktan sonra, istenen görüntüyü oluşturmak için donmuş bir yerleşimli metinden görüntü üretimi modelinden (Yerleşimli-T2I) yararlanılır. $iyiMi(R)$, görüntünün metin koşuluyla iyi hizalanıp hizalanmadığını belirleyen koşuludur. Sonunda, K tur içinde MLLM akıl-yürütme adımlına göre en iyi üretilen görüntü kabul edilir.

Tablo 6.1 MLLM rehberliği algoritması

Girdiler: Metin şartı C , en fazla tur K

Yerleştirme
 $O_0 \leftarrow \text{MLLM}-\text{yerleştirme}(C)$

Kendini düzeltme
 $k \leftarrow 0, O_1 \leftarrow \text{null}$

while $k < K$ **and** $O_k \neq O_{k+1}$ **do**

$O_{k+1} \leftarrow \text{MLLM}-\text{kendini-düzelme}(C, O_k)$

$k \leftarrow k + 1$

end while

$I_0 \leftarrow \text{Yerleşimli-T2I}(C, O_{k+1})$

Akil yürütmeye
 $R_{eniyi} \leftarrow 0$

for $k = 0$ **to** K **do**

$R_k \leftarrow \text{MLLM}-\text{akıl-yürütme}(C, I_k)$

if $iyiMi(R_k)$ **then**

$I_{eniyi} \leftarrow I_k$

break

else if $R_k > R_{eniyi}$ **then**

$R_{eniyi} \leftarrow R_k$

$I_{eniyi} \leftarrow I_k$

end if

Hızalama
 $O_{k+1} \leftarrow \text{MLLM}-\text{hızalama}(C, I_k)$

$I_{k+1} \leftarrow \text{Yerleşimli-T2I}(C, O_{k+1})$

end for

Çıktı: Görüntü I_{eniyi} .

6.4 Deneyler

6.4.1 Deney Kurulumu

6.4.1.1 Modeller

Yöntemimizin doğası gereği, hem görüntü oluşturma hem de rehberlik için herhangi bir ek eğitim gerekmektedir. Bu bağlamda her iki tarafta da önceden eğitilmiş hazır modeller kullanılabilir. Rehberlik için görsel yorumlama yeteneği olan

bir MLLM gerektiğinden, bu konuda GPT-4o [256] kullanılmıştır. Öte yandan, yerleşimli metinden görüntü üretimi modelleri, eğitildikleri veri kümeleri ve ince ayar stratejileri nedeniyle kendi özelliklerine sahiptir. Bu çalışmada, literatürde sıkça geçen GLIGEN [215] kullanılmıştır. GLIGEN'in ilk sürümü zayıf görsel kaliteye sahip olduğundan, IGLIGEN adaptörleri [257] ile eğitilen Stable Diffusion XL sürümünü [258] kullanılmıştır. Deneylerde, en fazla tur sayısı K=5 alınmış ve iyiMi(R) koşulu 5 puanı sağlayacak şekilde ayarlanmıştır.

6.4.1.2 Kıyaslamalar

Nesne ilişkileri bağlamında, T2I-CompBench [250] görevleri incelenmiş ve konumsal(spatial), 3 boyutlu konumsal(3d-spatial), konumsal olmayan(non-spatial), karmaşık(complex) kıyaslamaları deneylere dahil edilmiştir. Ek olarak, her biri 3 veya 4 nesne ve 1 veya 2 ilişki içeren gerçek yaşam sahneleri içeren 300 altyazıyla sahip yeni bir kıyaslama veri kümesi olan RLS-300 önerilmiştir. Gerçek yaşam sahnelerinde bazı tanık ilişkiler (örn. Bir öğrenci kulaklık takıyor ve bilgisayardan çalışıyor, Bir adam bir ağaçın altında oturuyor ve bir kuş ağaçta.) metinden görüntü üreten ilk modeller tarafından zahmetizce üretililebilirken, birçok gerçek yaşam sahnesi, en son teknoloji modellerle bile yakalanması zor olan birden fazla nesne ilişkisi içerir (örn. Bir öğrenci not defterine tahtadan notlar yazıyor, Bir kadın şemsiye tutuyor ve bir otobüse doğru koşuyor.) Bu kıyaslamada, modellerin birbirlerine olan üstünlüğünü görebilmek için her iki türden metin istemi de mevcuttur. Tüm veri seti Ek B'de verilmektedir.

6.4.1.3 Değerlendirme metrikleri

Görüntü kalitesi açısından, Başlangıç Puanı (IS) ve Frechet Başlangıç Mesafesi (FID) iyi bilinen metriklerdir ancak görüntünün metin koşuluyla ne kadar iyi hizalandığını ölçmezler. Önceden eğitilmiş gizli modeller, CLIP [234] ve BLIP [247] görüntüyü doğrudan metin koşuluyla karşılaştırmak için yaygın olarak kullanılmaktadır. Bu çalışmada, CLIPScore [248], BLIP-CLIP [223], ayrıstırılmış(disentangled) BLIP-VQA [250] ve [250]'de önerilen değerlendirme komutuyla GPT-4o [251] ele alınmıştır. Her değerlendirme metriğinde birden fazla nesne ilişkisi içeren metin istemleri için 50 görsele ait puanlar alınmış ve sonuçları incelenmiştir.

Değerlendirme metriklerini araştırmak için, 50 görüntü bizzat puanlanmış (tezin yazarı) ve her metriğin Kendall'ın τ ve Spearman'ın ρ ölçekleri açısından insan değerlendirmeleri ile korelasyonları Tablo 6.2'de incelenmiştir. GPT-4o değerlendirmelerinde daha yüksek korelasyonlar gözlemlenmiş ve bu nedenle,

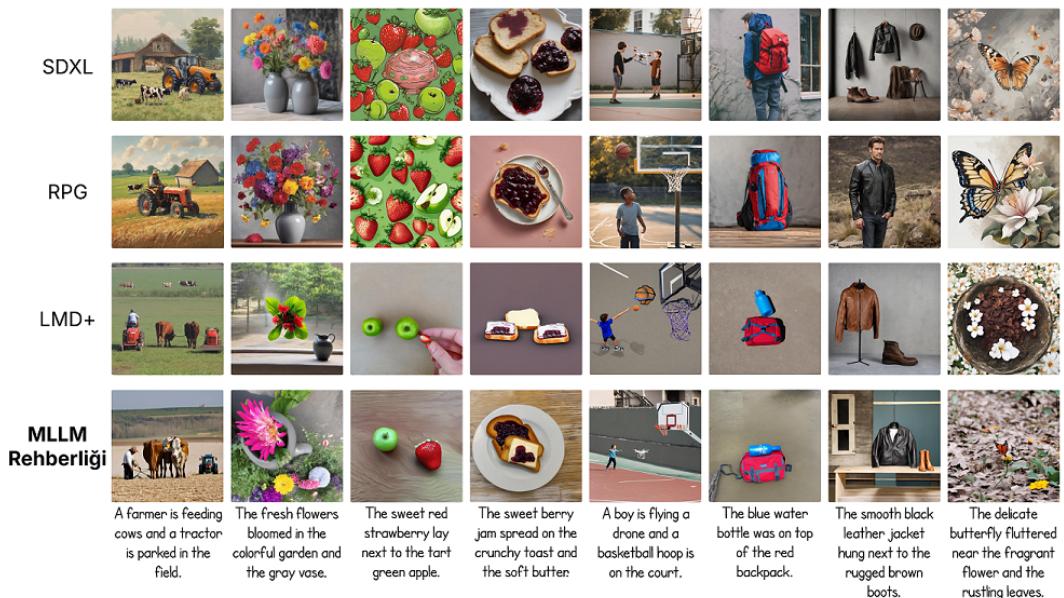
üretilen görüntüleri değerlendirmek için nicel deneylerde GPT-4o kullanılmıştır.

Tablo 6.2 Metriklerin insan değerlendirmeleri ile korelasyonları

	$\tau(\uparrow)$	$\rho(\uparrow)$
CLIPScore	0.2435	0.3017
BLIP-CLIP	0.2122	0.2759
BLIP-VQA	0.2444	0.3155
GPT-4o	0.583	0.6563

6.4.2 Nitel Deneyler

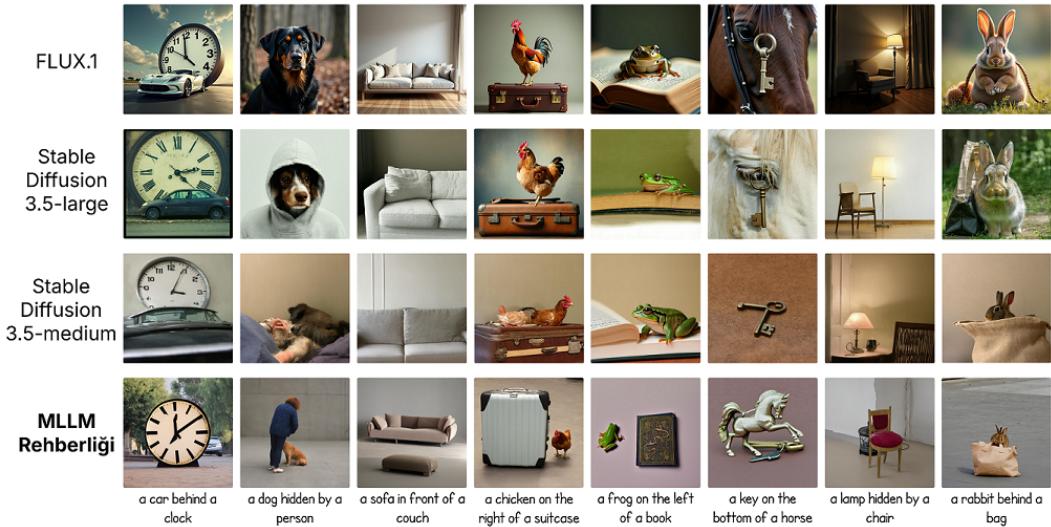
Nitel deneyler 2 kısım olarak verilecektir. Öncelikle, temel Stable Diffusion XL(SDXL) [252] ve literatürdeki benzer çalışmalar, RPG [255] ve LMD+ [220] önerilen yöntemle karşılaştırmalı olarak incelenmiştir. Şekil 6.3'teki her metin isteminde, niteliklere ve/veya ilişkilere sahip birden fazla nesne bulunmaktadır. MLLM rehberliği, metin istemindeki tüm talepleri karşılayarak önceki çalışmalarдан önemli ölçüde üstün olduğunu göstermiştir. RPG ve LMD+'nın eylemleri ve etkileşimleri yerine getirmede zayıf olduğu gözlemlenmiştir. Buna karşılık, MLLM rehberliği istenen sahneyi etkili bir şekilde planlayıp düzelterek ve sonrasında akıl yürütüp sahneyi yeniden planlayarak anlamsal tutarlılık açısından memnuniyeti artırmıştır. Mevcut çalışmalarla karşılaştırıldığında, önerilen işlem hattının gelişmiş anlamsal ifade yeteneklerinin üretim sonrası adımlarına, yani düşünce zinciri ile akıl yürütme ve hizalamaya bağlı olması kuvvetle muhtemeldir.



Şekil 6.3 MLLM Rehberliğinin SDXL ve benzer çalışmaları geride bırakması

Nitel deneylere son teknoloji FLUX.1[dev] [254], Stable Diffusion 3.5-large[253]

ve Stable Diffusion 3.5-medium [253] ile devam edilmiştir. Şekil 6.4’te, özellikle konumsal ve 3 boyutlu konumsal ilişkileri yakalamada MLLM rehberliğinin anlamsal uyumluluğu yakalamadaki başarısı görülmektedir. İlk 2 model parametre boyutu açısından çok daha büyük olmasına rağmen, nesnelerin göreceli konumlarını belirlemede başarısız olmuşlardır. Aksine, MLLM rehberliği metin istemini etkili bir şekilde planlayıp tatmin edici bir sonuç elde edilene kadar sahneyi yeniden düzenlemektedir.



Şekil 6.4 MLLM Rehberliğinin konumsal ve 3 boyutlu konumsal ilişkileri yakalamadaki üstünlüğü

6.4.3 Nicel Deneyler

MLLM Rehberliği, niceł deneylerde kapsamlı kıyaslama veri kümeleri üzerinde değerlendirilmiştir. Tablo 6.3 çeşitli model mimarilerinin ve mevcut çalışmaların görüntü-metin hizalama yeteneğini göstermektedir. Burada modeller mimarilerine göre ikiye ayrılmaktadır: geleneksel U-net tabanlı difüzyon modelleri (DM) ve son zamanlarda trend olan Çok Kipli Difüzyon DönüştürÜcÜleri (DiT). En iyi puanlar kalın, ikinci en iyi puanlar italic olarak gösterilmiştir. Parametre boyutları parantez içinde verilmiştir. Az kaynaklı MLLM Rehberliği tüm kıyaslamalarda birinci veya ikinci sırada yer alır.

MLLM Rehberliğinin, tüm kıyaslama ölçütlerinde aynı mimarideki tüm mevcut çalışmalarдан önemli ölçüde daha iyi performans gösterdiği görülmektedir. Literatürdeki diğer çalışmalar, konumsal olmayan ilişkiler içeren görevlerde (non-spatial, complex, RLS-300) temel modelin performansını düşürürken, önerilen yöntem üretim sonrası adımların etkisiyle anlamsal tutarlılık performansını önemli ölçüde artırmıştır.

