Unified Multimodal Discrete Diffusion

Alexander Swerdlow*   Mihir Prabhudesai*   Siddharth Gandhi

Deepak Pathak   Katerina Fragkiadaki

Carnegie Mellon University

Abstract

Multimodal generative models that can understand and generate across multiple modalities are dominated by autoregressive (AR) approaches, which process tokens sequentially from left to right, or top to bottom. These models jointly handle images, text, video, and audio for various tasks such as image captioning, question answering, and image generation. In this work, we explore discrete diffusion models as a unified generative formulation in the joint text and image domain, building upon their recent success in text generation. Discrete diffusion models offer several advantages over AR models, including improved control over quality versus diversity of generated samples, the ability to perform joint multimodal inpainting (across both text and image domains), and greater controllability in generation through guidance. Leveraging these benefits, we present the first **Uni**fied Multimodal **Disc**rete Diffusion (UniDisc) model which is capable of jointly understanding and generating text and images for a variety of downstream tasks. We compare UniDisc to multimodal AR models, performing a scaling analysis and demonstrating that UniDisc outperforms them in terms of both performance and inference-time compute, enhanced controllability, editability, inpainting, and flexible trade-off between inference time and generation quality. Code and additional visualizations are available at `https://unidisc.github.io`.

1. Introduction

Multimodal generative models—which can understand and generate a variety of modalities such as text, images, videos, and audio – can significantly improve the capabilities of an AI system, as these models can (1) leverage information from multiple sources to better understand the context (2) learn from any available data source, and (3) respond to a user's request in a flexible manner, thus dynamically generating text, images, or audio as required. Although the choice of model architecture—transformers—is currently clear, the optimal generative objective remains unclear.

Current multimodal models are typically trained jointly using (an approximation to) a maximum likelihood objective over sequences consisting of images, text, and other modalities. Autoregressive (AR) models typically quantize data from continuous modalities and optimize

---

[1]*Equal contribution
[2]Correspondence to: aswerdlow@cmu.edu, mprabhud@andrew.cmu.edu

Figure 1 | **We show UniDisc's ability to jointly inpaint image & text pairs**. We do not explicitly optimize for this objective but it is intrinsic to UniDisc's unified diffusion objective.

the exact likelihood through a series of conditionals; during generation, they use a fixed token order, e.g., left-to-right, top-to-bottom (raster order) for images. They have demonstrated strong performance in both text and image generation, making them the current workhorse for multimodal models. However, generating image tokens autoregressively is slow and wasteful as nearby tokens are highly correlated, and this process results in many unnecessary forward passes through the network Lu et al. (2022); Team (2024b); Team et al. (2023). Moreover, AR models are difficult to control Li et al. (2022), cannot inpaint or infill unless explicitly trained to, and cannot easily trade-off quality versus compute at inference time.

On the other hand, continuous diffusion models—which have been shown to work well for continuous modalities such as images—have fast inference, are highly controllable, and can easily trade-off quality vs. compute. These models corrupt data by adding Gaussian noise and are trained to denoise the data, maximizing a lower bound on the likelihood. However, these models have found to be significantly slower to train in text domain compared to AR models (by roughly 64 times) Gulrajani and Hashimoto (2024). Text is inherently discrete, and adding continuous Gaussian noise to text token embeddings does not correspond to meaningful changes in the actual text. These trade-offs between different modeling strategies across modalities raises the question: what is the right unified generative formulation across text, image, and other modalities?

To address this, we present UniDisc, a unified multimodal model based on discrete diffusion. While continuous Gaussian noise is inefficient with discrete data such as text and graphs, UniDisc corrupts data with discrete noise—specifically, randomly masking tokens—and learns to map mask tokens into multimodal tokens during inference. Discrete diffusion through masking has been explored separately for generating text Austin et al. (2021); Sahoo et al. (2024) and images Chang et al. (2022, 2023). Such explorations have resulted in different noise schedules, transition kernels, and loss functions across the text and image domains. In this paper, we explore a discrete diffusion formulation and its applicability in jointly modeling text and image modalities with a unified set of hyperparameters.

We propose a unified architecture that jointly tokenizes text and images, and uses full self-attention to learn to map a masked token sequence to a clean token sequence by sampling from a joint vocabulary of text and image tokens. We evaluate UniDisc across multimodal conditional and unconditional generation on multiple image-text datasets and compare to its AR counterpart. First, we find UniDisc achieves a higher FID and CLIP score than AR (Fig. 4), which we attribute to the effect of classifier-free guidance. We show that UniDisc exhibits strong joint image-text inpainting abilities that are not possible with prior unified generative models as seen in Fig. 1. Second, we find that UniDisc consistently outperforms its AR counterpart in inference efficiency: at a given inference compute budget, our model achieves generations of higher quality and diversity (Fig. 5). Third, we show UniDisc showcases stronger discriminative ability than AR on retrieval tasks due to its variable number of sampling steps (Fig. 7). Lastly, we scale UniDisc to a 1.4B parameter model, trained on web-scale image-text datasets.

Our code, model weights, and dataset are publicly available. More qualitative visualizations are available at `https://unidisc.github.io`.

2. Related Work

2.1. Unified Multi-Modal Models

In recent years, unified models for processing multiple modalities have advanced significantly. Models like Flamingo Alayrac et al. (2022) and PaLM-E Driess et al. (2023) demonstrate strong few-shot learning capabilities across tasks. LLAVA Liu et al. (2023) enhances LLaMa Touvron et al. (2023) with multimodal fine-tuning, but still uses separate encoders, limiting true unification and image generation. Recent efforts, like Perceiver IO Jaegle et al. (2021) and Unified-IO Lu et al. (2022), attempt modality unification but at a smaller scale. The Chameleon project Team (2024b) scales this up with a 34B parameter model trained on image-text data. However these approaches largely focus on autoregressive generation which is inefficient for high-dimensional data.

Relevant to our work, UniD3 Hu et al. (2023) considered discrete diffusion on image and text but made several design decisions that separated each modality, using both absorbing and uniform masking, decoupling the modalities inside the model with separate operations on each. Further we couldn't compare against their model—no training code is available and were unable to reproduce their reported results using their publicly available code.

2.2. Discrete Diffusion Models

Discrete diffusion models have emerged as a promising alternative to continuous diffusion for discrete data types. Sohl-Dickstein et al. introduced the first discrete diffusion model over binary variables, Hoogeboom et al. extended the noising process to categorical variables, demonstrating its effectiveness on image generation tasks. D3PM Austin et al. (2021) later extended discrete diffusion to a more general set of noising processes, allowing for more flexible noise schedules. Recent work by SEDD Lou et al. introduced score entropy, a novel loss function for discrete diffusion models that bridges the gap between continuous and discrete spaces, and more recently, Sahoo et al. (2024); Shi et al. (2024) showed text perplexity competitive with GPT-2. Most recently, Nie et al. (2024) looked at the scaling properties of discrete diffusion on text. While this approach shows promise for improving discrete diffusion models, these methods were primarily focused on language modeling tasks. Our work extends the application of discrete diffusion to multiple modalities and demonstrates its effectiveness in a unified architecture.

3. UniDisc: Unified Discrete Diffusion

3.1. Diffusion Models

Diffusion models Ho et al. (2020); Sohl-Dickstein et al.; Song et al. (2020) are a class of generative models that learn to construct a data distribution by gradually reversing a process that introduces noise into clean data samples. This approach models the transformation of a data sample x_0 from a clean state through increasingly noisy states until it reaches a pure noise distribution.

The forward diffusion process is described by a series of transitions, where each latent variable x_t at time step t is sampled from a Gaussian distribution as follows:

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

Here, $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ represents the cumulative product of noise levels, making x_t increasingly distant from x_0 as t increases. The variable x_t represents the noisy version of x_0 at time t, modeled to progressively approximate Gaussian noise as t approaches the final time step.

The reverse diffusion process then aims to reconstruct the original data by progressively denoising these samples. This involves learning the reverse transitions, with the goal to train the model $p_\theta(x_{t-1}|x_t)$ to approximate the true reverse process and effectively recover the original data point x_0 from the noisy samples.