Dönüştürücü tabanlı mimarilerin dil anlama gücüne rağmen, önerilen yöntem aynı boyuttaki Stable Diffusion 3.5-medium (SD 3.5-M)'u önemli ölçüde aşmıştır. Ayrıca, MLLM Rehberliği, konumsal ve 3 boyutlu konumsal ilişkileri yakalama bağlamında son teknoloji devleri Stable Diffusion 3.5-large (SD 3.5-L) ve FLUX.1'i önemli ölçüde geride bırakmış, karmaşık ve gerçek yaşam ilişkilerinde ise en iyi veya ikinci en iyi olmuştur.

Tablo 6.3 Kıyaslama sonuçları

Mim.	Model	Spatial	3d-spatial	Non-spatial	Complex	RLS-300
DiT	FLUX.1(12B)[254]	0.6925	0.6058	0.9108	0.8625	0.8525
	SD 3.5-L(8B)[253]	0.8108	0.6667	0.9375	0.8692	0.8275
	SD 3.5-M(2.6B)[253]	0.7267	0.6350	0.9067	0.8675	0.8083
DM	SDXL[252]	0.6908	0.5600	0.9208	0.7633	0.6892
	RPG[255]	0.6750	0.5333	0.8667	0.7500	0.6792
	LMD+[220]	0.7708	0.5900	0.8092	0.6725	0.6750
	MLLM Rehberliği	0.8392	0.7258	0.9217	0.8725	0.8517

6.5 Analiz

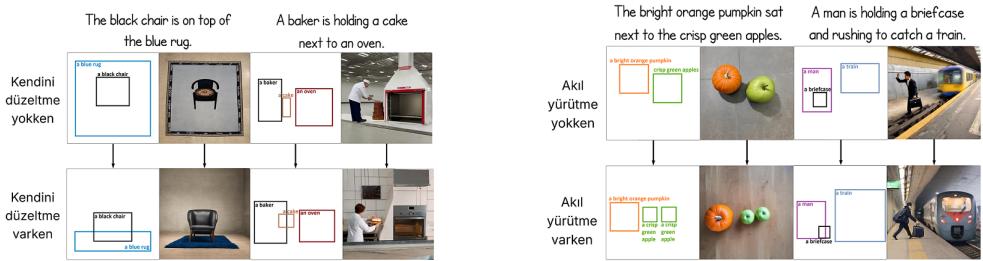
6.5.1 Ablasyon Çalışması

6.5.1.1 Kendini Düzeltmenin Etkisi

Ablasyon çalışmalarında ilk olarak, yerleşim bilgisini kendi kendine düzeltten ikinci adının etkisi Şekil 6.5 (a)'da incelenmiştir. İlk yerleşim bilgisi nesneler ve ilişkiler açısından tutarlı gibi görünse de, üretilen görüntüler incelendiğinde kendi kendini düzeltme adımı olmadan oluşturulan görüntülerin uyumlu olmadığı görülmektedir. Kendi kendini düzeltme yeteneğinden yararlanıldığında ilişkileri garantilemek için daha kesin yerleşim bilgisinin sağlandığı gözlemlenmiştir.

6.5.1.2 Akıl Yürütmeyenin Etkisi

Ablasyon çalışmalarının devamında, üçüncü adım olan düşünce zinciri ile akıl yürütmenin etkisi Şekil 6.5 (b)'de incelenmiştir. Burada, yerleşim bilgisi doğru ve tutarlı olmasına rağmen, düşünce zinciri ile akıl yürütme olmadan hizalama gerçekleştirildiğinde üretilen görüntülerin tutarsız olabileceği görülmüştür. Üçüncü adım, üretilen görüntünün metin istemiyle hizalamasını yorumlayarak daha tatmin edici görüntüler elde etmeyi sağlamaktadır.



Şekil 6.5 İkinci ve üçüncü adımların etkisini gösteren ablasyon çalışması

6.5.2 Genelleme Yeteneği

6.5.2.1 Farklı Rehberler

Önerilen yöntemi genelleştirmek için öncelikle farklı MLLM'lerin metinden görüntüyü üretimine rehberlik yapması incelenmiştir. Şekil 6.6 (a)'da, MLLM rehberliği işlem hattında açık kaynaklı bir MLLM Gemma-3-27b-it [259] ve başka bir amiral gemisi Gemini-2.5-flash-preview-04-17 [260]'yi kullanılmıştır. Burada, görüntülerin en yüksek puanı sahip olmasını sağlamak için akıl yürütme adımda en fazla tur sayısı genişletilerek daha küçük MLLM'lerin daha kolay ikna edildiği görülmüştür.

6.5.2.2 Farklı Görüntü Üreticiler

İkinci olarak MLLM rehberliği işlem hattında en son çıkan yerleşimli metinden görüntüyü üretimi modelleri olan InstanceDiffusion [261] ve Rich-context L2I [262]'nın kullanılması Şekil 6.6 (b)'de gösterilmiştir. Bu bağlamda, önerilen işlem hattının en son modellere kolayca uyarlanıldığı ve gelecekteki teknolojilerle uyumlu bir şekilde çalışmayı vaat ettiği gösterilmiştir.



Şekil 6.6 MLLM rehberliğinin çeşitli rehberler ve metinden görüntü üreten modellere genelleştirilmesi

6.6 Sonuç

Bu bölümde, metinden görüntü üretiminin anlamsal tutarlığını geliştirmek üzere Çok Kipli Büyük Dil Modelleri tarafından rehberlik yapılması önerilmiştir. Önerilen yöntem, herhangi bir model eğitimi/ince ayar gerektirmeden sadece giderek daha iyi hizalanmış bir görüntü elde etmeyi sağlayan dikkatlice tasarlanmış rehberlik adımlarıyla görüntü-metin hizalaması konusunda üstün bir başarı elde etmektedir. Yapılan deneylerde, önerilen yöntemin çeşitli kıyaslamalarda mevcut çalışmalardan önemli ölçüde daha iyi performans gösterdiği görülmüştür. Ek olarak, önerilen yöntem konumsal ve 3 boyutlu konumsal görevlerde son teknoloji devlerden daha üstün, karmaşık ve gerçek yaşam sahnelerinde ise rekabetçidir. Dahası, difüzyon dönüştürücülerindeki parametre sayısının etkisini açıkça gösteren yeni bir kıyaslama veri kümesi, RLS-300, önerilmiştir.

MLLM rehberliğindeki her adım daha iyi bir hizalama sağlar. Rehberlik adımları üretim öncesi ve üretim sonrası olarak ikiye ayrılsa, üretim sonrası adımların daha etkili olduğunu fark edilmiştir. Ayrıca yakın bir görüntüyü yorumlamadan, bir sonraki adım için daha iyi bir yerleşim bilgisi üretmeye yardımcı olabileceği gözlemlenmiştir. Bu nedenle MLLM akıl-yürütmenden sonra MLLM yerleştirme değil, görüntü yorumlanarak bir plan oluşturulan MLLM hizalama kullanılmıştır.

Burada bazı mevcut çalışmaların performansının hiperparametrelere güçlü bir şekilde bağlı olduğunu belirtmek gereklidir. Deneylerde gözlemlendiği üzere, temel istemin oranı(base prompt ratio), hiyerarşî sayısı gibi hiperparametrelere RPG [255]'nin performansını etkilemektedir. Deneylerde, önerilen oran ve LLM komutu kullanıldığından, bunlar tüm görevler ve tüm ilişki türleri için uygun olmayabilir. Dahası, nesne ilişkileri farklı alt bölgelerin birleştirilmesini gerektirir ancak bu RPG ile mümkün değildir. Mevcut çalışmalardan farklı olarak MLLM rehberliği herhangi bir hiperparametre ayarı gerektirmez, yalnızca ne kadar mükemmel hizalama tercih ettiğini gösteren en fazla döngü sayısı sabiti K ve $\text{iyiMi}(R)$ fonksiyonu ayarlanır.

Nesne ilişkileri dikkate alındığında sınırlayıcı kutulardan oluşan yerleşim planı bilgisi yerine piksel bazında yerleşim bilgisinin kullanılması düşünülebilir. MLLM'lerden sınırlayıcı kutu bilgisini elde etmek, her pikselin hangi nesneye ait olduğunu içeren segmentasyon haritaları sağlamaya göre zaman ve maliyet açısından çok daha fazla ekonomiktir. Yine de, GLIGEN için bilindik yerleşim bilgisi yerine, MLLM komutlarımız buna göre uyarlanabilir. Bu bilgiden yararlanmak için, ControlNet [263]'i MLLM rehberliği adımlarımızla birlikte kullanmak, istenen sahneleri oluşturmak için elverişli olacaktır.

MLLM'ler her gün gelişmekte ve yeni yetenekleri ortaya çıkarmaktadır. Bu çalışmadaki komutlarda sabit birkaç atımlı örnekler kullanılmıştır, ancak metin istemine yakın olan örnekler seçmek daha iyi bir yerleşim bilgisi sağlayabilir. Harici veritabanı bilgisinden yararlanmak için Bilgi getirmeyle zenginleştirilmiş üretim (Retrieval Augmented Generation-RAG) [264] uygulanabilir.

U-net tabanlı bir difüzyon modeli olan Stable Diffusion XL (SDXL) ile karşılaştırıldığında, aynı boyuttaki dönüştürücü tabanlı mimari modeli olan Stable Diffusion 3.5-Medium (SD 3.5-M) için özellikle karmaşık ve gerçek yaşam sahnelerinde önemli bir iyileşme görülmüştür. Bu gözleme göre, MLLM rehberliğinde dönüştürücü tabanlı mimarilerden yararlanmak, nesne ilişkilerini yakalamak adına daha etkili olacaktır.

7 SONUÇ

Metinden görüntü üreten modellerde görüntü-metin hizalamasını yeniden eğitim gerektirmeden artırmayı hedefleyen bu tez çalışması, fazladan donanımsal kaynak gerektirmeden iyileştirme yapmanın mümkün olduğunu göstermiştir. Tez kapsamında, üretilen görüntünün metin istemiyle uyumluluğu problemine odaklanılmış ve ek bir eğitim gerektirmeden modelin çıkışım zamanında yapılan geliştirmelerle daha iyi görüntüler elde edilmiştir.

İlk olarak çok kipli varyasyonlu otokodlayıcılarla metinden görüntü üretimi üzerine yapılan çalışmada kiplerin ağırlıklandırılmasının etkisi incelenmiştir. Kiplerin birleştirilmesi sırasında bazı kiplerin daha önemli olduğundan yola çıkarak hedef fonksiyonunda ağırlıklandırma yapılması önerilmiştir. Önerilen yöntemlerle mevcut yöntemlerin karşılaştırması yapılmış ve anlamlı sonuçlar elde edilmiştir.

Sonrasında yapay görüntü üretimi konusunda kısa sürede büyük bir ivme yakalayan difüzyon modelleri ile ilgili bir inceleme makalesi hazırlanmıştır. Difüzyon modellerinin temel çalışmalarını açıklayarak başlayan bu çalışmada mevcut incelemelerin aksine teorik gelişmelerin konularına göre kategorize edilmesine odaklanılmıştır. Gelişmelerin amaçlarına göre değil de konularına göre kategorize edilmesi her bir çalışmanın tek bir kategori altına girmesiyle net bir anlayış getirmiştir.

Tez çalışmasının devamında difüzyon modelleri ile metinden görüntü üretiminde görüntü-metin hizalamasını çıkışım zamanında iyileştirmeyi hedefleyen çalışmalar yapılmıştır. Bu kapsamında karmaşık metin istemlerinin parça parça üretimi için öncelikle çapraz-dikkat haritaları kullanılmış, sonrasında yerleşimli metinden görüntü üretimi modellerine odaklanılmış, sonunda ise büyük dil modeli rehberliğinde metinden görüntü üretimi konusunda çalışmalar yapılmıştır.

Büyük dil modelleri bir çok alanda akıl yürütme konusunda başarılı performanslar göstermiştir. Bu tez çalışmasında son olarak büyük dil modellerinin kendi cevabını

düzelme, düşünce zinciriyle akıl yürütme ve görsel yorumlama yeteneklerinden faydalalararak metinden görüntü üreten modellere rehberlik yapması önerilmiştir.

Sonuç olarak bu tez çalışması, farklı model ve tekniklerle, hedefinin gerçekleştirilebileceğini göstermiştir. Daha büyük veri/model kullanmadan ve ek bir donanım maliyeti gerektirmeden iyileştirme sağlamak için gelecekte yapılabilecek olan bazı çalışmalar şöyle sıralanabilir:

- Bilgi damıtımını(distillation) kullanarak küçük bir modelin büyük bir modeli taklit etmesini sağlamak
- Mevcut verileri zenginleştirme(augmentation) ve sentetik veri üretimi ile model performansını artırmak
- Kontrastlı(Contrastive) öğrenme [265] ile görüntü metin hizalaması iyi ve kötü olan eşleşmeler dikkate alınarak daha iyi öğrenme sağlamak
- Aktif öğrenme [266] ile modelin en çok hataya düştüğü örnekleri seçerek eğitim sürecine yön vermek
- Planlı(Curriculum) öğrenme [267] ile modeli önce basit ve net metin-görüntü çiftleri ile eğitip zamanla daha karmaşık metin istemlerini üretmek

KAYNAKÇA

- [1] D. P. Kingma, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [2] I. Goodfellow ve dig., “Generative adversarial networks,” *Communications of the ACM*, c. 63, no. 11, ss. 139–144, 2020.
- [3] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint arXiv:1308.0850*, 2013.
- [4] D. E. Rumelhart, G. E. Hinton R. J. Williams, “Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986,” *Biometrika*, c. 71, no. 599-607, s. 6, 1986.
- [5] S. Hochreiter, “Long Short-term Memory,” *Neural Computation MIT-Press*, 1997.
- [6] J. Chung, C. Gulcehre, K. Cho Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [7] A. Radford, “Improving language understanding by generative pre-training,” 2018.
- [8] A. Van Den Oord, N. Kalchbrenner K. Kavukcuoglu, “Pixel recurrent neural networks,” *International conference on machine learning*, PMLR, 2016, ss. 1747–1756.
- [9] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves ve dig., “Conditional image generation with pixelcnn decoders,” *Advances in neural information processing systems*, c. 29, 2016.
- [10] C.-C. Lin, A. Jaech, X. Li, M. R. Gormley J. Eisner, *Limitations of Autoregressive Models and Their Alternatives*, 2021. arXiv: 2010 . 11939 [cs.LG].
- [11] L. Dinh, D. Krueger Y. Bengio, “Nice: Non-linear independent components estimation,” *arXiv preprint arXiv:1410.8516*, 2014.
- [12] E. G. Tabak E. Vanden-Eijnden, “Density estimation by dual ascent of the log-likelihood,” *Communications in Mathematical Sciences*, c. 8, no. 1, ss. 217–233, 2010.
- [13] E. G. Tabak C. V. Turner, “A family of nonparametric density estimation algorithms,” *Communications on Pure and Applied Mathematics*, c. 66, no. 2, ss. 145–164, 2013.
- [14] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato F. Huang, “A tutorial on energy-based learning,” *Predicting structured data*, c. 1, no. 0, 2006.

- [15] Y. Song D. P. Kingma, “How to train your energy-based models,” *arXiv preprint arXiv:2101.03288*, 2021.
- [16] J. Xie, Y. Lu, S.-C. Zhu Y. Wu, “A theory of generative convnet,” *International conference on machine learning*, PMLR, 2016, ss. 2635–2644.
- [17] A. Vaswani ve diğ., “Attention is all you need,” *Advances in neural information processing systems*, c. 30, 2017.
- [18] P. Vincent, “A connection between score matching and denoising autoencoders,” *Neural computation*, c. 23, no. 7, ss. 1661–1674, 2011.
- [19] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” *International Conference on Machine Learning*, PMLR, 2015, ss. 2256–2265.
- [20] M. Suzuki, K. Nakayama Y. Matsuo, “Joint multimodal learning with deep generative models,” *arXiv preprint arXiv:1611.01891*, 2016.
- [21] M. Wu N. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” *Advances in neural information processing systems*, c. 31, 2018.
- [22] Y. Shi, B. Paige, P. Torr ve diğ., “Variational mixture-of-experts autoencoders for multi-modal deep generative models,” *Advances in neural information processing systems*, c. 32, 2019.
- [23] T. Sutter, I. Daunhawer J. Vogt, “Multimodal generative learning utilizing jensen-shannon-divergence,” *Advances in neural information processing systems*, c. 33, ss. 6100–6110, 2020.
- [24] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural computation*, c. 14, no. 8, ss. 1771–1800, 2002.
- [25] C. K. Williams F. V. Agakov, “Products of Gaussians and probabilistic minor component analysis,” *Neural Computation*, c. 14, no. 5, ss. 1169–1182, 2002.
- [26] Y. Cao D. J. Fleet, “Generalized product of experts for automatic and principled fusion of Gaussian process predictions,” *arXiv preprint arXiv:1410.7827*, 2014.
- [27] R. A. Jacobs, M. I. Jordan, S. J. Nowlan G. E. Hinton, “Adaptive mixtures of local experts,” *Neural computation*, c. 3, no. 1, ss. 79–87, 1991.
- [28] J. Ho, A. Jain P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, c. 33, ss. 6840–6851, 2020.
- [29] Y. Song S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in Neural Information Processing Systems*, c. 32, 2019.
- [30] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020.
- [31] F.-A. Croitoru, V. Hondru, R. T. Ionescu M. Shah, “Diffusion models in vision: A survey,” *arXiv preprint arXiv:2209.04747*, 2022.

- [32] A. Ulhaq, N. Akhtar G. Pogrebna, *Efficient Diffusion Models for Vision: A Survey*, 2022. arXiv: 2210.09292 [cs.CV].
- [33] X. Li ve diğ., “Diffusion Models for Image Restoration and Enhancement—A Comprehensive Survey,” *arXiv preprint arXiv:2308.09388*, 2023.
- [34] H. Zou, Z. M. Kim D. Kang, “Diffusion models in nlp: A survey,” *arXiv preprint arXiv:2305.14671*, 2023.
- [35] A. Kazerouni ve diğ., “Diffusion models for medical image analysis: A comprehensive survey,” *arXiv preprint arXiv:2211.07804*, 2022.
- [36] L. Lin, Z. Li, R. Li, X. Li J. Gao, “Diffusion models for time series applications: A survey,” *arXiv preprint arXiv:2305.00624*, 2023.
- [37] C. Zhang, C. Zhang, M. Zhang I. S. Kweon, “Text-to-image diffusion model in generative ai: A survey,” *arXiv preprint arXiv:2303.07909*, 2023.
- [38] C. Zhang ve diğ., “A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai,” *arXiv preprint arXiv:2303.13336*, c. 2, 2023.
- [39] H. Cao, C. Tan, Z. Gao, G. Chen, P.-A. Heng S. Z. Li, “A survey on generative diffusion model,” *arXiv preprint arXiv:2209.02646*, 2022.
- [40] L. Yang ve diğ., “Diffusion models: A comprehensive survey of methods and applications,” *arXiv preprint arXiv:2209.00796*, 2022.
- [41] Z. Chang, G. A. Koulieris H. P. H. Shum, *On the Design Fundamentals of Diffusion Models: A Survey*, 2023. arXiv: 2306.04542 [cs.LG].
- [42] Y. Qiu, L. Zhang X. Wang, “Unbiased contrastive divergence algorithm for training energy-based latent variable models,” *International Conference on Learning Representations*, 2019.
- [43] Y. Song, S. Garg, J. Shi S. Ermon, “Sliced score matching: A scalable approach to density and score estimation,” *Uncertainty in Artificial Intelligence*, PMLR, 2020, ss. 574–584.
- [44] C.-W. Huang, J. H. Lim A. C. Courville, “A variational perspective on diffusion-based generative models and score matching,” *Advances in Neural Information Processing Systems*, c. 34, ss. 22 863–22 876, 2021.
- [45] E. Nachmani, R. S. Roman L. Wolf, “Non gaussian denoising diffusion models,” *arXiv preprint arXiv:2106.07582*, 2021.
- [46] E. Nachmani, R. S. Roman L. Wolf, *Denoising Diffusion Gamma Models*, 2021. arXiv: 2110.05948 [eess.SP].
- [47] Z. Xiao, K. Kreis A. Vahdat, “Tackling the generative learning trilemma with denoising diffusion gans,” *arXiv preprint arXiv:2112.07804*, 2021.
- [48] G. Batzolis, J. Stanczuk, C.-B. Schönlieb C. Etmann, *Non-Uniform Diffusion Models*, 2022. arXiv: 2207.09786 [cs.LG].
- [49] V. Voleti, C. Pal A. Oberman, *Score-based Denoising Diffusion with Non-Isotropic Gaussian Noise Models*, 2022. arXiv: 2210.12254 [cs.LG].

- [50] G. Daras, M. Delbracio, H. Talebi, A. G. Dimakis P. Milanfar, “Soft diffusion: Score matching for general corruptions,” *arXiv preprint arXiv:2209.05442*, 2022.
- [51] E. Hoogeboom T. Salimans, *Blurring Diffusion Models*, 2022. arXiv: 2209.05557 [cs.LG].
- [52] S. Rissanen, M. Heinonen A. Solin, *Generative Modelling With Inverse Heat Dissipation*, 2023. arXiv: 2206.13397 [cs.CV].
- [53] D. Chen, Z. Zhou, J.-P. Mei, C. Shen, C. Chen C. Wang, *A Geometric Perspective on Diffusion Models*, 2023. arXiv: 2305.19947 [cs.CV].
- [54] A. Q. Nichol P. Dhariwal, “Improved denoising diffusion probabilistic models,” *International Conference on Machine Learning*, PMLR, 2021, ss. 8162–8171.
- [55] Z. Kong W. Ping, “On fast sampling of diffusion probabilistic models,” *arXiv preprint arXiv:2106.00132*, 2021.
- [56] J. Song, C. Meng S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.
- [57] Q. Zhang Y. Chen, “Diffusion normalizing flow,” *Advances in Neural Information Processing Systems*, c. 34, ss. 16 280–16 291, 2021.
- [58] D. Kingma, T. Salimans, B. Poole J. Ho, “Variational diffusion models,” *Advances in neural information processing systems*, c. 34, ss. 21 696–21 707, 2021.
- [59] M. W. Lam, J. Wang, D. Su D. Yu, “BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis,” *arXiv preprint arXiv:2203.13508*, 2022.
- [60] Y. Benny L. Wolf, *Dynamic Dual-Output Diffusion Models*, 2022. arXiv: 2203.04304 [cs.CV].
- [61] R. San-Roman, E. Nachmani L. Wolf, *Noise Estimation for Generative Diffusion Models*, 2021. arXiv: 2104.02600 [cs.LG].
- [62] S. Lin, B. Liu, J. Li X. Yang, *Common Diffusion Noise Schedules and Sample Steps are Flawed*, 2023. arXiv: 2305.08891 [cs.CV].
- [63] Z. Duan, C. Wang, C. Chen, J. Huang W. Qian, “Optimal Linear Subspace Search: Learning to Construct Fast and High-Quality Schedulers for Diffusion Models,” *arXiv preprint arXiv:2305.14677*, 2023.
- [64] T. Chen, *On the Importance of Noise Scheduling for Diffusion Models*, 2023. arXiv: 2301.10972 [cs.CV].
- [65] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim S. Yoon, *Perception Prioritized Training of Diffusion Models*, 2022. arXiv: 2204.00227 [cs.CV].
- [66] A. Bansal ve diğ., “Cold diffusion: Inverting arbitrary image transforms without noise,” *arXiv preprint arXiv:2208.09392*, 2022.
- [67] H. Zheng, P. He, W. Chen M. Zhou, “Truncated diffusion probabilistic models,” *stat*, c. 1050, s. 7, 2022.
- [68] H. Zheng M. Zhou, “ACT: Asymptotic Conditional Transport,” 2020.

- [69] Z. Lyu, X. Xu, C. Yang, D. Lin B. Dai, “Accelerating Diffusion Models via Early Stop of the Diffusion Process,” *arXiv preprint arXiv:2205.12524*, 2022.
- [70] K. Pandey, A. Mukherjee, P. Rai A. Kumar, “Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents,” *arXiv preprint arXiv:2201.00308*, 2022.
- [71] D. Kim, S. Shin, K. Song, W. Kang I.-C. Moon, “Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation,” *International Conference on Machine Learning*, PMLR, 2022, ss. 11 201–11 228.
- [72] H. Chung, B. Sim J. C. Ye, “Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, ss. 12 413–12 422.
- [73] G. Franzese ve diğ., “How much is enough? a study on diffusion times in score-based generative models,” *arXiv preprint arXiv:2206.05173*, 2022.
- [74] A. Jolicoeur-Martineau, R. Piché-Taillefer, R. T. d. Combes I. Mitliagkas, “Adversarial score matching and improved sampling for image generation,” *arXiv preprint arXiv:2009.05475*, 2020.
- [75] K. Deja, A. Kuzina, T. Trzciński J. M. Tomczak, *On Analyzing Generative and Denoising Capabilities of Diffusion-based Deep Generative Models*, 2022. arXiv: 2206.00070 [cs.LG].
- [76] Y. Lee, J.-Y. Kim, H. Go, M. Jeong, S. Oh S. Choi, *Multi-Architecture Multi-Expert Diffusion Models*, 2023. arXiv: 2306.04990 [cs.CV].
- [77] W. Cho ve diğ., *Towards Enhanced Controllability of Diffusion Models*, 2023. arXiv: 2302.14368 [cs.CV].
- [78] M. Yi, J. Sun Z. Li, *On the Generalization of Diffusion Model*, 2023. arXiv: 2305.14712 [cs.LG].
- [79] Z. Wu, P. Zhou, K. Kawaguchi H. Zhang, *Fast Diffusion Model*, 2023. arXiv: 2306.06991 [cs.CV].
- [80] Y. Xu, S. Tong T. Jaakkola, *Stable Target Field for Reduced Variance Score Estimation in Diffusion Models*, 2023. arXiv: 2302.00670 [cs.LG].
- [81] M. Ning, E. Sangineto, A. Porrello, S. Calderara R. Cucchiara, *Input Perturbation Reduces Exposure Bias in Diffusion Models*, 2023. arXiv: 2301.11706 [cs.LG].
- [82] F. V. S. Jothiraj A. Mashhadi, *Phoenix: A Federated Generative Diffusion Model*, 2023. arXiv: 2306.04098 [cs.LG].
- [83] E. Luhman T. Luhman, “Knowledge distillation in iterative generative models for improved sampling speed,” *arXiv preprint arXiv:2101.02388*, 2021.
- [84] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui Y. Ren, “Prodiff: Progressive fast diffusion model for high-quality text-to-speech,” *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, ss. 2595–2605.