Given T timesteps of diffusion, the loss using the Evidence Lower Bound (ELBO) for the diffusion process equals[1]:

$$\mathcal{L}_{\text{diff}} = \underbrace{-\mathbb{E}_{q(x_1|x_0)}\left[\log p_\theta(x_0|x_1)\right]}_{\text{reconstruction term}} + \tag{1}$$

$$\underbrace{\sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[D_{KL}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t))\right]}_{\text{denoising matching term}}$$

3.2. Discrete Diffusion Models

Building on the foundations of continuous diffusion models, discrete diffusion models adapt these concepts to structures that are inherently discrete. Unlike their continuous counterparts that model transitions of x_t given x_{t-1} with Gaussian distributions, discrete models define transitions using categorical distributions. The forward process for discrete models is thus characterized as:

$$q(x_t|x_0) = \text{Cat}(x_t; x_0 \cdot \bar{Q}_t) \tag{2}$$

Here, $\bar{Q}_t = \prod_{t=0}^{t=t} Q_t$ is a $N \times N$ matrix where N is the size of the vocabulary. $\bar{Q}_t$ represents the cumulative transition matrix at each discrete time step t, and Q_t is a transition matrix $[Q_t]_{ij} = q(x_t = j \mid x_{t-1} = i)$ dictating the probabilities of moving from one discrete state (a token in the vocabulary) x_{t-1} to another x_t discrete state (a token in the vocabulary), x_0 is a one-hot vector of the input data sample. D3PM Austin et al. (2021) generalizes this framework over various transition matrices (Q_t), the popular ones mainly include uniform and absorbing transition matrix. In UniDisc, we use the absorbing transition matrix as empirically it has been found to work the best across text and images Austin et al. (2021); Lou et al. (2024). Absorbing

---

[1]We skip the prior matching term from the loss as it is zero.
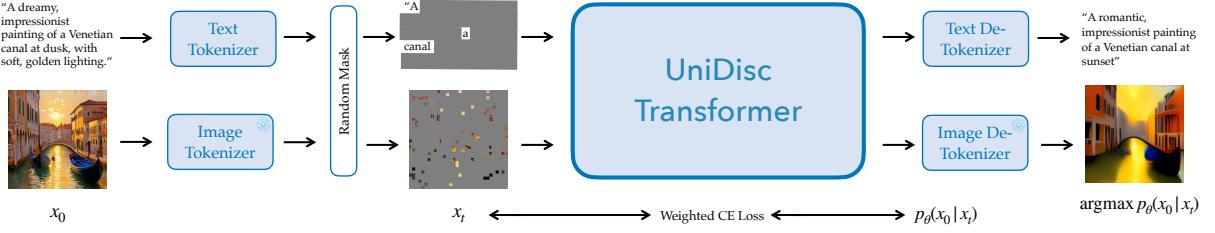
Figure 2 | UniDisc is a unified multimodal discrete diffusion model that can jointly process and generate text and images. Each modality is converted into a sequence of discrete tokens and we jointly denoise, supervising with a weighted cross-entropy loss. At inference time we begin with a set of [MASK] tokens and iteratively unmask tokens.

transition matrix requires having an absorbing state namely the [MASK] token. The matrix is represented as $Q_t = \alpha_t I + (1 - \alpha_t) \mathbb{1} e_m^T$, where $\mathbb{1}$ is a column vector of ones and e_m is a one-hot vector with one on the mask state m. This ends up being a matrix with all zeros except $i = j \neq m$ is α and $j = m, i \neq m$ is $1 - \alpha$ and $i = j = m$ is 1.

Intuitively this means that during the forward transition, the probability of an input token x_0 to stay the same is α, the probability of it being masked is $1 - \alpha$, and the probability of a masked token to be unmasked is 0.

Given the forward diffusion in Eq. (2), Sohl-Dickstein et al. uses the same objective function as Eq. (1) to optimize their model, where $q(x_{t-1}|x_t)$ ends up being a Bernoulli distribution instead of a Gaussian distribution. MDLM Sahoo et al. (2024) simplifies this objective function by considering continuous time-diffusion and applying loss only on the masked tokens. The final loss simply ends up being a re-weighted masked generative modeling loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t \sim \mathcal{U}(0,1), q(x_t|x)} \left[\frac{\alpha'_t}{1 - \alpha_t} \log p_\theta(x_0 \mid x_t) \right] \tag{3}$$

where $\alpha'_t = \alpha_t - \alpha_{t-1}$, and α_t is the probability of the token not being masked. MaskGIT Chang et al. (2022) and Muse, state-of-the-art masked image generative model use the same loss as Eq. (3), except there is no reweighting term and the time is discrete time instead of continuous time. The noising schedule α_t is also different, while language discrete diffusion models such as Austin et al. (2021); Sahoo et al. (2024) use a linear-time schedule, MaskGIT and Muse Chang et al. (2022, 2023) use a cosine schedule. We ablate these different design choices in Appendix F.

3.3. Unified Training via UniDisc

We train a bidirectional decoder-only transformer Vaswani et al. (2017) using 2D RoPE embeddings Liu et al. (2024) for all image tokens, 1D RoPE Su et al. (2023) embeddings for text tokens, and add learned modality-specific embeddings to each token. This allows our model both flexibility in resolution at inference, and the ability to use compute effectively by performing the majority of training at a lower resolution. We use the same objective function as Eq. (3), except for us x_0 is $[x_0^{img}, x_0^{txt}]$

Classifier-Free guidance (CFG) Ho and Salimans (2022) has been used in continuous diffusion models to trade-off between quality and diversity of generation. We apply this idea to discrete diffusion, with a probability of 0.1 we set all the tokens of a random modality to be mask tokens, this allows UniDisc to learn unconditional likelihood for image and text modality. During inference we use CFG for conditional generation (image-to-text or text-to-image) to trade-off

between quality and diversity of generation as shown in Fig. 5.

To improve training stability, we use Query-Key Normalization Wortsman et al. (2023) and use RMSNorm Zhang and Sennrich (2019) for all other norms. We use Sandwich Normalization—normalization before and after each FFN, as we found this helps control activations in deeper layers as previously reported in Ding et al. (2021); Zhuo et al. (2024).

To further improve the convergence speed of discrete diffusion we analyze the noising schedule and find that linear schedule in Austin et al. (2021); Sahoo et al. (2024) results in excessively high weighting for early timesteps, impairing the convergence speed. Following Min-SNR trick in continuous diffusion Hang et al. (2023), we limit the minimum weighting to 5. An architecture diagram is provided in Fig. 2 and pseudo-code for training procedure is provided in Appendix A.1.

3.4. Unified Sampling via UniDisc

Sampling in masked discrete diffusion, involves mapping a set of masked tokens m to a set of visible tokens x_0 using T timesteps of denoising. A variety of sampling strategies have been previously proposed Austin et al. (2021); Chang et al. (2022); Lou et al. (2024); Sahoo et al. (2024); Sohl-Dickstein et al.; Zheng et al. (2024) for masked discrete diffusion. MaskGIT Chang et al. (2022) proposes a confidence-based sampling, where they decode the most confident tokens at each step of denoising. D3PM Austin et al. (2021) and MLDM Sahoo et al. (2024) uses a sampling mechanism similar to Ho et al. (2020) except applied to a bernoulli distribution, which we refer to as DDPM sampling. This results in a random set of tokens being decoded, instead of the most confident ones as in MaskGIT. We ablate these sampling strategies in Fig. 5 and find the confidence-based sampling proposed in MaskGIT to work the best for unified modeling.

4. Experiments

We compare UniDisc against an autoregressive (AR) baseline across various tasks, metrics and datasets. We use the same architecture and hyper-parameters, and data, only differing in the attention mask and respective loss functions. For our autoregressive baseline we use a standard language model architecture from Chameleon Team (2024b)—that is a decoder-only transformer with causal attention and rotary positional embeddings. To enable classifier-free guidance, we dropout modalities with 10% probability during training. For UniDisc, we dropout both modalities and for the AR baseline we dropout only the first modality in the input sequence as in Liu et al. (2024).

Our experiments aim to answer the following questions:

1. How does UniDisc compare against AR models in unconditional and conditional multi-modal generation of image/text pairs?
2. How effective is classifier-free guidance in conditional generation for AR models and for UniDisc?
3. How does UniDisc compare against AR models in terms of training efficiency with varying the ratio of image-text tokens?
4. How do various sampling strategies for UniDisc affect its generation results and inference speed?
5. How does UniDisc compare against AR models across image-language reasoning tasks?
6. How do various design choices of UniDisc contribute to its performance?

Lastly, we show that we can successfully scale UniDisc, to a 1.4B parameter model, trained on
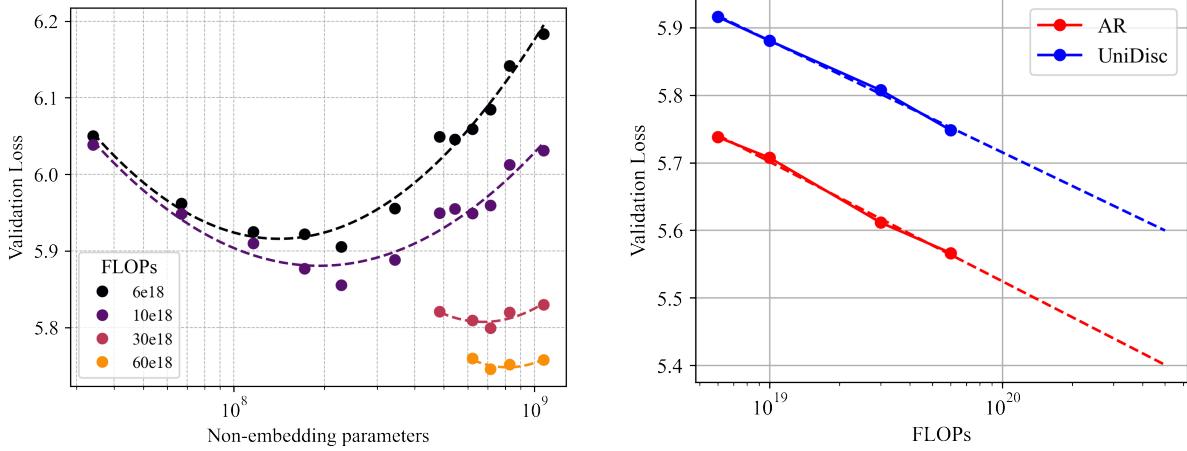
Figure 3 | **Scaling Analysis for AR and UniDisc models**: (Left) IsoFLOP curves for UniDisc, plotting varying model size for a fixed FLOP budget. (Right) Estimating optimal parameter size for each budget - minima of fitted parabola, we plot scaling laws for both AR and UniDisc. We find 13.2x more compute is required for UniDisc to achieve the same overall loss as AR.