- [85] G. Hinton, O. Vinyals J. Dean, *Distilling the Knowledge in a Neural Network*, 2015. arXiv: 1503.02531 [stat.ML].
- [86] D. Ryu J. C. Ye, *Pyramidal Denoising Diffusion Probabilistic Models*, 2022. arXiv: 2208.01864 [cs.CV].
- [87] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi T. Salimans, “Cascaded Diffusion Models for High Fidelity Image Generation.,” *J. Mach. Learn. Res.*, c. 23, ss. 47–1, 2022.
- [88] X. Han, H. Zheng M. Zhou, “CARD: Classification and regression diffusion models,” *arXiv preprint arXiv:2206.07275*, 2022.
- [89] J. Ho T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [90] V. T. Hu, D. W. Zhang, Y. M. Asano, G. J. Burghouts C. G. M. Snoek, *Self-Guided Diffusion Models*, 2023. arXiv: 2210.06462 [cs.CV].
- [91] A. Blattmann, R. Rombach, K. Oktay, J. Müller B. Ommer, “Semi-Parametric Neural Image Synthesis,” *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave K. Cho, ed., 2022. erişim adresi: <https://openreview.net/forum?id=Bqk9c0wBNrZ>.
- [92] A. Blattmann, R. Rombach, K. Oktay, J. Müller B. Ommer, “Retrieval-augmented diffusion models,” *Advances in Neural Information Processing Systems*, c. 35, ss. 15 309–15 324, 2022.
- [93] S. Sheynin ve diğ., *KNN-Diffusion: Image Generation via Large-Scale Retrieval*, 2022. arXiv: 2204.02849 [cs.CV].
- [94] R. Rombach, A. Blattmann B. Ommer, *Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models*, 2022. arXiv: 2207.13038 [cs.CV].
- [95] W. Chen, H. Hu, C. Saharia W. W. Cohen, *Re-Imagen: Retrieval-Augmented Text-to-Image Generator*, 2022. arXiv: 2209.14491 [cs.CV].
- [96] D. Kim, B. Na, S. J. Kwon, D. Lee, W. Kang I.-c. Moon, “Maximum Likelihood Training of Parametrized Diffusion Model,” 2021.
- [97] D. Kim, B. Na, S. J. Kwon, D. Lee, W. Kang I.-C. Moon, “Maximum Likelihood Training of Implicit Nonlinear Diffusion Models,” *arXiv preprint arXiv:2205.13699*, 2022.
- [98] Y. Song, C. Durkan, I. Murray S. Ermon, “Maximum likelihood training of score-based diffusion models,” *Advances in Neural Information Processing Systems*, c. 34, ss. 1415–1428, 2021.
- [99] C. Lu, K. Zheng, F. Bao, J. Chen, C. Li J. Zhu, “Maximum likelihood training for score-based diffusion odes by high order denoising score matching,” *International Conference on Machine Learning*, PMLR, 2022, ss. 14 429–14 460.
- [100] T. Dockhorn, A. Vahdat K. Kreis, “Score-based generative modeling with critically-damped langevin diffusion,” *arXiv preprint arXiv:2112.07068*, 2021.

- [101] A. Vahdat, K. Kreis J. Kautz, “Score-based generative modeling in latent space,” *Advances in Neural Information Processing Systems*, c. 34, ss. 11 287–11 302, 2021.
- [102] R. Rombach, A. Blattmann, D. Lorenz, P. Esser B. Ommer, “High-resolution image synthesis with latent diffusion models,” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, ss. 10 684–10 695.
- [103] B. Jing, G. Corso, R. Berlinghieri T. Jaakkola, “Subspace diffusion generative models,” *arXiv preprint arXiv:2205.01490*, 2022.
- [104] K. Pandey S. Mandt, *Generative Diffusions in Augmented Spaces: A Complete Recipe*, 2023. arXiv: 2303.01748 [cs.LG].
- [105] C. H. Wu F. D. la Torre, *Unifying Diffusion Models’ Latent Space, with Applications to CycleDiffusion and Guidance*, 2022. arXiv: 2210.05559 [cs.CV].
- [106] M. Bounoua, G. Franzese P. Michiardi, *Multi-modal Latent Diffusion*, 2023. arXiv: 2306.04445 [cs.LG].
- [107] H. Zhang ve diğ., “Dimensionality-Varying Diffusion Process,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, ss. 14 307–14 316.
- [108] A. Campbell, W. Harvey, C. Weilbach, V. D. Bortoli, T. Rainforth A. Doucet, *Trans-Dimensional Generative Modeling via Jump Diffusion Models*, 2023. arXiv: 2305.16261 [stat.ML].
- [109] Y. Xu, Z. Liu, M. Tegmark T. Jaakkola, “Poisson flow generative models,” *Advances in Neural Information Processing Systems*, c. 35, ss. 16 782–16 795, 2022.
- [110] Y. Xu, Z. Liu, Y. Tian, S. Tong, M. Tegmark T. Jaakkola, “Pfgm++: Unlocking the potential of physics-inspired generative models,” *arXiv preprint arXiv:2302.04265*, 2023.
- [111] H. Phung, Q. Dao A. Tran, “Wavelet Diffusion Models are fast and scalable Image Generators,” *arXiv preprint arXiv:2211.16152*, 2022.
- [112] V. De Bortoli, A. Doucet, J. Heng J. Thornton, “Simulating diffusion bridges with score matching,” *arXiv preprint arXiv:2111.07243*, 2021.
- [113] V. Khrulkov, G. Ryzhakov, A. Chertkov I. Oseledets, “Understanding DDPM Latent Codes Through Optimal Transport,” *The Eleventh International Conference on Learning Representations*, 2023. erişim adresi: <https://openreview.net/forum?id=6PIrhAx1j4i>.
- [114] X. Su, J. Song, C. Meng S. Ermon, “Dual diffusion implicit bridges for image-to-image translation,” *arXiv preprint arXiv:2203.08382*, 2022.
- [115] S. Lee, B. Kim J. C. Ye, “Minimizing trajectory curvature of ode-based generative models,” *arXiv preprint arXiv:2301.12003*, 2023.
- [116] E. Heitz, L. Belcour T. Champon, “Iterative α -Blending: a Minimalist Deterministic Diffusion Model,” *arXiv preprint arXiv:2305.03486*, 2023.

- [117] X. Liu, C. Gong Q. Liu, “Flow straight and fast: Learning to generate and transfer data with rectified flow,” *arXiv preprint arXiv:2209.03003*, 2022.
- [118] M. S. Albergo E. Vandenberg-Eijnden, “Building normalizing flows with stochastic interpolants,” *arXiv preprint arXiv:2209.15571*, 2022.
- [119] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel M. Le, “Flow matching for generative modeling,” *arXiv preprint arXiv:2210.02747*, 2022.
- [120] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré M. Welling, “Argmax flows and multinomial diffusion: Learning categorical distributions,” *Advances in Neural Information Processing Systems*, c. 34, ss. 12 454–12 465, 2021.
- [121] J. Austin, D. D. Johnson, J. Ho, D. Tarlow R. van den Berg, “Structured denoising diffusion models in discrete state-spaces,” *Advances in Neural Information Processing Systems*, c. 34, ss. 17 981–17 993, 2021.
- [122] E. Hoogeboom, A. A. Gritsenko, J. Bastings, B. Poole, R. v. d. Berg T. Salimans, “Autoregressive diffusion models,” *arXiv preprint arXiv:2110.02037*, 2021.
- [123] S. Gu ve diğ., “Vector quantized diffusion model for text-to-image synthesis,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, ss. 10 696–10 706.
- [124] A. Van Den Oord, O. Vinyals ve diğ., “Neural discrete representation learning,” *Advances in neural information processing systems*, c. 30, 2017.
- [125] T. Chen, R. Zhang G. Hinton, *Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning*, 2023. arXiv: 2208 . 04202 [cs . CV].
- [126] J. E. Santos, Z. R. Fox, N. Lubbers Y. T. Lin, *Blackout Diffusion: Generative Diffusion Models in Discrete-State Spaces*, 2023. arXiv: 2305 . 11089 [cs . LG].
- [127] A. Campbell, J. Benton, V. De Bortoli, T. Rainforth, G. Deligiannidis A. Doucet, “A Continuous Time Framework for Discrete Denoising Models,” *arXiv preprint arXiv:2205.14987*, 2022.
- [128] C. Niu, Y. Song, J. Song, S. Zhao, A. Grover S. Ermon, “Permutation invariant graph generation via score-based generative modeling,” *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, ss. 4474–4484.
- [129] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner G. Monfardini, “The graph neural network model,” *IEEE transactions on neural networks*, c. 20, no. 1, ss. 61–80, 2008.
- [130] J. Jo, S. Lee S. J. Hwang, “Score-based Generative Modeling of Graphs via the System of Stochastic Differential Equations,” *arXiv preprint arXiv:2202.02514*, 2022.
- [131] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon J. Tang, “Geodiff: A geometric diffusion model for molecular conformation generation,” *arXiv preprint arXiv:2203.02923*, 2022.

- [132] C. Shi, S. Luo, M. Xu J. Tang, “Learning gradient fields for molecular conformation generation,” *International Conference on Machine Learning*, PMLR, 2021, ss. 9558–9568.
- [133] H. Huang, L. Sun, B. Du, Y. Fu W. Lv, “Graphgdp: Generative diffusion processes for permutation invariant graph generation,” *2022 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2022, ss. 201–210.
- [134] X. Chen, Y. Li, A. Zhang L.-p. Liu, “Nvdiff: Graph generation through the diffusion of node vectors,” *arXiv preprint arXiv:2211.10794*, 2022.
- [135] T. Luo, Z. Mo S. J. Pan, “Fast graph generative model via spectral diffusion,” *arXiv preprint arXiv:2211.08892*, 2022.
- [136] C. Fefferman, S. Mitter H. Narayanan, “Testing the manifold hypothesis,” *Journal of the American Mathematical Society*, c. 29, no. 4, ss. 983–1049, 2016.
- [137] V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh A. Doucet, “Riemannian score-based generative modeling,” *arXiv preprint arXiv:2202.02763*, 2022.
- [138] C.-W. Huang, M. Aghajohari, A. J. Bose, P. Panangaden A. Courville, “Riemannian Diffusion Models,” *arXiv preprint arXiv:2208.07949*, 2022.
- [139] L. Luzi, A. Siahkoohi, P. M. Mayer, J. Casco-Rodriguez R. Baraniuk, “Boomerang: Local sampling on image manifolds using diffusion models,” *arXiv preprint arXiv:2210.12100*, 2022.
- [140] X. Cheng, J. Zhang S. Sra, “Theory and Algorithms for Diffusion Processes on Riemannian Manifolds,” *arXiv preprint arXiv:2204.13665*, 2022.
- [141] P. Zhuang, S. Abnar, J. Gu, A. Schwing, J. M. Susskind M. Á. Bautista, *Diffusion Probabilistic Fields*, 2023. arXiv: 2303 . 00165 [cs . CV].
- [142] J. Thornton, M. Hutchinson, E. Mathieu, V. D. Bortoli, Y. W. Teh A. Doucet, *Riemannian Diffusion Schrödinger Bridge*, 2022. arXiv: 2207 . 03024 [stat . ML].
- [143] Y.-H. Park, M. Kwon, J. Jo Y. Uh, *Unsupervised Discovery of Semantic Latent Directions in Diffusion Models*, 2023. arXiv: 2302 . 12469 [cs . CV].
- [144] A. Hyvärinen P. Dayan, “Estimation of non-normalized statistical models by score matching.,” *Journal of Machine Learning Research*, c. 6, no. 4, 2005.
- [145] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman I. Mitliagkas, “Gotta go fast when generating data with score-based models,” *arXiv preprint arXiv:2105.14080*, 2021.
- [146] C. Meng, Y. Song, W. Li S. Ermon, *Estimating High Order Gradients of the Data Distribution by Denoising*, 2021. arXiv: 2111 . 04726 [cs . LG].
- [147] P. Verma, V. Adam A. Solin, *Variational Gaussian Process Diffusion Processes*, 2023. arXiv: 2306 . 02066 [cs . LG].
- [148] S. Li, W. Chen D. Zeng, *SciRE-Solver: Accelerating Diffusion Models Sampling by Score-integrand Solver with Recursive Difference*, 2023. arXiv: 2308 . 07896 [stat . ML].

- [149] M. F. Hutchinson, “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines,” *Communications in Statistics-Simulation and Computation*, c. 18, no. 3, ss. 1059–1076, 1989.
- [150] J. Skilling, “The eigenvalues of mega-dimensional matrices,” içinde *Maximum Entropy and Bayesian Methods*, Springer, 1989, ss. 455–466.
- [151] Q. Zhang, M. Tao Y. Chen, “gDDIM: Generalized denoising diffusion implicit models,” *arXiv preprint arXiv:2206.05564*, 2022.
- [152] Q. Zhang Y. Chen, “Fast Sampling of Diffusion Models with Exponential Integrator,” *arXiv preprint arXiv:2204.13902*, 2022.
- [153] T. Karras, M. Aittala, T. Aila S. Laine, “Elucidating the Design Space of Diffusion-Based Generative Models,” *arXiv preprint arXiv:2206.00364*, 2022.
- [154] U. M. Ascher L. R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations*. Siam, 1998, c. 61.
- [155] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li J. Zhu, “DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps,” *arXiv preprint arXiv:2206.00927*, 2022.
- [156] T. Pang, C. Lu, C. Du, M. Lin, S. Yan Z. Deng, *On Calibrating Diffusion Probabilistic Models*, 2023. arXiv: 2302 .10688 [cs .LG].
- [157] L. Liu, Y. Ren, Z. Lin Z. Zhao, “Pseudo numerical methods for diffusion models on manifolds,” *arXiv preprint arXiv:2202.09778*, 2022.
- [158] Y. Xu, M. Deng, X. Cheng, Y. Tian, Z. Liu T. Jaakkola, “Restart Sampling for Improving Generative Processes,” *arXiv preprint arXiv:2306.14878*, 2023.
- [159] Y. Cao, J. Chen, Y. Luo X. Zhou, *Exploring the Optimal Choice for Generative Processes in Diffusion Models: Ordinary vs Stochastic Differential Equations*, 2023. arXiv: 2306 .02063 [cs .LG].
- [160] D. Watson, J. Ho, M. Norouzi W. Chan, “Learning to efficiently sample from diffusion probabilistic models,” *arXiv preprint arXiv:2106.03802*, 2021.
- [161] D. Watson, W. Chan, J. Ho M. Norouzi, “Learning fast samplers for diffusion models by differentiating through sample quality,” *International Conference on Learning Representations*, 2021.
- [162] M. Bińkowski, D. J. Sutherland, M. Arbel A. Gretton, “Demystifying mmd gans,” *arXiv preprint arXiv:1801.01401*, 2018.
- [163] F. Bao, C. Li, J. Zhu B. Zhang, “Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models,” *arXiv preprint arXiv:2201.06503*, 2022.
- [164] F. Bao, C. Li, J. Sun, J. Zhu B. Zhang, “Estimating the optimal covariance with imperfect mean in diffusion probabilistic models,” *arXiv preprint arXiv:2206.07309*, 2022.
- [165] M. Ning, M. Li, J. Su, A. A. Salah I. O. Ertugrul, “Elucidating the Exposure Bias in Diffusion Models,” *arXiv preprint arXiv:2308.15321*, 2023.

- [166] H. Zheng, W. Nie, A. Vahdat, K. Azizzadenesheli A. Anandkumar, “Fast sampling of diffusion models via operator learning,” *International Conference on Machine Learning*, PMLR, 2023, ss. 42 390–42 402.
- [167] M. Li, T. Qu, W. Sun M.-F. Moens, *Alleviating Exposure Bias in Diffusion Models through Sampling with Shifted Time Steps*, 2023. arXiv: 2305 . 15583 [cs.CV].
- [168] G. Fang, X. Ma X. Wang, *Structural Pruning for Diffusion Models*, 2023. arXiv: 2305 . 10924 [cs.LG].
- [169] D. Kim, Y. Kim, S. J. Kwon, W. Kang I.-C. Moon, *Refining Generative Process with Discriminator Guidance in Score-based Diffusion Models*, 2023. arXiv: 2211 . 17091 [cs.CV].
- [170] B. Kawar, R. Ganz M. Elad, *Enhancing Diffusion-Based Image Synthesis with Robust Classifier Guidance*, 2023. arXiv: 2208 . 08664 [cs.CV].
- [171] G. Giannone, D. Nielsen O. Winther, “Few-Shot Diffusion Models,” *arXiv preprint arXiv:2205.15463*, 2022.
- [172] V. Sehwag, C. Hazirbas, A. Gordo, F. Ozgenel C. C. Ferrer, *Generating High Fidelity Data from Low-density Regions using Diffusion Models*, 2022. arXiv: 2203 . 17260 [cs.CV].
- [173] S. Hong, G. Lee, W. Jang S. Kim, *Improving Sample Quality of Diffusion Models Using Self-Attention Guidance*, 2023. arXiv: 2210 . 00939 [cs.CV].
- [174] B. Kawar, M. Elad, S. Ermon J. Song, “Denoising diffusion restoration models,” *arXiv preprint arXiv:2201.11793*, 2022.
- [175] A. Shih, S. Belkhale, S. Ermon, D. Sadigh N. Anari, *Parallel Sampling of Diffusion Models*, 2023. arXiv: 2305 . 16317 [cs.LG].
- [176] C. Xiang, F. Bao, C. Li, H. Su J. Zhu, *A Closer Look at Parameter-Efficient Tuning in Diffusion Models*, 2023. arXiv: 2303 . 18181 [cs.CV].
- [177] E. Aiello, D. Valsesia E. Magli, “Fast inference in denoising diffusion models via mmd finetuning,” *arXiv preprint arXiv:2301.07969*, 2023.
- [178] J. Choi, S. Kim, Y. Jeong, Y. Gwon S. Yoon, *ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models*, 2021. arXiv: 2108 . 02938 [cs.CV].
- [179] P. Dhariwal A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in Neural Information Processing Systems*, c. 34, ss. 8780–8794, 2021.
- [180] A. Graikos, S. Yellapragada D. Samaras, *Conditional Generation from Unconditional Diffusion Models using Denoiser Representations*, 2023. arXiv: 2306 . 01900 [cs.CV].
- [181] B. Wallace, A. Gokul, S. Ermon N. Naik, *End-to-End Diffusion Latent Optimization Improves Classifier Guidance*, 2023. arXiv: 2303 . 13703 [cs.CV].
- [182] M. W. Shen ve diğ., *Conditional Diffusion with Less Explicit Guidance via Model Predictive Control*, 2022. arXiv: 2210 . 12192 [cs.LG].

- [183] J. Gou, B. Yu, S. J. Maybank D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, c. 129, no. 6, ss. 1789–1819, 2021.
- [184] T. Salimans J. Ho, “Progressive distillation for fast sampling of diffusion models,” *arXiv preprint arXiv:2202.00512*, 2022.
- [185] C. Meng ve diğ., “On distillation of guided diffusion models,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, ss. 14 297–14 306.
- [186] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, c. 1, no. 2, s. 3, 2022.
- [187] C. Saharia ve diğ., “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in neural information processing systems*, c. 35, ss. 36 479–36 494, 2022.
- [188] D. Berthelot ve diğ., *TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation*, 2023. arXiv: 2303 . 04248 [cs . LG].
- [189] W. Sun, D. Chen, C. Wang, D. Ye, Y. Feng C. Chen, “Accelerating diffusion sampling with classifier-based feature distillation,” *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, ss. 810–815.
- [190] B. Poole, A. Jain, J. T. Barron B. Mildenhall, “Dreamfusion: Text-to-3d using 2d diffusion,” *arXiv preprint arXiv:2209.14988*, 2022.
- [191] J. Gu, S. Zhai, Y. Zhang, L. Liu J. Susskind, *BOOT: Data-free Distillation of Denoising Diffusion Models with Bootstrapping*, 2023. arXiv: 2306 . 05544 [cs . CV].
- [192] Y. Song, P. Dhariwal, M. Chen I. Sutskever, “Consistency models,” 2023.
- [193] D. Kim ve diğ., “Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion,” *arXiv preprint arXiv:2310.02279*, 2023.
- [194] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, c. 29, 2016.
- [195] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, ss. 248–255.
- [196] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens Z. Wojna, “Rethinking the inception architecture for computer vision,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, ss. 2818–2826.
- [197] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, c. 30, 2017.
- [198] A. Razavi, A. Van den Oord O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2,” *Advances in neural information processing systems*, c. 32, 2019.

- [199] A. Krizhevsky, G. Hinton ve diğ., “Learning multiple layers of features from tiny images,” 2009.
- [200] T. Chen, G.-H. Liu E. A. Theodorou, “Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory,” *arXiv preprint arXiv:2110.11291*, 2021.
- [201] T. Dockhorn, A. Vahdat K. Kreis, “GENIE: Higher-Order Denoising Diffusion Solvers,” *arXiv preprint arXiv:2210.05475*, 2022.
- [202] R. Gao, Y. Song, B. Poole, Y. N. Wu D. P. Kingma, “Learning energy-based models by diffusion recovery likelihood,” *arXiv preprint arXiv:2012.08125*, 2020.
- [203] Y. Song S. Ermon, “Improved techniques for training score-based generative models,” *Advances in neural information processing systems*, c. 33, ss. 12 438–12 448, 2020.
- [204] Z. Liu, P. Luo, X. Wang X. Tang, “Deep learning face attributes in the wild,” *Proceedings of the IEEE international conference on computer vision*, 2015, ss. 3730–3738.
- [205] D. P. Kingma R. Gao, *Understanding Diffusion Objectives as the ELBO with Simple Data Augmentation*, 2023. arXiv: 2303.00848 [cs.LG].
- [206] F. Vargas, W. Grathwohl A. Doucet, *Denoising Diffusion Samplers*, 2023. arXiv: 2302.13834 [cs.LG].
- [207] K. Oko, S. Akiyama T. Suzuki, *Diffusion Models are Minimax Optimal Distribution Estimators*, 2023. arXiv: 2303.01861 [stat.ML].
- [208] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim A. R. Zhang, *Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions*, 2023. arXiv: 2209.11215 [cs.LG].
- [209] D. McAllester, *On the Mathematics of Diffusion Models*, 2023. arXiv: 2301.11108 [cs.LG].
- [210] F. Bao ve diğ., *All are Worth Words: A ViT Backbone for Diffusion Models*, 2023. arXiv: 2209.12152 [cs.CV].
- [211] X. Yang, S.-M. Shih, Y. Fu, X. Zhao S. Ji, *Your ViT is Secretly a Hybrid Discriminative-Generative Diffusion Model*, 2022. arXiv: 2208.07791 [cs.CV].
- [212] G. Daras, K. Shah, Y. Dagan, A. Gollakota, A. G. Dimakis A. Klivans, *Ambient Diffusion: Learning Clean Distributions from Corrupted Data*, 2023. arXiv: 2305.19256 [cs.LG].
- [213] G. Stein ve diğ., *Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models*, 2023. arXiv: 2306.04675 [cs.LG].
- [214] S. Patil, P. Cuenca, N. Lambert P. von Platen, “Stable Diffusion with Diffusers,” *Hugging Face Blog*, 2022.
- [215] Y. Li ve diğ., “Gligen: Open-set grounded text-to-image generation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, ss. 22 511–22 521.

- [216] Z. Yang ve diğ., “Reco: Region-controlled text-to-image generation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, ss. 14 246–14 255.
- [217] J. Xie ve diğ., “Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, ss. 7452–7461.
- [218] L. Qu, S. Wu, H. Fei, L. Nie T.-S. Chua, “Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation,” *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, ss. 643–654.
- [219] W. Feng ve diğ., “Layoutgpt: Compositional visual planning and generation with large language models,” *Advances in Neural Information Processing Systems*, c. 36, 2024.
- [220] L. Lian, B. Li, A. Yala T. Darrell, “Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models,” *arXiv preprint arXiv:2305.13655*, 2023.
- [221] W. Feng ve diğ., “Training-free structured diffusion guidance for compositional text-to-image synthesis,” *arXiv preprint arXiv:2212.05032*, 2022.
- [222] Z. Tang, Z. Yang, C. Zhu, M. Zeng M. Bansal, “Any-to-Any Generation via Composable Diffusion,” *arXiv preprint arXiv:2305.11846*, 2023.
- [223] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf D. Cohen-Or, “Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models,” *ACM Transactions on Graphics (TOG)*, c. 42, no. 4, ss. 1–10, 2023.
- [224] Y. Li, M. Keuper, D. Zhang A. Khoreva, “Divide & bind your attention for improved generative semantic nursing,” *arXiv preprint arXiv:2307.10864*, 2023.
- [225] O. Avrahami ve diğ., “Spatext: Spatio-textual representation for controllable image generation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, ss. 18 370–18 380.
- [226] O. Bar-Tal, L. Yariv, Y. Lipman T. Dekel, “Multidiffusion: Fusing diffusion paths for controlled image generation,” 2023.
- [227] T. Brooks, A. Holynski A. A. Efros, “Instructpix2pix: Learning to follow image editing instructions,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, ss. 18 392–18 402.
- [228] O. Patashnik, D. Garibi, I. Azuri, H. Averbuch-Elor D. Cohen-Or, “Localizing object-level shape variations with text-to-image diffusion models,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, ss. 23 051–23 061.
- [229] M. Chen, I. Laina A. Vedaldi, “Training-free layout control with cross-attention guidance,” *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, ss. 5343–5353.
- [230] T. Brown ve diğ., “Language models are few-shot learners,” *Advances in neural information processing systems*, c. 33, ss. 1877–1901, 2020.