500B tokens. We qualitatively evaluate this model, to demonstrate its capabilities.

Datasets: In Section 4.1, 4.2, and 4.3, we conduct experiments with different train and validation sets. Our training set includes DataComp1B Gadre et al. (2024), CC12M Changpinyo et al. (2021), CLEVR-math Lindström and Abraham (2022), and CLEVR-Ref Liu et al. (2019). Our evaluation datasets include a held-out validation set of DataComp1B and CC12M, along with Flickr Plummer et al. (2016), MS-COCO30k Chen et al. (2015), and Winoground Thrush et al. (2022).

4.1. Evaluation of Multimodal Generation

We evaluate UniDisc and AR models in unconditional and conditional generation tasks.

Evaluation metrics: We consider the following three evaluation metrics, most commonly used in previous works: **i) Joint perplexity** indicates a model's ability to fit to different validation sets. Note that this metric is jointly calculated across image-text tokens. The perplexity values from the autoregressive Chameleon baseline are exact likelihoods, the values from UniDisc are upper bounds. While perplexity is a good metric for assessing the fitting ability of a model, it cannot be used to evaluate its generation ability. **ii) Fréchet inception distance (FID)** Heusel et al. (2017) is a popular metric in image-generation to quantify the quality and diversity of image generation.**iii) CLIP score** is used for calculating image-text coherence. While we could not find an equivalent to the FID metric for text, we use CLIP score to evaluate generated image-text coherence, conditioning our model on an input image.

Experimental details:

We show conditional image-text generation results in Fig. 4. We condition on an image to generate the corresponding language description, and vice versa, condition on the language description to generate the corresponding image. For unconditional results please refer to Fig. 10 in the Appendix. We use a dataset comprising 30M image-text pairs from DataComp1B Gadre et al. (2024) and CC12M Changpinyo et al. (2021), please refer to Appendix B.1 for further details.

UniDisc significantly outperforms AR in conditional generation while performing equally well in unconditional generation). We attribute this performance gap to classifier-free guidance (CFG). As
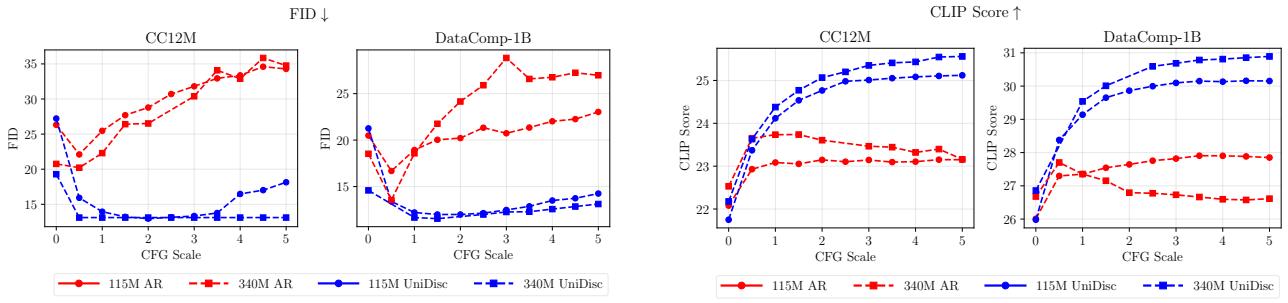
Figure 4 | **Conditional generation results for both FID and CLIP metrics, across a range of CFG values.** We find that AR is more sensitive to the CFG weighting, with a narrower optimal range.

can be seen in Fig. 4, the results without CFG (Scale=0) are similar between AR and UniDisc, but increasing CFG disproportionately benefits UniDisc.

The iterative generation process of diffusion makes it easy to blend conditional and unconditional predictions to guide the output. Autoregressive models, on the other hand, generate data sequentially in a fixed order, without any iterative refinement, which makes it difficult to mix in unconditional predictions to guide generation. We study this in detail in Section I in Appendix.

Joint Image-Text Inpainting In Fig. 1, we show that UniDisc can inpaint in a joint text and image space—without any fine-tuning. Currently, none of the popular generative models have this capability, because most multimodal generative models are either autoregressive Team (2024a) or use mixed modeling Zhou et al. (2024), which prevents them from easily inpainting jointly over image and text. In Appendix B.4 we explicitly fine-tune an AR model for joint inpainting, UniDisc zero-shot still shows far better performance. For more qualitative results, please refer to Fig. 18, and Fig. 19 in the Appendix.

4.2. Training Efficiency and Inference Speed

With the ever growing scale of recent generative models, an important aspect of their performance is their compute efficiency. Prior works Hoffmann et al. (2022); Kaplan et al. (2020) have extensively measured the training scaling laws of AR models, finding a power-law relationship between compute cost and distribution fitting, measured by negative log likelihood (NLL). In contrast, there has been little work that has measured the training efficiency of discrete diffusion models: the closest work is that of Gulrajani and Hashimoto (2024), which finds that the training efficiency of continuous diffusion models is approximately 64x worse than AR models on text. Recently, concurrent work Zheng et al. (2024) studied discrete diffusion models, again only on text, and found a scaling factor of 16x compared to AR models.

Although discrete diffusion is thought to be comparatively more efficient on other modalities such as images, we are not aware of prior work that has empirically measured this. We perform an ISOFlop analysis Hoffmann et al. (2022) of UniDisc and our AR baseline, changing only the attention mask and loss function. As in prior work, we select a set of compute budgets C_i and, within each budget, vary the non-embedding parameters (incl. LM head) N, and total tokens during training D, keeping the total compute, measured in FLOPs, fixed using an approximation of $C \approx 6ND$.

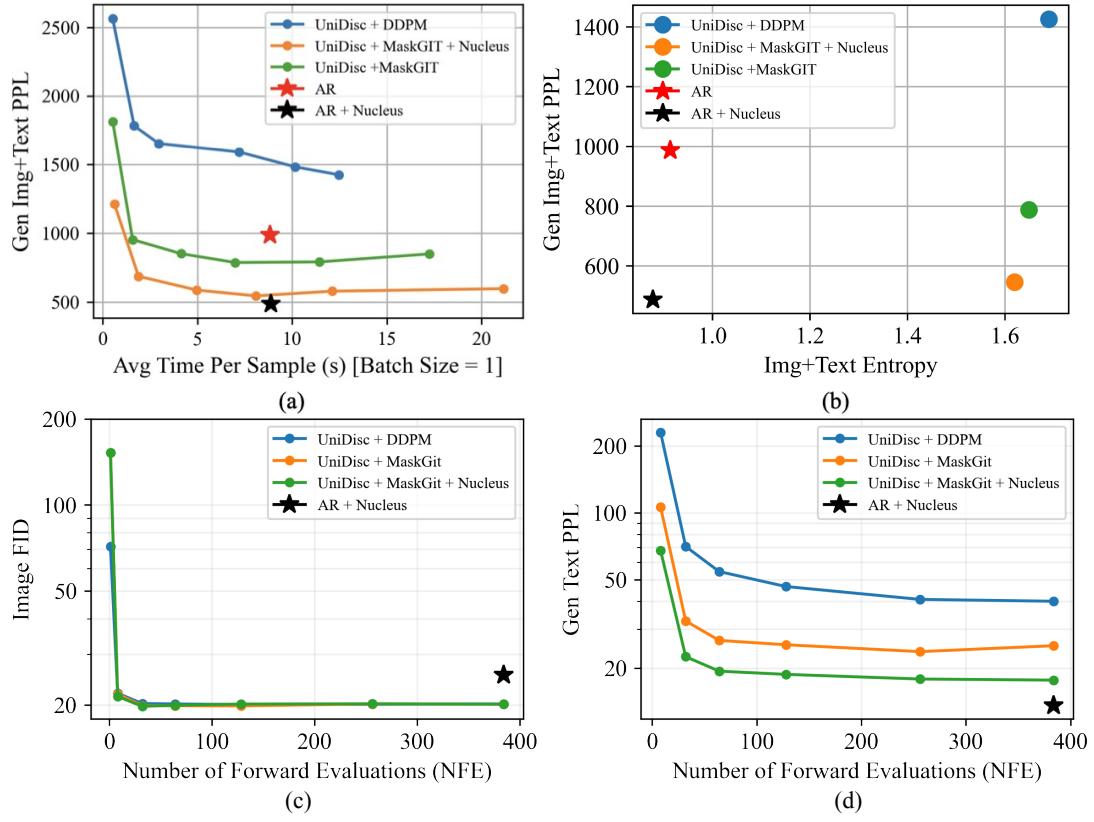We compare the training efficiency of UniDisc and our AR baseline in Figure 3 (right) and

8

Figure 5 | **Inference Comparisons for UniDisc and AR baseline**: (a) Chameleon Perplexity (Text+Image) vs. Time - we perform similar to best AR method, (b) Chameleon Perplexity vs. Entropy - UniDisc has high diversity and low perplexity, while AR has significantly lower diversity, (c) Image FID vs. NFE, showing image generation saturates quickly with NFE (≈ 32), (d) GPT2 Generative Text Perplexity vs. NFE showing text generation benefits from more sampling steps (diminishing).

find that the rough training-inefficiency factor for discrete diffusion to that of AR models for unified training is about 13.2—i.e. one needs to train UniDisc 13.2x longer to achieve the same loss. Additional experimental details are available in Appendix C.