- [231] Z. Zhao, E. Wallace, S. Feng, D. Klein S. Singh, “Calibrate before use: Improving few-shot performance of language models,” *International conference on machine learning*, PMLR, 2021, ss. 12 697–12 706.
- [232] J. Wei ve diğ., “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, c. 35, ss. 24 824–24 837, 2022.
- [233] L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang W. Y. Wang, “Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies,” *arXiv preprint arXiv:2308.03188*, 2023.
- [234] A. Radford ve diğ., “Learning transferable visual models from natural language supervision,” *International Conference on Machine Learning*, PMLR, 2021, ss. 8748–8763.
- [235] J.-B. Alayrac ve diğ., “Flamingo: a visual language model for few-shot learning,” *Advances in neural information processing systems*, c. 35, ss. 23 716–23 736, 2022.
- [236] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang N. Duan, “Visual chatgpt: Talking, drawing and editing with visual foundation models,” *arXiv preprint arXiv:2303.04671*, 2023.
- [237] Z. Yang ve diğ., “Mm-react: Prompting chatgpt for multimodal reasoning and action,” *arXiv preprint arXiv:2303.11381*, 2023.
- [238] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis A. Smola, “Multimodal chain-of-thought reasoning in language models,” *arXiv preprint arXiv:2302.00923*, 2023.
- [239] K. Lee ve diğ., “Aligning text-to-image models using human feedback,” *arXiv preprint arXiv:2302.12192*, 2023.
- [240] T.-H. Wu, L. Lian, J. E. Gonzalez, B. Li T. Darrell, “Self-correcting llm-controlled diffusion models,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, ss. 6327–6336.
- [241] Y. Lu, X. Yang, X. Li, X. E. Wang W. Y. Wang, “Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation,” *Advances in Neural Information Processing Systems*, c. 36, 2024.
- [242] C. Raffel ve diğ., “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, c. 21, no. 140, ss. 1–67, 2020.
- [243] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, ss. 658–666.
- [244] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye D. Ren, “Distance-IoU loss: Faster and better learning for bounding box regression,” *Proceedings of the AAAI conference on artificial intelligence*, c. 34, 2020, ss. 12 993–13 000.

- [245] D. Mahan, R. Carlow, L. Castricato, N. Cooper C. Laforte, *Stable Beluga models*. erişim adresi: [https : / / huggingface . co / stabilityai/StableBeluga2] (<https://huggingface.co/stabilityai/StableBeluga2>).
- [246] H. Touvron ve dig., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, 2023. arXiv: 2307.09288 [cs.CL].
- [247] J. Li, D. Li, C. Xiong S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” *International conference on machine learning*, PMLR, 2022, ss. 12 888–12 900.
- [248] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning,” *arXiv preprint arXiv:2104.08718*, 2021.
- [249] J. Li, D. Li, S. Savarese S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *International conference on machine learning*, PMLR, 2023, ss. 19 730–19 742.
- [250] K. Huang, K. Sun, E. Xie, Z. Li X. Liu, “T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation,” *Advances in Neural Information Processing Systems*, c. 36, ss. 78 723–78 747, 2023.
- [251] A. Hurst ve dig., “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [252] D. Podell ve dig., “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.
- [253] P. Esser ve dig., “Scaling rectified flow transformers for high-resolution image synthesis,” *Forty-first international conference on machine learning*, 2024.
- [254] BlackForest, *FLUX*, [https : / / github . com / black - forest - labs / flux](https://github.com/black-forest-labs/flux), 2024.
- [255] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon B. Cui, “Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms,” *Forty-first International Conference on Machine Learning*, 2024.
- [256] R. OpenAI, “Gpt-4 technical report. arxiv 2303.08774,” *View in Article*, c. 2, no. 5, 2023.
- [257] L. Lian, B. Shi, A. Yala, T. Darrell B. Li, “Llm-grounded video diffusion models,” *arXiv preprint arXiv:2309.17444*, 2023.
- [258] J. T. Hoe, X. Jiang, C. S. Chan, Y.-P. Tan W. Hu, “InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [259] G. Team, *Gemma 3 Technical Report*, 2025. arXiv: 2503 . 19786 [cs.CL]. erişim adresi: <https://arxiv.org/abs/2503.19786>.

- [260] G. Team, *Gemini: A Family of Highly Capable Multimodal Models*, 2025. arXiv: 2312.11805 [cs.CL]. erişim adresi: <https://arxiv.org/abs/2312.11805>.
- [261] X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar I. Misra, “Instancediffusion: Instance-level control for image generation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, ss. 6232–6242.
- [262] J. Cheng, Z. Zhao, T. He, T. Xiao, Z. Zhang Y. Zhou, “Rethinking the training and evaluation of rich-context layout-to-image generation,” *Advances in Neural Information Processing Systems*, c. 37, ss. 62 083–62 107, 2024.
- [263] L. Zhang, A. Rao M. Agrawala, “Adding conditional control to text-to-image diffusion models,” *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, ss. 3836–3847.
- [264] P. Lewis ve diğ., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, c. 33, ss. 9459–9474, 2020.
- [265] P. Khosla ve diğ., “Supervised contrastive learning,” *Advances in neural information processing systems*, c. 33, ss. 18 661–18 673, 2020.
- [266] B. Settles, “Active learning literature survey,” 2009.
- [267] Y. Bengio, J. Louradour, R. Collobert J. Weston, “Curriculum learning,” *Proceedings of the 26th annual international conference on machine learning*, 2009, ss. 41–48.

A BÖLÜM 6'DA ÖNERİLEN KOMUTLAR

A.1 Yerleştirme için MLLM Komutu

"""You are an expert bounding box generator . I will provide you a user prompt for a photo , image , or painting . Your task is to generate the bounding boxes for the objects mentioned in the user prompt. Please pay attention to the actions/relations while generating bounding boxes. The images are 1x1. The top - left corner has coordinate [0 , 0]. The bottom - right corner has coordinate [1 , 1]. The bounding boxes should not overlap or go beyond the image boundaries. Each bounding box should be in the format of object name [top - left x coordinate , top - left y coordinate , bottom - right x coordinate , bottom - right y coordinate] and should not include more than one object. Pay attention that the minimum size for each bounding box is 0.1 x 0.1. If needed, you can make reasonable guesses. Please do not include any other explanations. Please refer to the example below for the desired format.

User prompt: A realistic image of landscape scene depicting a green car parking on the left of a blue truck , with a red air balloon and a bird in the sky Objects: ['a green car[0.041, 0.549, 0.453, 0.859]', 'a blue truck[0.525, 0.553, 0.934, 0.865]', 'a red air balloon[0.129, 0.016, 0.412, 0.279]', 'a bird[0.578, 0.082, 0.857, 0.277]']

User prompt: A realistic top - down view of a wooden table with two apples on it Objects: ['a wooden table[0.039, 0.289, 0.961, 0.711]', 'an apple[0.293, 0.441, 0.488, 0.637]', 'an apple[0.547, 0.441, 0.742, 0.637]']

User prompt: A realistic scene of three skiers standing in a line on the snow near a palm tree Objects: ['a skier[0.01, 0.297, 0.281, 0.625]', 'a skier[0.543, 0.375, 0.779, 0.684]', 'a skier[0.289, 0.338, 0.531, 0.641]', 'a palm tree[0.789, 0.205, 0.99, 0.695]']

User prompt: An oil painting of a pink dolphin jumping on the left of a steam boat on the sea Objects: ['a steam boat[0.453, 0.439, 0.955, 0.73]', 'a jumping pink dolphin[0.041, 0.486, 0.41, 0.727]']

User prompt: A cute cat and an angry dog without birds Objects: ['a cute cat[0.1, 0.131, 0.629, 0.764]', 'an angry dog[0.59, 0.232, 1.002, 0.678]']

User prompt: Two pandas in a forest without flowers Objects: ['a panda[0.059, 0.334, 0.473, 0.775]', 'a panda[0.516, 0.338, 0.949, 0.77]']

User prompt: An oil painting of a living room scene without chairs with a painting mounted on the wall , a cabinet below the painting , and two flower vases on the cabinet Objects: ['a painting[0.172, 0.166, 0.826, 0.562]', 'a cabinet[0.111, 0.602, 0.9, 0.994]', 'a flower vase[0.324, 0.434, 0.504, 0.645]', 'a flower vase[0.641, 0.434, 0.82, 0.645]']

User prompt: """

A.2 Kendini Düzeltme için MLLM Komutu

""You are an expert bounding box adjuster. I will provide you a user prompt with current objects' bounding boxes. Your objective is manipulating bounding boxes according to the user prompt while maintaining visual accuracy. Given user prompts are containing interactions between objects. Please pay attention to these interactions while adjusting bounding boxes. First read and understand the user prompt. After that review the current objects' bounding boxes and adjust to meet user specifications.

Guidelines for adjustment: Alignment: Follow the user's prompt, keeping the specified object count and attributes. If the described object lacks specified attributes adjust it with its attributes. Boundary Adherence: Keep bounding box coordinates within [0, 1]. Minimal Modifications: Change bounding boxes only if they don't match the user's prompt (i.e., don't modify matched objects). Overlap Reduction: Minimize intersections in new boxes and remove the smallest, least overlapping objects. Explain Adjustments: Justify the reasons behind each alteration and ensure every adjustment abides by the guidelines. Output the Result: Present the reasoning first (max. 50 words), then give bounding boxes in the updated objects section.

You should give updated objects with bounding boxes according to the criteria below: Images are 1x1 with top-left at [0, 0] and bottom-right at [1, 1]. Object bounding boxes should be presented like object name [Top-left x, Top-left y, Bottom-right x, Bottom-right y]. Pay attention that the minimum size for each bounding box is 0.1 x 0.1. Pay attention that a bounding box do not overlap more than the half of another bounding box, or more than the half of a bounding box is not inside another bounding box.

Please refer to the examples below for the desired format. Pay attention to the given format for updated objects and please do not include any other explanations:

User prompt: A realistic image of landscape scene depicting a green car parking on the left of a blue truck, with a red air balloon and a bird in the sky Current Objects: ['green car [0.027, 0.365, 0.275, 0.207]', 'blue truck [0.350, 0.368, 0.272, 0.208]', 'red air balloon [0.086, 0.010, 0.189, 0.176]'] Reasoning: To add a bird in the sky as per the prompt, ensuring all coordinates and dimensions remain within [0, 1]. Updated Objects: ['green car [0.027, 0.365, 0.275, 0.207]', 'blue truck [0.350, 0.369, 0.272, 0.208]', 'red air balloon [0.086, 0.010, 0.189, 0.176]', 'bird [0.385, 0.054, 0.186, 0.130]']

User prompt: A realistic image of landscape scene depicting a green car parking on the right of a blue truck, with a red air balloon and a bird in the sky Current Output Objects: ['green car [0.027, 0.365, 0.275, 0.207]', 'blue truck [0.350, 0.369, 0.272, 0.208]', 'red air balloon [0.086, 0.010, 0.189, 0.176]'] Reasoning: The relative positions of the green car and blue truck do not match the prompt. Swap positions of the green car and blue truck to match the prompt, while keeping all coordinates and dimensions within [0, 1]. Updated Objects: ['green car [0.350, 0.369, 0.275, 0.207]', 'blue truck [0.027, 0.365, 0.272, 0.208]', 'red air balloon [0.086, 0.010, 0.189, 0.176]', 'bird [0.485, 0.054, 0.186, 0.130]']

User prompt: An oil painting of a pink dolphin jumping on the left of a steam boat on the sea Current Objects: ['steam boat [0.302, 0.293, 0.335, 0.194]', 'pink dolphin [0.027, 0.324, 0.246, 0.160]', 'blue dolphin [0.158, 0.454, 0.376, 0.290]'] Reasoning: The prompt mentions only one dolphin, but two are present. Thus, remove one dolphin to match the prompt, ensuring all coordinates and dimensions stay within [0, 1]. Updated Objects: ['steam boat [0.302, 0.293, 0.335, 0.194]', 'pink dolphin [0.027, 0.324, 0.246, 0.160]']

User prompt: An oil painting of a pink dolphin jumping on the left of a steam boat on the sea Current Objects: ['steam boat [0.302, 0.293, 0.335, 0.194]', 'dolphin [0.027, 0.324, 0.246, 0.160]'] Reasoning: The prompt specifies a pink dolphin, but there's only a generic one. The attribute needs to be changed. Updated Objects: ['steam boat [0.302, 0.293, 0.335, 0.194]', 'pink dolphin [0.027, 0.324, 0.246, 0.160]']

User prompt: A realistic photo of a scene with a brown bowl on the right and a gray dog on the left Current Objects: ['gray dog [0.186, 0.592, 0.449, 0.408]', 'brown bowl [0.376, 0.194, 0.624,

0.502]'] Reasoning: The leftmost coordinate (0.186) of the gray dog's bounding box is positioned to the left of the leftmost coordinate (0.376) of the brown bowl, while the rightmost coordinate (0.186 + 0.449) of the bounding box has not extended beyond the rightmost coordinate of the bowl. Thus, the image aligns with the user's prompt, requiring no further modifications. Updated Objects: ['gray dog [0.186, 0.592, 0.449, 0.408]', 'brown bowl [0.376, 0.194, 0.624, 0.502]']

.....