While training efficiency is important, inference efficiency is equally—if not more—important as we deploy these models at wide scale. Thus, we compare the inference efficiency of UniDisc and our AR baseline in Figure 5 (a), (c) and (d). In (a), we measure the joint generative perplexity using Chameleon, In (c) we measure the Image FID and in (d) we measure the Text Perplexity. While it might appear from (a) and (d) that AR does better than UniDisc, in Figure 5 (b), we find that UniDisc has far higher entropy at a given perplexity.

We note that solely looking at the generative perplexity is not sufficient, as it has been previously found Zheng et al. (2024) that very low perplexity can be achieved by repeating the same tokens, which we find often happens with AR w/nucleus sampling and low temperature. We demonstrate such degenerate cases in Appendix B.3. Therefore, Generative Perplexity + Entropy should be considered jointly for evaluating the quality of generation results.

4.3. Multimodal Discriminative Performance

Generative models can act as strong discriminative models as shown in several recent works Jaini et al. (2024); Li et al. (2023); Prabhudesai et al. (2023). Moreover, Rambhatla and Misra (2023) show the discriminative ability of a generative model can be a good metric for its generation performance. In this section, we compare the discriminative capabilities of AR models and UniDisc on cross-model retrieval tasks (image/text/joint).

We evaluate on Winoground Thrush et al. (2022) and a held-out DataComp1B validation set Gadre et al. (2024), using 18M text/image pairs from DataComp1B as our training set. To enable text retrieval during inference for the AR model, we train with flipping the order of modalities, putting the image first 20% of the time, following Zhou et al. (2024). We find that this improves the retrieval for the AR model. All other hyperparameters follow those in Section 4.1. Details of evaluations on CLEVR-VQA and CLEVR-Ref are available in Appendix B.5

Figure 6 | **Uniform Concept Generation**: We perform joint generation given only masked text input (left). We use a language conditioned segmentation model and find that UniDisc generates uniformly in *concept space* (right).

	Clevr-VQA	Clevr-Ret	Datacomp	Winoground
Text Retrieval				
AR	0.60	0.81	0.85	0.24
UniDisc	**0.63**	**0.94**	0.85	**0.31**
Image Retrieval				
AR	N/A	0.06	**0.96**	0.25
UniDisc	N/A	**0.25**	0.95	**0.27**
Joint Retrieval				
AR	N/A	0.06	0.17	0.06
UniDisc	N/A	**0.5**	**0.64**	**0.20**

Figure 7 | **Image-Text Reasoning measured by QA and retrieval accuracy across datasets.**

In Fig. 7, we report the image retrieval, text retrieval and joint retrieval accuracy for AR and UniDisc. For image retrieval, the model is given a text caption paired with 16 images, out of which only one image is correctly paired and the rest are random. The goal is to accurately classify the correct image. To evaluate the model's retrieval accuracy we check if the correct image has the highest $p(x^{img}|x^{txt})$ among all other images. We do the same for text retrieval, where we check $p(x^{txt}|x^{img})$. For joint retrieval, only a single pair has the correct mapping, and every other pair has a random image and text. We check if the correct pair has the highest joint probability $p(x_{img}, x_{txt})$

We find that UniDisc significantly outperforms the AR model on all retrieval tasks. To further investigate this, we measure the joint retrieval accuracy across denoising steps & CFG values in Fig. 13 in Appendix. We find CFG and the number of denoising steps to play a large role in UniDisc's retrieval accuracy. While the number of denoising steps in an AR model is fixed to the sequence length, the denoising steps for UniDisc can be much higher.

10

4.4. Scaling UniDisc

We show that UniDisc scales well across parameters and dataset size. We train a 1.4B parameter model with web-scale data. Our model is trained in two stages, with a low-resolution pre-training stage and a second high-resolution fine-tuning stage. Our first-stage consists of 250M image/caption pairs at 256×256 resolution. We curate our dataset from several sources, with 200M open-web images from Gadre et al. (2024), which were re-captioned by a VLM to create higher-quality descriptions by Li et al. (2024b). We also add a set of smaller datasets consisting of Pixel-Prose Singla et al. (2024), JourneyDB Sun et al. (2023), and Cambrian-10M Tong et al. (2024). In addition, we construct a high-quality, custom dataset of 18M synthetic images, following findings by Sehwag et al. (2024); Zhuo et al. (2024) on the importance of image/caption alignment for image generation. We construct our dataset by prompting an LLM to augment a set of 250K human prompts and use Esser et al. (2024) for generations. In both stages, we account for dataset imbalance and sample more from higher-quality sources. Finally, we fine-tune our model in a second stage, interpolating the RoPE 2D embeddings to train at 512×512 on 30M image/caption pairs.

Impressionist painting of a woman walking down a street in Paris

liberty lady statue, crowded street scene, mosaic style

the 4 willows are swaying really hard over the pond at the beach

a majestic, weeping willow tree in a field, with a lake in the background.

Figure 8 | **Zero-shot Image Editing**: UniDisc can take corrupted and mismatched image/text pairs (left) and produce an aligned, high-quality pair (right), using the model's own likelihood as a scoring function.

Further due to lack of space, we ablate several architecture and objective design choices on a smaller model in Appendix F and we show the training curve of our 1.4B model is available in Fig. 15. We also compare UniDisc to recent multimodal models on standard image generation benchmarks in Appendix H. Qualitative results from the model are available in Appendix G, demonstrating zero-shot text-conditioned image inpainting (Fig. 18), standard text-to-image (Fig. 16), and image-to-text (Fig. 17) generation. Moreover, we demonstrate a form of image editing in Fig. 8 and Appendix G.1, showing that UniDisc can, without any specialized fine-tuning, *automatically* improve a text & image pair by noising and denoising, using the model's likelihood as a judge. Additionally, we analyze the joint generation of UniDisc in Fig. 6 and Appendix G.2 , finding that the model generates images roughly *uniformly in concepts instead of in area*.

5. Conclusion

In this paper, we introduced UniDisc, the first large-scale unified multimodal discrete diffusion model capable of generating, inpainting and editing both images and text. By leveraging discrete diffusion processes, we showed that UniDisc surpasses autoregressive models in both inference efficiency and quality. Our model unifies various design choices in discrete diffusion space, across modalities, through extensive ablations and analysis. We hope that our work inspires future research in this direction.
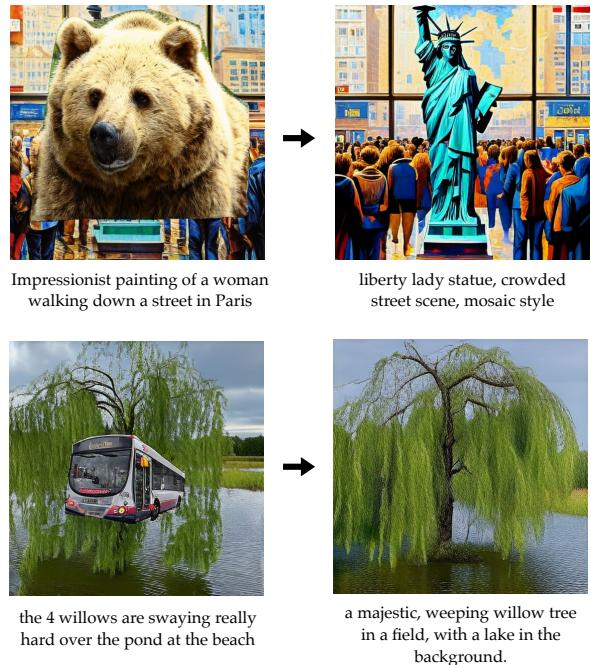
6. Acknowledgment

References

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. v. d. Berg. Structured denoising diffusion models in discrete state-spaces. 2021. URL http://arxiv.org/pdf/2107.03006.

H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. Maskgit: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11315–11325, 2022.

H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan. Muse: Text-to-image generation via masked generative transformers, 2023. URL https://arxiv.org/abs/2301.00704.

S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021. URL https://arxiv.org/abs/2102.08981.

X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. URL https://arxiv.org/abs/1504.00325.

K. T. Chitty-Venkata, S. Raskar, B. Kale, F. Ferdaus, A. Tanikanti, K. Raffenetti, V. Taylor, M. Emani, and V. Vishwanath. Llm-inference-bench: Inference benchmarking of large language models on ai accelerators, 2024. URL https://arxiv.org/abs/2411.00136.

T. Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In The Twelfth International Conference on Learning Representations, Oct. 2023.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, and J. Tang. Cogview: Mastering text-to-image generation via transformers. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 19822–19835, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a4d92e2cd541fca87e4620aba658316d-Abstract.html.

D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. H. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. R. Florence. Palm-e: An embodied multimodal language model. In International Conference on Machine Learning, 2023. URL `https://api.semanticscholar.org/CorpusID:257364842`.

P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first International Conference on Machine Learning, 2024.

S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems, 36, 2024.

I. Gat, T. Remez, N. Shaul, F. Kreuk, R. T. Q. Chen, G. Synnaeve, Y. Adi, and Y. Lipman. Discrete Flow Matching, Nov. 2024.

D. Ghosh, H. Hajishirzi, and L. Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL `https://arxiv.org/abs/2310.11513`.

I. Gulrajani and T. B. Hashimoto. Likelihood-based diffusion language models. Advances in Neural Information Processing Systems, 36, 2024.

T. Hang, S. Gu, C. Li, J. Bao, D. Chen, H. Hu, X. Geng, and B. Guo. Efficient diffusion training via min-snr weighting strategy. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7441–7451, 2023.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

J. Ho and T. Salimans. Classifier-free diffusion guidance. 2022. URL `http://arxiv.org/pdf/2207.12598`.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. URL `https://arxiv.org/abs/2006.11239`.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022. URL `https://arxiv.org/abs/2203.15556`.

E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems, 34:12454–12465. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf`.

M. Hu, C. Zheng, Z. Yang, T.-J. Cham, H. Zheng, C. Wang, D. Tao, and P. N. Suganthan. Unified discrete diffusion for simultaneous vision-language generation. In The Eleventh International Conference on Learning Representations, 2023.

A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795, 2021.

P. Jaini, K. Clark, and R. Geirhos. Intriguing properties of generative classifiers, 2024. URL https://arxiv.org/abs/2309.16779.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. CoRR, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak. Your diffusion model is secretly a zero-shot classifier, 2023. URL https://arxiv.org/abs/2303.16203.

D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024a.

X. Li, H. Tu, M. Hui, Z. Wang, B. Zhao, J. Xiao, S. Ren, J. Mei, Q. Liu, H. Zheng, et al. What if we recaption billions of web images with llama-3? arXiv preprint arXiv:2406.08478, 2024b.

X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto. Diffusion-lm improves controllable text generation, 2022. URL https://arxiv.org/abs/2205.14217.

A. D. Lindström and S. S. Abraham. Clevr-math: A dataset for compositional language, visual, and mathematical reasoning, 2022. URL https://arxiv.org/abs/2208.05358.

D. Liu, S. Zhao, L. Zhuo, W. Lin, Y. Qiao, H. Li, and P. Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657, 2024.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

R. Liu, C. Liu, Y. Bai, and A. Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions, 2019. URL https://arxiv.org/abs/1901.00850.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization, Jan. 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs, math].

A. Lou, C. Meng, and S. Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Forty-first International Conference on Machine Learning.

A. Lou, C. Meng, and S. Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In The Eleventh International Conference on Learning Representations, 2022.

Z. Luo, F. Shi, Y. Ge, Y. Yang, L. Wang, and Y. Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. arXiv preprint arXiv:2409.04410, 2024.

S. Mehta, M. H. Sekhavat, Q. Cao, M. Horton, Y. Jin, C. Sun, I. Mirzadeh, M. Najibi, D. Belenko, P. Zatloukal, et al. Openelm: An efficient language model family with open-source training and inference framework. arXiv preprint arXiv:2404.14619, 2024.

S. Nie, F. Zhu, C. Du, T. Pang, Q. Liu, G. Zeng, M. Lin, and C. Li. Scaling up masked diffusion models on text, 2024. URL https://arxiv.org/abs/2410.18514.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL `https://arxiv.org/abs/1505.04870`.

M. Prabhudesai, T.-W. Ke, A. C. Li, D. Pathak, and K. Fragkiadaki. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback, 2023. URL `https://arxiv.org/abs/2311.16102`.

S. S. Rambhatla and I. Misra. Selfeval: Leveraging the discriminative nature of generative models for evaluation, 2023. URL `https://arxiv.org/abs/2311.10708`.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov. Simple and effective masked diffusion language models. arXiv preprint arXiv:2406.07524, 2024.

V. Sehwag, X. Kong, J. Li, M. Spranger, and L. Lyu. Stretching each dollar: Diffusion training from scratch on a micro-budget. arXiv preprint arXiv:2407.15811, 2024.

J. Shi, K. Han, Z. Wang, A. Doucet, and M. K. Titsias. Simplified and generalized masked diffusion for discrete data. arXiv preprint arXiv:2406.04329, 2024.

V. Singla, K. Yue, S. Paul, R. Shirkavand, M. Jayawardhana, A. Ganjdanesh, H. Huang, A. Bhatele, G. Somepalli, and T. Goldstein. From pixels to prose: A large dataset of dense image captions, 2024. URL `https://arxiv.org/abs/2406.10328`.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. pages 2256–2265. URL `http://proceedings.mlr.press/v37/sohl-dickstein15.pdf`.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2020.

J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL `https://arxiv.org/abs/2104.09864`.

K. Sun, J. Pan, Y. Ge, H. Li, H. Duan, X. Wu, R. Zhang, A. Zhou, Z. Qin, Y. Wang, J. Dai, Y. Qiao, L. Wang, and H. Li. Journeydb: A benchmark for generative image understanding, 2023. URL `https://arxiv.org/abs/2307.00716`.

P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525, 2024.

Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal. Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems, 36, 2024.

C. Team. Chameleon: Mixed-modal early-fusion foundation models. 2024a. URL `http://arxiv.org/pdf/2405.09818`.

C. Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024b. doi: 10.48550/arXiv.2405.09818. URL `https://github.com/facebookresearch/chameleon`.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL `https://arxiv.org/abs/2204.03162`.

S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. arXiv preprint arXiv:2406.16860, 2024.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. Technical Report arXiv:1706.03762, arXiv, Dec. 2017. URL `http://arxiv.org/abs/1706.03762`. arXiv:1706.03762 [cs] type: article.

M. Wortsman, P. J. Liu, L. Xiao, K. Everett, A. Alemi, B. Adlam, J. D. Co-Reyes, I. Gur, A. Kumar, R. Novak, et al. Small-scale proxies for large-scale transformer training instabilities. arXiv preprint arXiv:2309.14322, 2023.

L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737, 2023.

B. Zhang and R. Sennrich. Root mean square layer normalization. Advances in Neural Information Processing Systems, 32, 2019.

K. Zheng, Y. Chen, H. Mao, M.-Y. Liu, J. Zhu, and Q. Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling, 2024.

C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2024. URL `http://arxiv.org/pdf/2408.11039`.

L. Zhuo, R. Du, H. Xiao, Y. Li, D. Liu, R. Huang, W. Liu, L. Zhao, F.-Y. Wang, Z. Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. arXiv preprint arXiv:2406.18583, 2024.

A. Ziv, I. Gat, G. L. Lan, T. Remez, F. Kreuk, A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. Masked Audio Generation using a Single Non-Autoregressive Transformer, Mar. 2024.

A. UniDisc

A.1. UniDisc Training

We describe the detailed algorithm for unified discrete diffusion training on image and text below in Algorithm 1.

Algorithm 1 UniDisc Training

1: **Require:** Training data x
2: **Require:** Noising Schedule α_t i.e., Linear or Cosine
3: **Require:** Unconditional probability p_{uncond}
4: **Initialize:** Model parameters θ
5: **repeat**
6:      $[x_0^{img}, x_0^{txt}] = x_0 \sim p(x, c)$ ▷ Sample image and text data
7:      $t \sim \mathcal{U}(0, 1)$ ▷ Sample random timestep
8:
9:      $x_t^v \sim q(x_t^v \mid x_0^v) = \alpha_t x_0 + (1 - \alpha_t) e_m \quad$ for $v \in \{\text{img}, \text{txt}\}$ ▷ mask all tokens
10:
11:      **With probability** p_{uncond} ▷ For Classifier-Free Guidance:
12:          **If** rand() < 0.5:   ▷ Randomly set one of the modalities to mask tokens
13:              $x_t^{img} \leftarrow m$
14:          **Else**:
15:              $x_t^{txt} \leftarrow m$
16:      $x_0^{pred} = p_\theta([x_t^{img}, x_t^{txt}])$ ▷ Estimate model prediction from masked sequence
17:      Compute loss as: $\mathcal{L}_{\text{diff}} = \frac{\alpha_t'}{1 - \alpha_t} \log \langle x_0^{pred}, x_0 \rangle$ ▷ Loss function over the logits of inputs
18:      Perform gradient step on $\mathcal{L}$ to update θ
19: **until** converged

A.2. Sampling Algorithms

Here we describe the implementations of UniDisc's sampling algorithm and MaskGIT Chang et al. (2022).