A.3 Akıl Yürütme için MLLM Komutu

""You are an expert text-to-image evaluator. I will give you an image with a text prompt. Your objective is to evaluate how well the image aligns with the given text prompt. First, analyze the user input and identify the objects and relationships that should be in the scene. Then examine the given image and check if objects and relations in your analysis exist. Finally, give a score from 1 to 5, while scoring take the following criteria into account: 1: the image did not depict any elements or actions/relations in the caption. 2: the image depicted some elements, but ignored the actions/relations in the caption. 3: the image portrayed most of the elements and at least one action/relation in the caption. 4: the image depicted all of the elements and most of the actions/relations in the caption. 5: the image portrayed all of the elements and all of the actions/relations in the caption.

Please explain the reasoning behind your evaluation in maximum 20 words before scoring. Refer to the examples below for the desired format. Pay attention that do not include any other explanations in the score section.

User prompt: A girl is planting flowers and there is a watering can. Reasoning: Girl, flowers and watering can are visible. But the girl is not seems to be planting. Score: 4

User prompt: A black cat meanders up the walkway in front of a motorcycle in front of a house. Reasoning: The image shows a motorcycle and a house but lacks the black cat. Score: 3

User prompt: A beautiful woman sitting on a bench next to a stone building. Reasoning: The image features a woman, but lacks a bench and the stone building context. Score: 2

User prompt: A piece of dutch chocolate cake with a fork on a plate. Reasoning: The cake and fork are presented on a plate, perfectly aligned with the text prompt. Score: 5

User prompt: A child is holding a magnifying glass and examining a bug. Reasoning: Image shows a child with an object in his hand depicts a magnifying glass, but no bug is visible. Score: 3

User prompt: A woman is holding a present and walking towards a car. Reasoning: Image is including a woman with a present, directed towards a car. Perfectly aligning with the prompt. Score: 5

User prompt: A man is holding a map and navigating a boat on a lake. Reasoning: Image portrays a man in a boat on a lake, a map is also visible but the man is not holding it. Score: 4

User prompt: ""

A.4 Hızalama için MLLM Komutu

""You are an expert Text-to-Image Aligner. I will give you a text prompt with an image. Your objective is designing a layout to obtain a more compatible image with the user prompt.

Here is what need to you do step-by-step: First you should read and understand the user prompt. After that you should review the current image. Pay attention to the locations and relationships of

the objects. Then evaluate the alignment: If the image perfectly aligns with the user prompt please leave the updated objects section blank. If the image does not align with the user prompt then design a layout to adjust the locations and relationships.

Instructions for the layout is given below: If an object in the image lacks specified attributes and/or actions update it with its attribute and/or action. Please pay attention to not include another object. If some objects do not show the relation between them please carefully design their locations as they overlap but do not cover the most part of each other. If necessary you can do operations like object addition, deletion and repositioning.

You should give updated objects with bounding boxes according to the criteria below: Images are 1x1 with top-left at [0, 0] and bottom-right at [1, 1]. Bounding boxes should be presented like object name [Top-left x, Top-left y, Bottom-right x , Bottom-right y] and should not include more than one object. Pay attention that the minimum size for each bounding box is 0.1 x 0.1. Pay attention that a bounding box do not overlap more than the half of another bounding box, or more than the half of a bounding box is not inside another bounding box. Please justify the reasons behind your design in maximum 50 words, then give the layout in the updated objects section.

Pay attention to the format given below for the updated objects. Please do not include additional explanation in the updated objects section.

User prompt: A person is holding a baby kangaroo and feeding it from a bottle. Reasoning: The prompt specifies that a person is holding a baby kangaroo and feeding it from a bottle, so their bounding boxes should overlap, indicating the interaction between them. Updated Objects: ['a person [0.059, 0.137, 0.475, 0.898]', 'a baby kangaroo [0.41, 0.684, 0.625, 0.898]', 'a bottle [0.46, 0.742, 0.723, 0.821']]

User prompt: An oil painting of a pink dolphin jumping on the left of a steam boat on the sea
Reasoning: The image specifies a pink dolphin, but the jumping action needs to be added. Updated Objects: ['steam boat [0.302, 0.293, 0.335, 0.194]', 'a jumping pink dolphin [0.027, 0.324, 0.246, 0.160']']

User prompt: A woman is holding a basket of laundry and heading to the washing machine. Reasoning: The prompt specifies a woman holding a basket of laundry and heading toward the washing machine. The current image needs to reflect the relation between the woman and the basket. Updated Objects: ['a woman [0.059, 0.293, 0.391, 0.781]', 'a basket of laundry [0.291, 0.586, 0.723, 0.859]', 'a washing machine [0.742, 0.508, 0.957, 0.801']']

User prompt: A dog is fetching a stick and bringing it back to its owner. Reasoning: The prompt specifies that the dog is fetching and bringing back a stick to its owner, but there is no owner in the image. The dog's bounding box also needs slight adjustment to show the relation with the stick. Updated Objects: ['a dog [0.195, 0.391, 0.391, 0.566]', 'a stick [0.293, 0.508, 0.391, 0.566]', 'a person [0.600, 0.300, 0.800, 0.500']']

User prompt: A child is playing with a toy doctor's kit and checking their stuffed animals' health. Reasoning: To ensure the image accurately reflects the user's prompt, an additional stuffed animal needs to be added, as the prompt specifies "stuffed animals", implying more than one. Updated Objects: ['a child [0.059, 0.234, 0.391, 0.625]', 'a toy doctor's kit [0.41, 0.645, 0.615, 0.879]', 'a stuffed animal [0.684, 0.488, 0.879, 0.684]', 'a stuffed animal [0.200, 0.100, 0.350, 0.300']']

User prompt: The baby is crawling towards the toy. Reasoning: The image currently shows the baby not oriented towards the toy. The baby's position should be adjusted to face the toy, implying crawling motion toward it. Updated Objects: ['a crawling baby [0.151, 0.403, 0.352, 0.652]', 'a toy [0.551, 0.506, 0.653, 0.606']']

.....

B BÖLÜM 6'DA ÖNERİLEN VERİ KÜMESİ

1. A woman is holding a shopping bag and admiring a new dress.
2. A woman is holding a tray of cookies and offering them to guests.
3. A man is holding a briefcase and rushing to catch a train.
4. A man is standing on a street corner and waiting for a bus.
5. A girl is sitting on a swing with a kitten nearby.
6. A woman is walking with her child while a dog is passing by.
7. A student is writing notes to their notebook from the board.
8. A woman is holding a baby while a baby stroller nearby.
9. A father is walking with his daughter while a kite flies in the sky.
10. A sister holds her brother's hand as a car passes them.
11. A man is sitting under a tree while a bird is in the tree.
12. A child is on a rocking horse and a cat is walking by.
13. A woman is holding an umbrella and running towards a bus.
14. A boy is throwing a stick and playing with his dog.
15. A boy is shooting a soccer ball to goal.
16. A baker is holding a cake next to an oven.
17. A man is fixing a bicycle with a tire pump nearby.
18. A girl is planting flowers and a blank pot is around there.
19. A woman is setting up a tripod and there are lights around.
20. A man is assembling a tent and there is a backpack around.
21. A child is holding a guitar and there is a music stand nearby.
22. A woman is holding a gift and walking towards a car.
23. A boy is flying a kite and there is a tree nearby.
24. A man is fixing a bicycle and a toolbox is lying beside him.
25. A woman is watering flowers and there's a bench under the tree.
26. A student is typing on a laptop and has a notebook open next to it.
27. A chef is grilling meat and there's a bowl of salad on the table.
28. A child is reading a book and a lamp is glowing beside the chair.
29. A painter is cleaning brushes and a canvas is leaning against the wall.
30. A man is pouring coffee and a newspaper is spread out on the table.
31. A woman is jogging with headphones and there's a water bottle on the bench.
32. A boy is kicking a soccer ball and a goalpost is standing nearby.
33. A teacher is writing on the board and there's a globe on her desk.
34. A man is polishing his car and a bucket of water is next to him.
35. A girl is feeding a rabbit and there's a basket of carrots nearby.
36. A child is painting a picture and a jar of water is on the table.

37. A woman is sewing a dress and a mannequin is standing by the window.
38. A man is fishing in a lake and there's a cooler on the dock.
39. A boy is climbing a tree and a rope swing is hanging nearby.
40. A chef is chopping onions and a pot is simmering on the stove.
41. A photographer is adjusting the camera and a light reflector is set up.
42. A student is highlighting a textbook and a cup of coffee is steaming nearby.
43. A father is pushing a stroller and a diaper bag is hanging from the handle.
44. A woman is vacuuming the floor and a pile of magazines is on the couch.
45. A boy is riding a skateboard and a backpack is resting on a bench.
46. A gardener is trimming hedges and a watering can is placed on the grass.
47. A man is installing shelves and a toolbox is open on the floor.
48. A child is building a sandcastle and a plastic shovel is lying next to it.
49. A woman is brushing her hair and a makeup kit is on the dresser.
50. A man is lifting weights and a towel is draped over a chair.
51. A boy is launching a model rocket and a control pad is in his hands.
52. A woman is wrapping a scarf and a mirror is hanging on the wall.
53. A painter is outlining a mural and a ladder is propped nearby.
54. A chef is plating a dish and a basket of bread is on the counter.
55. A man is reading a map and a suitcase is by his feet.
56. A girl is blowing up balloons and a box of decorations is on the floor.
57. A child is bouncing a basketball and a water bottle is on the bench.
58. A woman is lighting candles and a bouquet of flowers is on the table.
59. A boy is adjusting his helmet and a skateboard is on the ground.
60. A man is planting seeds and a rake is leaning against the fence.
61. A man is washing his car and there's a motorcycle parked nearby.
62. A woman is setting up a tent and there's a kayak by the riverbank.
63. A boy is riding a bicycle and a scooter is leaning against a fence.
64. A gardener is watering bushes and a wheelbarrow is resting beside the garden.
65. A hiker is pitching a tent and a canoe is floating on the lake.
66. A farmer is loading hay onto a trailer and a tractor is parked nearby.
67. A woman is arranging chairs and a large table is standing in the center.
68. A worker is painting a wall and a ladder is set up against the building.
69. A man is repairing a truck and a trailer is hitched behind it.
70. A firefighter is spraying water and a fire engine is parked close by.
71. A child is playing inside a treehouse and a swing set is standing nearby.
72. A sailor is securing ropes on a sailboat and a dinghy is floating beside it.
73. A woman is cleaning windows and a balcony is attached to the building.
74. A construction worker is moving bricks and a cement mixer is running beside him.
75. A delivery driver is unloading boxes and a van is parked at the curb.
76. A cyclist is checking the tires and a parked car is next to the bike rack.
77. A man is mowing the lawn and a shed is standing at the edge of the yard.
78. A woman is parking her car and a shopping cart is left near the entrance.
79. A boy is climbing a jungle gym and a slide is set up beside it.
80. A surfer is waxing his board and a lifeguard tower is on the beach.
81. A camper is lighting a campfire and a caravan is parked behind the trees.
82. A worker is stacking crates and a forklift is parked in the warehouse.
83. A boy is flying a drone and a basketball hoop is set up on the court.
84. A man is boarding a bus and a bicycle rack is attached to the front.

85. A woman is jogging on a track and a set of bleachers is nearby.
86. A pilot is inspecting a small airplane and a hangar is open behind it.
87. A mechanic is changing a tire and a tow truck is parked on the roadside.
88. A skier is adjusting his boots and a ski lift is running up the mountain.
89. A zookeeper is feeding elephants and a large water trough is next to them.
90. A lifeguard is watching swimmers and a rescue boat is anchored nearby.
91. A man is washing a car and a van is parked nearby.
92. A woman is assembling a tent and a bicycle is leaning against a tree.
93. A boy is climbing a tree and a bench is standing below.
94. A farmer is driving a tractor and a barn is in the background.
95. A girl is painting a fence and a ladder is resting against the wall.
96. A hiker is setting up a hammock and a cabin is nearby.
97. A man is rowing a boat and a pier is extending into the lake.
98. A worker is lifting a box and a pallet is stacked in the corner.
99. A boy is kicking a football and a goalpost stands at the edge of the field.
100. A woman is opening a car door and a trailer is hitched behind.
101. A child is climbing a slide and a sandbox is beside the play area.
102. A mechanic is fixing a motorcycle and a pickup truck is parked nearby.
103. A gardener is trimming a hedge and a greenhouse is behind it.
104. A skier is tightening his boots and a ski lift is running nearby.
105. A sailor is cleaning a sailboat and a dinghy is tied alongside.
106. A driver is filling a truck with gas and a bus is parked at the station.
107. A woman is planting flowers in a garden and a bench is beside the path.
108. A builder is stacking bricks and a cement mixer is idle nearby.
109. A boy is riding a scooter and a basketball hoop is nearby.
110. A photographer is adjusting a tripod and a backdrop is hanging behind.
111. A firefighter is testing a hose and a fire truck is stationed nearby.
112. A man is painting a wall and a scaffold is standing beside it.
113. A camper is lighting a firepit and a tent is pitched close by.
114. A girl is riding a horse and a stable is in the background.
115. A man is cleaning a window and a ladder is standing by the wall.
116. A cyclist is repairing a flat tire and a parked car is by the curb.
117. A woman is locking her bicycle and a streetlamp is on the corner.
118. A boy is throwing a frisbee and a picnic table is nearby.
119. A zookeeper is washing an elephant and a water tank is nearby.
120. A man is checking a trailer hitch and a parked truck is beside him.
121. A man is parking a motorcycle and a car is parked nearby.
122. A woman is adjusting a satellite dish and a shed is standing beside the house.
123. A boy is inflating a tire and a bicycle is leaning against a pole.
124. A lifeguard is checking a rescue boat and a flagpole is beside the tower.
125. A farmer is feeding cows and a tractor is parked in the field.
126. A painter is cleaning a ladder and a scaffold is standing next to the wall.
127. A worker is unloading crates and a forklift is stationed nearby.
128. A hiker is rolling up a sleeping bag and a tent is pitched close by.
129. A sailor is lowering an anchor and a buoy is floating nearby.
130. A man is starting a generator and a trailer is parked behind him.
131. A child is climbing a rope and a slide is next to the playground.
132. A mechanic is checking the oil in a car and a motorcycle is parked beside it.