Algorithm 2 MaskGIT Sampling

1: **Initialize:** $x_T \leftarrow [m, m, \ldots, m]$ ▷ All tokens are masked
2: **Require:** Sampling steps T
3: **Require:** Num Tokens to Unmask: $f(t)$. We set $f(t)$ as $\frac{1 - \alpha_t}{\sum_{t=1}^T 1 - \alpha_t}$
4: **for** $t = T$ **down to** 1 **do**
5:      $p_{x_0} \leftarrow p_\theta(x_0 \mid x_t)$ ▷ Model prediction
6:      $p_{x_0}^{(p)} \leftarrow \text{Top}_p(p_{x_0})$ ▷ Top-p (Nucleus) sampling on logits
7:      $p_{x_0}^{(k)} \leftarrow \frac{p_{x_0}^{(k)}}{\tau(t)}$ ▷ Apply temperature annealing
8:      Sample $x_{\text{new}} \sim \text{Categorical}(p_{x_0}^{(k)})$ ▷ Sample new tokens
9:      $M \leftarrow \lfloor f(t) \times N \rfloor$ ▷ Determine number of tokens to unmask
10:      Select M most confident tokens based on $p_{x_0}^{(k)}$
11:      Update $x_{t-1}[i] \leftarrow x_{\text{new}}[i] \quad \forall i \in$ selected positions
12:      **Keep** previously unmasked tokens unchanged
13: **end for**

B. Additional Experiment Details

B.1. Conditional and Unconditional Experiment Details

For unconditional and conditional results in Fig. 10 and 11 we use a dataset of 11B tokens comprising 30M images from DataComp1B Gadre et al. (2024) and CC12M Changpinyo et al. (2021) as our training set, with a fraction of 20% text tokens and 80% image tokens after excluding pad tokens. For faster convergence, we train only on DataComp1B for results in Fig. 5 and Fig. 13. We tokenize the image and text tokens using separate tokenizers. We use lookup-free quantization (LFQ) from Luo et al. (2024); Yu et al. (2023) for as our image tokenizer, and use the tokenizer from Touvron et al. (2023) as our text tokenizer. We use an image resolution of 256×256, and a downsampling ratio of 16, resulting in a sequence length of 384 with 256 with image tokens and 128 text tokens. Note that we use the same tokenizers for all the baselines, ensuring fair comparisons. We train UniDisc for 300 L40S GPU hours and train the autoregressive model for a proportionate amount of time such that it achieves the same validation loss. Our model comprises 115M/340M non-embedding parameters and we use a batch size of 512, a learning rate of 3e−4, and weight decay of 0.05, following Sun et al. (2024).

B.2. Conditional and Unconditional Evaluations

We extend Fig. 4, adding results on Flickr-30K and MS-COCO below in Fig. 9. We show unconditional results in Fig. 10 and conditional results (taking the optimal CFG weight for both UniDisc and the AR model) in Fig. 11.

B.3. Generative Perplexity — Qualitative

Text	Chameleon Perplexity	GPT2 Perplexity
"ICLR is globally renowned for presenting..." (Continued)	32.836	35.780
"This is simple. This is simple." (Repeated)	8.423	3.930
"Words Words Words Words" (Repeated)	2.226	3.583
"AAAAAAAAAAAA" (Repeated)	2.732	1.904
"(Spaces Repeated)"	80.240	1.095

Table 1. We demonstrate how generative perplexity is an imperfect metric requiring calibration with entropy.

Figure 9. **Conditional generation results for both FID and CLIP metrics, across a range of CFG values.** We find that AR is more sensitive to the CFG weighting, with a narrower optimal range.

B.4. Quantitative Inpainting Comparison w/autoregressive models

To demonstrate the tradeoff between the pre-training objectives of UniDisc and AR models, we evaluate both models on inpainting. We fine-tune the 340M parameter AR model on a standard set of multimodal datasets (CC12M, Recap-DataComp-1B, LAION 400M) and evaluate UniDisc in a zero shot manner—without any fine-tuning. Specifically, for the AR model, we use a linear masking schedule for the prefix sequence consisting of a randomly masked text and image pair and then predict and supervise the clean sequence, doubling the overall sequence length. In Fig. 12, we evaluate at multiple noise levels, showing the degradation in performance as the original sequence is increasingly masked.

	CC12M	DataComp	Flickr	MS-COCO
Image + Text Perplexity				
Chameleon	541.2	156.8	1254.9	1128.3
UniDisc	494.5	154.8	1115.0	982.2
Image - FID				
Chameleon	**30.5**	**20.49**	**75.70**	**70.67**
UniDisc	35.78	22.97	88.88	77.43
Text - CLIP				
Chameleon	23.70	**26.08**	23.70	23.64
UniDisc	**25.01**	25.98	**24.92**	**25.01**

Figure 10. Unconditional multimodal generation results for UniDisc and AR baseline at 115M parameters - both models perform similarly.

	CC12M	DataComp	Flickr	COCO
Text to Image - FID				
Chameleon 115M w/o CFG	26.32	20.49	46.13	56.46
Chameleon 340M w/o CFG	20.75	18.53	36.24	42.41
Chameleon 115M w/ CFG (0.5)	22.10	16.68	46.06	47.58
Chameleon 340M w/ CFG (0.5)	20.22	13.55	32.74	30.62
UniDisc 115M w/o CFG	27.22	21.26	43.46	54.21
UniDisc 340M w/o CFG	19.28	14.59	34.37	37.73
UniDisc 115M w/ CFG (1.5)	**13.21**	**12.00**	33.79	**31.94**
UniDisc 340M w/ CFG (1.5)	**13.11**	**11.55**	**26.83**	**23.77**
Image to Text - CLIP				
Chameleon 115M w/o CFG	22.08	26.01	22.50	23.02
Chameleon 340M w/o CFG	22.53	26.68	23.51	24.46
Chameleon 115M w/ CFG (0.5)	22.93	27.30	23.38	24.03
Chameleon 340M w/ CFG (0.5)	23.65	27.70	24.95	25.99
UniDisc 115M w/o CFG	21.75	25.98	22.44	22.88
UniDisc 340M w/o CFG	22.18	26.86	23.18	24.44
UniDisc 115M w/ CFG (1.5)	**24.54**	**29.65**	**25.42**	**26.24**
UniDisc 340M w/ CFG (1.5)	**24.77**	**30.01**	**26.63**	**27.82**

Figure 11. Conditional generation results for UniDisc and AR baseline. Our model significantly outperforms the AR model when classifier free guidance is used.

Figure 12. We compare UniDisc with an AR model fine-tuned for joint inpainting and evaluate on a subset of DataComp1B.

B.5. Discriminative Evaluations

For evaluations on CLEVR-VQA and CLEVR-Ref Liu et al. (2019) we use their respective train-val splits. Note that for CLEVR-VQA and CLEVR-Ref, we do not follow the training scaling factor found in Fig. 3, we instead train both the models until convergence, i.e multiple epochs. The small size of these datasets makes it possible to train until convergence. For CLEVR images, we find that none of the existing tokenizers work well, so we fine-tune our own tokenizer on CLEVR images. We use images of 128×128 resolution, with a total sequence length of 320 (256 image tokens and 64 text tokens). For text, we use a standard BERT tokenizer Devlin et al. (2019). In Figure Fig. 13, we ablate the role CFG and the number of denoising steps play in UniDisc's retrieval accuracy. While the number of denoising steps in an AR model is fixed to the sequence length, the denoising steps for UniDisc can be much higher.

Figure 13. **Joint Retrieval Accuracy on DataComp1B.** We outperform AR given the task of retrieving one correct image-text pair out of 16 possible pairs, implying better learnt representations.

C. Scaling Experiment Details

As in prior experiments, all implementation details are shared between UniDiscand AR training configurations except: (1) causal vs. full attention, (2) masking of the input sequence for UniDisc, and (3) weighting the CE loss as in Eq. (3).

We use similar hyperparameters as in our small-scale experiments in Section 4.1, but repeat them here for clarity. For images, we use lookup-free quantization (LFQ) Luo et al. (2024); Yu et al. (2023) with a image resolution of 256×256, and a downsampling ratio of 16, resulting in a sequence length of 256 image tokens. We use a BPE tokenizer Touvron et al. (2023) for text with 128 text tokens, resulting in a total sequence length of 384. We report only non-embedding parameters and data tokenization is identical across all models.

We use a batch size of 512 and use AdamW Loshchilov and Hutter (2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay $\lambda = 0.05$. We use a max learning rate of 3e−4 with a linear warmup followed by a cosine decay schedule, ending with zero at the final step for a given training run.

We list all model variants in Table 2.

Parameters (M)	n_layers	n_heads	d_model
34	11	6	384
67	11	9	576
116	12	12	768
172	20	12	768
228	20	14	896
343	24	16	1024
484	22	10	1280
543	17	12	1536
622	29	10	1280
713	23	12	1536
826	27	12	1536
1074	26	14	1792
1290	24	16	2048

Table 2. Model variants. The FFN hidden size is always 4x the overall d_model

D. Training Details

D.1. Additional Training Implementation Details

We use flash attention for all models except as noted below, using the popular Flash-Attention 2 library Dao (2023). For all AR models at inference, we use K/V caching and take advantage of specially optimized functions for this in FlashAttention 2.

E. Fine-tuning An Autoregressive model for Discrete Diffusion

As we already have a plethora of large-scale AR models Team (2024b); Touvron et al. (2023), it would be useful to have the ability to fine-tune them for a discrete diffusion objective. While the naive method for fine-tuning would be to change the objective function to discrete diffusion while using AR's pre-trained weights. We find that a better idea is to left-shift the output targets of the diffusion objective such that instead of having the masked token predict its respective visible token, we have the token before the masked token predict it. In this way, we more closely match the original AR next-token prediction objective. In Fig. 14 we show that this strategy works well and we can effectively fine-tune a pre-trained autoregressive language model using discrete diffusion loss. We demonstrate this result on a 270M parameter language model Mehta et al. (2024), OpenELM, which is trained with an AR objective. We compare against training from scratch and training AR without the shift. We find the shifting strategy converges faster.

E.1. Large Scaling Training Curve

We show the training curve for the large scale experiments described in Section 4.4 in Fig. 15.

Figure 14. Fine-tuning a pre-trained 270M parameter AR model on LM1B.

Figure 15. Training Loss Curve vs. Tokens on our 1.4B model.

F. Ablations

We validate our design choices by running small-scale experiments on a subset of our primary dataset, taking 18M image/caption pairs on DataComp1B. We train on lower-resolution images at 128×128 and obtain a 1:1 ratio of text to image tokens, with 64 text and 64 image tokens for a total sequence length of 128, with all other hyperparameters the same as in our primary experiments.

We examine the influence of several design choices for our model in Table 3 and reach several conclusions. First, architecture changes to improve training stability—namely adding QK Normalization and using RMSNorm instead of LayerNorm—do not substantially affect convergence in this setting.

Another natural design choice is to parameterize the model such that we provide the modality of a given token to the model. With this relaxation we can drastically reduce the output space and, in theory, simplify the objective for our model. However, we find that this reparametrization only marginally reduces overall perplexity, even at this smaller-scale. We hypothesize that the modality-specific embeddings added to each token allows the model to learn the correct output space with minimal added parameters.

	DataComp1B Validation PPL
UniDisc	93.8
w/o QK Norm	92.7
w/ Zero-linear init	93.8
w/o RMSNorm	93.8
w/o -inf for invalid tokens	94.7
w/o Softmin SNR	109.6
None	111.2

Table 3. Ablation w/115M parameter model of QK Norm, zero initialization of linear layers, RMSNorm, setting invalid tokens to $-\infty$ during training and generation, and Softmin SNR.

	DataComp1B Validation FID
UniDisc	11.4
w/cosine noising schedule	11.5
w/o CE loss weighting	11.35
w/discrete time (T=1000)	13.8

Table 4. Ablation w/115M parameter model on different objective level decisions such as noising schedule, loss weighting and whether to use discrete time.

G. Large Scale Qualitative Results

We show additional results on tasks such as joint inpainting, image captioning and image generation. We note that none of these tasks were explicitly trained or optimized for by our model. This is an intrinsic property due to the nature of UniDisc's unified diffusion based objective. In Fig. 16 we show standard text-to-image generation and in Fig. 17 we show standard image-to-text generation. In Fig. 18 we show zero-shot text-conditioned inpainting, and in Fig. 19 we show zero-shot *multimodal* inpainting.

Figure 16. UniDisc's ability to generate an image, given unseen text as input.

A whimsical penguin in a sports outfit, playing with balls and floating above a city skyline, inspired by the steampunk. Inspired by M.C. Escher.

A bright red ceramic cup filled with a vibrant lemonade next to an old fashioned wooden baby wagon in a sunny backyard setting.

a grove of trees at dusk with scattered branches.

On a quiet summer afternoon, a pair of classic wooden Adir-ack chairs are placed on the porch.

Figure 17. UniDisc's ability to generate text (captioning), given unseen image as input.

a detailed pencil drawing of an anatomy-inspired flower design

a Baroque-style etching of a parrot

Vintage photograph, ginger cat, green eyes, curious expression

young king, 14, newly crowned, ultra-realistic, colorful, godlike presence, close-up

Figure 18. Zero-shot text-conditioned inpainting. UniDisc inpaints a masked region given a user-provided text prompt.

A close-up of a corgi's face

A close-up of a corgi's face with a big smile on red lips, set against a background covered in silly doodles and scribbles.

In a dense forest, a wise old a gnarled tree branch.

In a dense forest, a wise old owl and mongoose sit side by side, both supporting a gnarled tree branch.
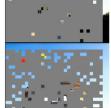
A group of hot above the Colosseum.

A group of hot air balloons float above the Colosseum.

tropical fish

tropical fish swimming in shallow coral reef

Figure 19. Zero-shot multimodal inpainting. UniDisc jointly inpaints in both image and text spaces.

G.1. Zero-shot image editing of UniDisc

A clear benefit of diffusion models is the ability to perform zero-shot editing without specific paired data—which is often difficult to obtain. We demonstrate one such method in Figure 20, showing that UniDisc can automatically improve a user provided image and caption.

We augment real images by overlaying random objects from the COCO dataset. Similarly, we augment captions by asking an LLM to generate purposely incorrect variations. We then randomly mask the image and text inputs and unmask as described above, automatically removing these undesired image artifacts and generating the correct caption. We adopt a best-of-n sampling strategy with n distinct noise masks. We unroll each generation using the model's own likelihood to select the best generation.

Figure 20. Zero-shot multimodal editing. We provide a *clean* image and text pair and UniDisc automatically enhances both the image and text. In the final row, we fix the text and allow only the image to change.

G.2. Analyzing the joint image-text generation of UniDisc

In Figure 21, we visualize how the model iteratively infills both image and text. This raises the question - does UniDisc follow a certain strategy during generation (for example, generating entire background first then moving to subject or generating text first before image), or does it generate everything at once jointly. To analyze this, we take the final model generated image, semantically segment it (using Grounded SAM 2 in our case) and then see which concepts get generated at what timesteps. This is visualized in Figure 22. We find that UniDisc generates

all concepts at once proportional to the overall fraction of the image the concept occupies. We also investigated if the UniDisc has any strong positional bias, such as first generating tokens in the middle and radially filling out. However we find no such positional strategy and that UniDisc is positionally invariant. Intuitively, this means that at any denoising step, all positions are equally likely to be decoded.

Figure 21. We show how UniDisc jointly infills both image and text. argmax $p_\theta(x_0 \mid x_t)$

tropical fish, shallow coral garden, sunlit

A dragonfly zooms out of a window, a koala climbs up a palm tree, and a toucan sits on a branch.
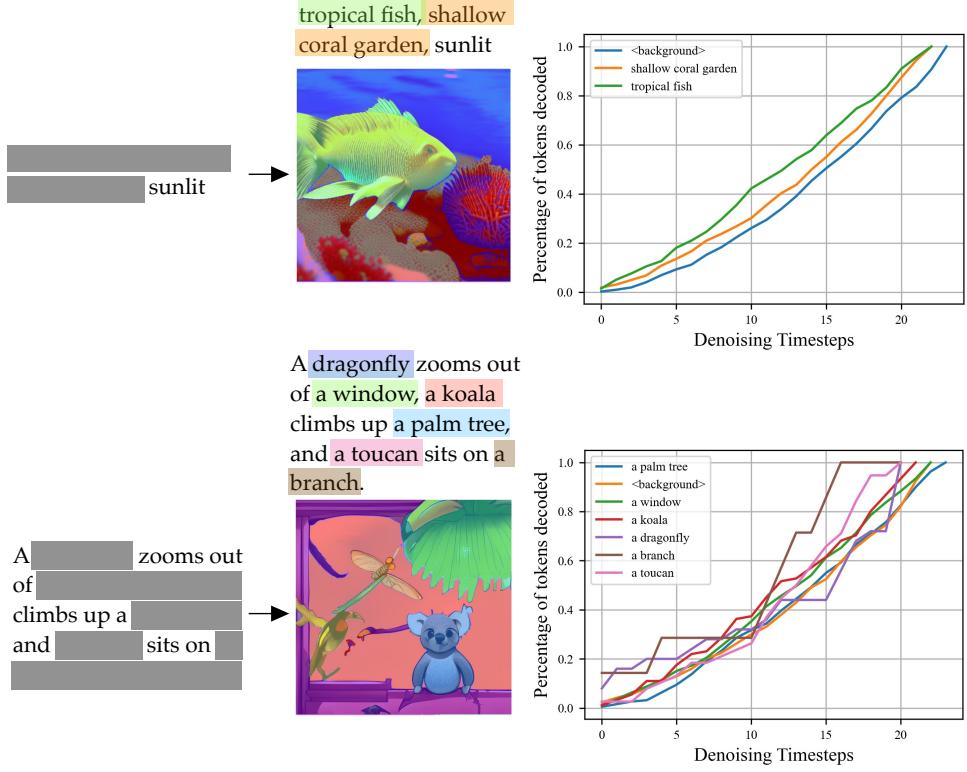
Figure 22. We demonstrate that UniDisc uniformly generates all concepts at once.

G.3. Zero-shot length extrapolation of UniDisc

In this section, we demonstrate the ability of UniDisc to perform zero-shot flexible resolution generation thanks to the use of RoPE embeddings on both text and image tokens. UniDisc model was fine-tuned on 512x512 images—resulting in each image using 1024 tokens—but is able to infill at 1024x1024—resulting in 4096 tokens per image—without further training. We demonstrate this in Fig. 23.