133. A gardener is raking leaves and a wheelbarrow is sitting on the path.
134. A firefighter is coiling a hose and a hydrant is standing on the sidewalk.
135. A boy is flying a drone and a basketball hoop is on the court.
136. A woman is arranging patio chairs and a grill is set up nearby.
137. A man is cleaning a swimming pool and a diving board is at the edge.
138. A girl is riding a pony and a fence is surrounding the field.
139. A delivery worker is loading packages and a van is parked at the curb.
140. A cyclist is pumping air into the tires and a bench is nearby.
141. A man is moving a piano and a bookshelf is against the wall.
142. A boy is paddling a kayak and a dock is extending into the water.
143. A photographer is focusing a camera and a light stand is beside him.
144. A worker is painting a gate and a lamppost is at the corner.
145. A woman is washing a horse and a stable is behind her.
146. A student is unlocking a bike and a bus stop is nearby.
147. A sailor is folding a sail and a mast is standing tall.
148. A man is stacking firewood and a log cabin is in the background.
149. A gardener is watering bushes and a statue is standing in the yard.
150. A mechanic is jacking up a car and a van is parked beside it.
151. A man is unloading a kayak and a camper van is parked nearby.
152. A woman is securing a bike to a rack and a bench is beside the path.
153. A boy is pitching a tent and a picnic table is set up nearby.
154. A gardener is pruning a tree and a fountain is in the center of the yard.
155. A firefighter is inspecting a ladder and a fire truck is parked nearby.
156. A sailor is hoisting a sail and a lighthouse stands on the shore.
157. A man is towing a trailer and a van is parked next to the road.
158. A woman is cleaning a barbecue grill and a table is under the gazebo.
159. A boy is climbing a rock wall and a flag is waving at the top.
160. A mechanic is adjusting a car engine and a toolbox is resting nearby.
161. A worker is painting shutters and a ladder is leaning on the wall.
162. A child is playing inside a treehouse and a swing is hanging nearby.
163. A man is filling a water tank and a truck is parked beside it.
164. A woman is packing a car trunk and a suitcase is standing on the driveway.
165. A student is locking a scooter and a bicycle is parked next to it.
166. A farmer is loading hay onto a wagon and a barn is behind him.
167. A painter is mixing colors on a palette and an easel is standing nearby.
168. A lifeguard is securing a rescue board and a tower is overlooking the beach.
169. A man is washing windows and a balcony is above him.
170. A cyclist is adjusting handlebars and a bike rack is bolted to the sidewalk.
171. A gardener is planting shrubs and a trellis is set up nearby.
172. A zookeeper is feeding giraffes and an enclosure is behind them.
173. A camper is setting up chairs and a firepit is ready for use.
174. A woman is locking her car and a bicycle is parked at the rack.
175. A boy is throwing a football and a set of bleachers is behind the field.
176. A hiker is unpacking a backpack and a cabin is in the distance.
177. A man is wiping down a boat and a dock stretches into the water.
178. A firefighter is attaching a hose and a fire hydrant is at the corner.
179. A child is pedaling a tricycle and a slide is in the playground.
180. A woman is hanging laundry and a shed is in the backyard.

181. A man is loading bikes onto a rack and a minivan is parked nearby.
182. A woman is sweeping the porch and a rocking chair is placed beside the door.
183. A boy is paddling a canoe and a dock is stretching out into the lake.
184. A worker is securing a ladder and a scaffolding is standing by the wall.
185. A child is climbing a jungle gym and a merry-go-round is nearby.
186. A sailor is folding a sail and a buoy is floating in the water.
187. A woman is unlocking a gate and a shed is standing in the yard.
188. A man is starting a chainsaw and a stack of logs is piled up.
189. A cyclist is fixing a chain and a row of parked bikes is beside the rack.
190. A gardener is mulching a flower bed and a trellis is next to the fence.
191. A firefighter is loading hoses onto a truck and a ladder is extended nearby.
192. A photographer is setting up a light stand and a backdrop is hanging behind.
193. A boy is balancing on a beam and a basketball hoop is standing nearby.
194. A painter is sketching on a canvas and a stool is set up beside him.
195. A woman is washing her SUV and a boat trailer is parked nearby.
196. A mechanic is replacing a tire and a van is on the next lift.
197. A man is tying a boat to the dock and a fishing net is coiled nearby.
198. A worker is sanding a wooden door and a ladder is standing against the house.
199. A boy is riding a dirt bike and a wooden ramp is set up nearby.
200. A woman is decorating a large tree and a bench is beneath the branches.
201. A zookeeper is hosing down the elephant pen and a water tank is next to the gate.
202. A gardener is planting a row of bushes and a garden arch is nearby.
203. A camper is folding a tent and a canoe is pulled up on the shore.
204. A lifeguard is adjusting binoculars and a rescue board is propped against the tower.
205. A man is mowing a soccer field and a set of goalposts is at the far end.
206. A woman is unpacking her car and a picnic table is under the trees.
207. A cyclist is checking the brakes and a parking sign is beside the rack.
208. A boy is tossing a basketball and a metal fence encloses the court.
209. A sailor is cleaning the deck and a life preserver is hanging on the railing.
210. A man is loading a kayak onto a truck and a paddleboard is leaning against the wall.
211. A man is folding a tarp and a trailer is parked beside the campsite.
212. A woman is polishing a motorcycle and a pickup truck is parked nearby.
213. A boy is climbing a ladder and a playhouse is standing at the top.
214. A gardener is watering hedges and a stone fountain is in the middle of the yard.
215. A hiker is packing a backpack and a cabin is visible nearby.
216. A firefighter is spraying water on a building and a ladder truck is parked behind him.
217. A sailor is untying ropes from a sailboat and a pier stretches into the water.
218. A man is washing a trailer and a large SUV is parked next to it.
219. A boy is adjusting his helmet and a mountain bike is standing nearby.
220. A mechanic is fixing a car engine and a tow truck is waiting outside.
221. A woman is dusting a bookshelf and a sofa is placed against the wall.
222. A construction worker is hammering wooden planks and scaffolding is set up nearby.
223. A man is boarding a train and a luggage cart is standing on the platform.
224. A camper is starting a barbecue grill and a picnic table is next to the tent.
225. A boy is bouncing a basketball and bleachers are beside the court.
226. A cyclist is cleaning the wheels and a parked bus is behind the rack.
227. A woman is stretching on a yoga mat and a bench is by the wall.
228. A sailor is coiling a rope and a lifeboat is tied up nearby.

229. A man is organizing boxes in a van and a hand truck is beside it.
230. A firefighter is testing the siren and a rescue ladder is mounted on the truck.
231. A gardener is laying down turf and a gazebo is standing nearby.
232. A boy is climbing monkey bars and a slide is part of the playground.
233. A zookeeper is opening a gate and a feeding trough is inside the enclosure.
234. A woman is vacuuming the car interior and a garage door is open behind her.
235. A man is parking a forklift and a stack of pallets is in the corner.
236. A cyclist is oiling the chain and a streetlamp is standing nearby.
237. A sailor is mopping the deck and a radar tower is on the ship.
238. A boy is raking leaves and a treehouse is in the big oak tree.
239. A worker is sanding a door and a scaffold is standing by the wall.
240. A woman is hanging up a sign and a bicycle rack is near the entrance.
241. A man is loading a generator onto a truck and a trailer is hitched behind.
242. A woman is unfolding a picnic blanket and a large tree is shading the spot.
243. A boy is pumping air into a soccer ball and a set of goalposts is nearby.
244. A worker is stacking boxes in a warehouse and a forklift is parked at the side.
245. A hiker is unfolding a map and a trail marker is beside the path.
246. A farmer is securing a fence and a water trough is in the field.
247. A sailor is lowering a dinghy into the water and a sailboat is anchored nearby.
248. A boy is climbing onto a horse and a stable is in the background.
249. A gardener is trimming a hedge and a stone bench is along the path.
250. A firefighter is checking the ladder and a rescue truck is parked beside it.
251. A woman is rolling up a yoga mat and a bench is against the wall.
252. A man is attaching a kayak to a roof rack and a minivan is parked in the driveway.
253. A camper is lighting a lantern and a tent is pitched under a tree.
254. A mechanic is aligning car wheels and a stack of tires is nearby.
255. A boy is polishing his bike and a ramp is set up in the yard.
256. A woman is setting up a projector and a whiteboard is behind her.
257. A worker is mixing cement and a pile of bricks is stacked beside him.
258. A photographer is packing a camera bag and a lighting rig is standing nearby.
259. A man is coiling a hose and a garden shed is behind him.
260. A sailor is adjusting the mast and a lifeboat is secured at the stern.
261. A zookeeper is opening a transport cage and a feeding area is visible behind.
262. A cyclist is replacing a tire and a repair stand is next to him.
263. A gardener is planting a tree and a bench is positioned nearby.
264. A boy is skating on a ramp and a chain-link fence surrounds the park.
265. A woman is scrubbing a patio table and a barbecue grill is standing nearby.
266. A man is leveling paving stones and a wheelbarrow is resting on the path.
267. A worker is wrapping plastic around pallets and a truck is backed up to the dock.
268. A child is climbing a rope ladder and a slide is next to the playground.
269. A man is parking a snowmobile and a shed is at the edge of the lot.
270. A woman is painting a fence and a pergola is standing nearby.
271. A man is loading firewood into a truck and a cabin is standing nearby.
272. A woman is washing a horse and a hay bale is stacked by the fence.
273. A boy is kicking a soccer ball and a bicycle is parked beside the field.
274. A worker is setting up a ladder and a cherry picker is parked nearby.
275. A camper is pitching a tent and a canoe is resting on the shore.
276. A sailor is tying a rope to the dock and a lighthouse is standing nearby.

277. A gardener is pruning a vine and a garden arch is overhead.
278. A boy is climbing onto a quad bike and a trailer is hitched behind.
279. A photographer is mounting a camera and a light box is beside the stand.
280. A firefighter is loading gear onto a truck and a water tank is positioned nearby.
281. A man is cleaning a barbecue grill and a patio table is under an umbrella.
282. A woman is unzipping a tent and a picnic table is next to the site.
283. A child is jumping on a trampoline and a treehouse is in the branches.
284. A sailor is folding a tarp and a lifeboat is secured to the railing.
285. A mechanic is replacing brake pads and a jack stand is supporting the car.
286. A worker is unloading timber and a cement mixer is parked beside the site.
287. A woman is spraying paint on a wall and a scaffold is set up nearby.
288. A man is tightening bolts on a bike and a streetlamp is on the corner.
289. A boy is stacking wood and a stone fireplace is built nearby.
290. A gardener is weeding a flower bed and a fountain is bubbling in the yard.
291. A man is adjusting a satellite dish and a tool chest is beside him.
292. A woman is hanging up a banner and a ladder is leaning against the wall.
293. A child is riding a pony and a hitching post is nearby.
294. A camper is unrolling a sleeping bag and a campfire ring is set up beside the tent.
295. A photographer is changing lenses and a light reflector is propped against a chair.
296. A man is laying down stepping stones and a bench is in the corner of the yard.
297. A cyclist is pumping tires and a metal railing lines the path.
298. A sailor is checking navigation equipment and a buoy is floating nearby.
299. A woman is dusting shelves and a large wardrobe is against the wall.
300. A mechanic is securing a bumper and a workbench is set up beside the car.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. Melike Nur Yeğin and Mehmet Fatih Amasyalı, "Generative diffusion models: A survey of current theoretical developments", Neurocomputing, Volume 608, 2024, 128373, ISSN 0925-2312,
<https://doi.org/10.1016/j.neucom.2024.128373>

Konferans Bildirisi

1. M. N. Yeğin and M. F. Amasyalı, "Modality Weighting in Multimodal Variational Autoencoders," 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), Antalya, Turkey, 2022, pp. 1-6,
<https://doi.org/10.1109/ASYU56188.2022.9925305>