A whimsical, dreamlike painting of a forest, with a destroyed AT-AT in the distance, surrounded by towering mushrooms and glowing fireflies

an eye-catching graphic art poster featuring a majestic winged lion from mythology, surrounded by flames and magic spells

A group of friends hiking together through the misty fog, with a beautiful lake or river in the distance
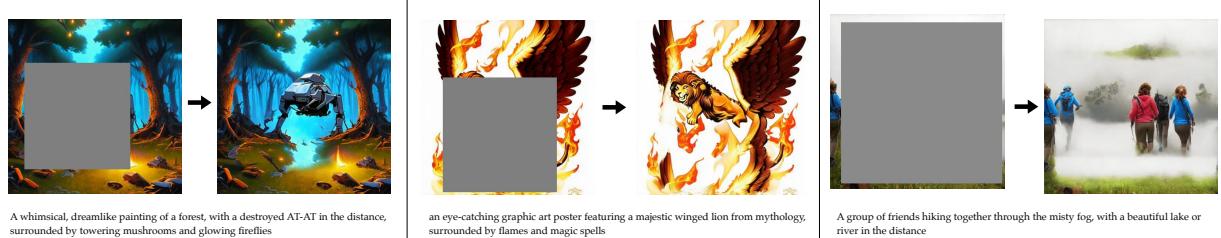
Figure 23. We train UniDisc on 512x512 resolution images but demonstrate zero-shot inpainting at 1024x1024.

H. Large Scale Quantitative Comparisons

H.1. Quantitative Generation Comparison with recent mulitmodal models

In Table 5, we evaluate UniDisc on the popular GenEval Ghosh et al. (2023) benchmark which looks at how well a generated image adheres to the prompt in terms of a set of predefined attributes (e.g., color, positioning). In Table 6, we compare FID on the popular MS-COCO 30K Chen et al. (2015) dataset on MJHQ-30K Li et al. (2024a), which contains a higher proportion of highly-aesthetic images.

We also compare to the reported results from UniD3 Hu et al. (2023), which most closely resembles our work.

Method	Sing. Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
SDv1.5 Rombach et al. (2022)	0.97	0.38	0.35	0.76	0.04	0.06	0.43
CoDI Tang et al. (2024)	0.89	0.16	0.16	0.65	0.02	0.01	0.31
Lumina-mGPT Liu et al. (2024)	-	-	-	-	-	-	0.32
UniDisc	0.92	0.47	0.15	0.67	0.13	0.19	0.42

Table 5. We evaluate UniDisc on the GenEval Ghosh et al. (2023) benchmark.

Method	MSCOCO-30K FID ↓	MJHQ-30K FID ↓
SDv1.5 Rombach et al. (2022)	11.12	-
CoDi Tang et al. (2024)	**22.26**	19.87
UniD3 Hu et al. (2023)[2]	25.11	-
UniDisc (Ours)	23.86	**18.67**

Table 6. We evaluate the 1.4B version of UniDisc on FID. We use evaluate on MS-COCO 30K Chen et al. (2015) and MJHQ-30K Li et al. (2024a).

Method	Params	CUB200 FID ↓
UniD3 Hu et al. (2023)	637M	17.38
UniDisc (Ours)	330M	**11.03**

Table 7. We compare our model to Hu et al. (2023) on CUB200.

---

[2]Trained only on MS-COCO. Other works listed in this table trained on a broader set of datasets (possibly including MS-COCO). In most cases, training on additional datasets likely harms dataset-specific FID.

I. Understanding the effect of Classifier Free Guidance (CFG)

In Table 11, we observe that CFG is a significant factor in the performance difference between UniDisc and the AR baseline. We hypothesize that this is because CFG is most useful in decoding the first few tokens, with diminishing utility in later tokens. To examine this, we look at intermediate predictions by storing $\arg\max p_\theta(x_0 \mid x_t)$ at each sampling step. As an AR model cannot directly capture this distribution without an intractable rollout, we opt to use the same UniDisc model but with an autoregressive inference strategy, decoding from left to right. This allows us to directly compare the performance of different inference strategies and how they interact with classifier-free guidance.

We visualize this in Figure 24, where we visualize the difference between the conditional and unconditional image generated at different percentages of decoded tokens. We notice two things: (a) the difference diminishes as more tokens are decoded and (b) UniDisc consistently has higher distances between the logits than AR, which flattens out more quickly.
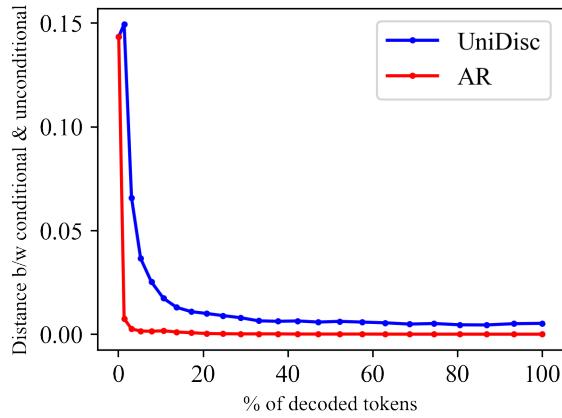

Figure 24. L2 distance between unconditional and conditional logits on currently masked tokens as sampling steps increase.

Steps	CLIP Score
[1 − 3]	0.301
[12 − 14]	0.293
[22 − 24]	0.283
All (24)	0.312

Figure 25. Comparing CLIP scores by applying CFG only on specific steps. This shows CFG has the most impact on the initial denoising steps (total steps = 24).

Intuitively, this means UniDisc extracts much more discriminating signal from CFG compared to AR. We believe this is because UniDisc has much more flexibility to decode tokens initially based on confidence, compared to AR which is forced to decode in a left to right manner and thus, can course correct quickly and more effectively. This can be seen in Table 25, where we selectively apply CFG only on a few steps and notice that CLIP score when CFG is applied on steps 1-3 almost matches applying CFG on all, while applying on the last few steps doesn't affect things much at all.

Given the differences in CFG between UniDisc and AR models, we conduct a hyperparameter sweep over guidance scales in Figure Fig. 4. We compute FID and CLIP scores over four datasets, and at both 115/340M parameters. We find that our AR baseline benefits from a weight of $w = 0.5$ but sees far less improvement than UniDisc with CFG. For UniDisc, we choose an overall weight of $w = 1.5$, but note that the CLIP score scales cleanly with the guidance scale, demonstrating the trade-off between visual quality and prompt adherence.

Finally, in Fig. 26, we show the effect of CFG on the generated image. We increase the weight of the classifier-free guidance from $w = 0$ to $w = 8$ and observe the effect on the generated image.

Figure 26. We show the effect of classifier-free guidance from left-to-right, starting with $w = 0$, and increasing linearly to $w = 8$ on the right, where output logits are

$$l_{\text{cfg}} = (1 + w)l_{\text{cond}} + w * l_{\text{uncond}}.$$

Caption: "crab meditating, surfboard, orange sun setting, rainbow clouds, zen beach"

J. Inference: Generation time vs. batch size

We analyze the quality of the generation versus time in Figure 27. We make a similar observation as in prior work Gat et al. (2024); Ziv et al. (2024) on discrete diffusion, finding that the ability to obtain predictions with varying sampling steps allows lower latencies. However, with current implementations, KV caching in AR models results in higher throughput as the batch size increases. This tradeoff can be explained by looking at the number of function evaluations (NFEs) and the cost of each in both cases. In AR generation w/KV caching, we have a fixed NFE, but each forward pass is substantially less expensive than in the NAR case. In contrast, in NAR, we can use substantially fewer NFEs, but each is more costly. Modern GPUs only reach peak throughput at larger batch sizes Chitty-Venkata et al. (2024); as we decrease the batch size, the difference in computation per function evaluation diminishes, resulting in NAR having favorable performance.
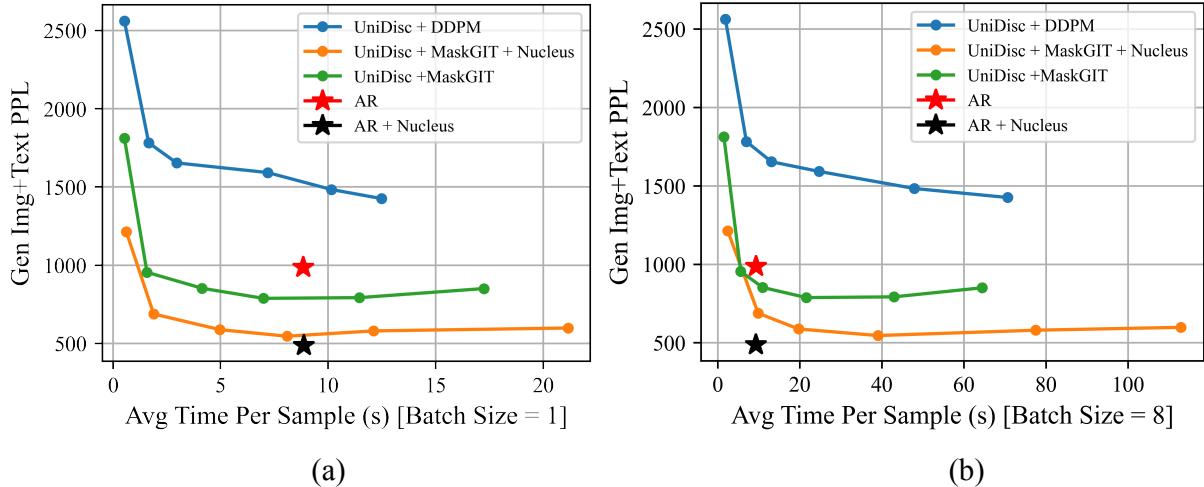

(a)

(b)

Figure 27. Generative Perplexity vs. Time with various models and sampling strategies.