GlueGen: Plug and Play Multi-modal Encoders for X-to-image Generation

Can Qin[*][*], Ning Yu[†], Chen Xing[†], Shu Zhang[†], Zeyuan Chen[†],

Stefano Ermon[‡], Yun Fu[*], Caiming Xiong[†], Ran Xu[†]

[*]Northeastern University, [†]Salesforce AI Research, [‡]Stanford University
qin.ca@northeastern.edu, ermon@cs.stanford.edu, yunfu@ece.neu.edu,
{ning.yu, cxing, shu.zhang, zeyuan.chen, cxiong, ran.xu}@salesforce.com

Abstract

Text-to-image (T2I) models based on diffusion processes have achieved remarkable success in controllable image generation using user-provided captions. However, the tight coupling between the current text encoder and image decoder in T2I models makes it challenging to replace or upgrade. Such changes often require massive fine-tuning or even training from scratch with the prohibitive expense. To address this problem, we propose GlueGen, which applies a newly proposed GlueNet model to align features from single-modal or multi-modal encoders with the latent space of an existing T2I model. The approach introduces a new training objective that leverages parallel corpora to align the representation spaces of different encoders. Empirical results show that GlueNet can be trained efficiently and enables various capabilities beyond previous state-of-the-art models: 1) multilingual language models such as XLM-Roberta can be aligned with existing T2I models, allowing for the generation of high-quality images from captions beyond English; 2) GlueNet can align multi-modal encoders such as AudioCLIP with the Stable Diffusion model, enabling sound-to-image generation; 3) it can also upgrade the current text encoder of the latent diffusion model for challenging case generation. By the alignment of various feature representations, the GlueNet allows for flexible and efficient integration of new functionality into existing T2I models and sheds light on X-to-image (X2I) generation.[1]

1. Introduction

Text-to-image (T2I) generative models have made great progress in the last few years thanks to algorithmic advances
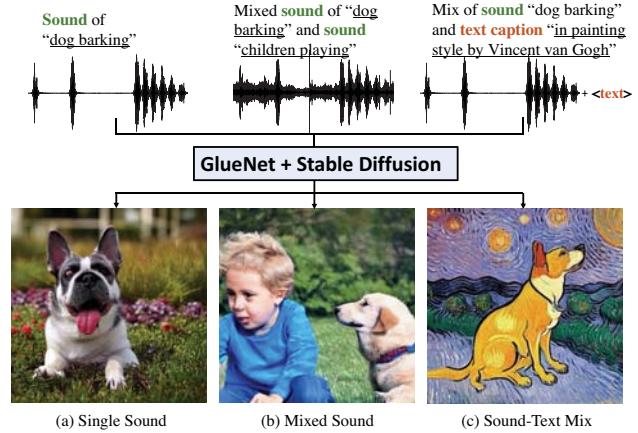

Figure 1. Setting of GlueGen. GlueNet is trying to provide an adaptable portal for the Stable Diffusion model to input multi-modal data, such as text, audio, *i.e.*, (a) and (b), or text-audio hybrid signals, *i.e.* (c), for X-to-image generation.

and the availability of large-scale paired training datasets [42, 70, 45, 50, 49]. Diffusion-based T2I generative models in particular have achieved remarkable results in terms of image quality [24, 36, 3, 47, 35, 73]. Despite these strong results, controllable generation for these methods is still challenging: generated images are often not faithful to the captions, compositional capabilities are lacking, and prompt engineering is often required to achieve the desired results [37]. Moreover, most large-scale models have only been trained on English text captions, greatly limiting their use across the world.

Recent research has emphasized the crucial role of text encoders in improving Text-to-Image (T2I) models' performance, and their ability to comprehend and represent text is considered a bottleneck for image generation [47, 10]. However, the current T2I models' text encoders are often trained on short image captions, which limits their performance on complex prompts and challenges their qual-

ity of feature extraction [45]. Furthermore, T2I models' capacities are limited to generating images from text, and they cannot incorporate multimodal conditions such as sound and audio easily. Nevertheless, replacing the text encoder in existing T2I models is challenging since the text encoder and image generator's representation spaces are tightly coupled [45, 42]. This severe domain gap between the new conditions and the existing model impedes the image generation's final performance, and training the entire T2I model from scratch, with higher quality image-caption pairs, would be prohibitively expensive [14].[2]

As seen in Fig. 1, we propose GlueNet to address the challenge of efficiently replacing or upgrading the text encoder in existing diffusion-based T2I models. With GlueNet, off-the-shelf pre-trained language models and multimodal encoders can be easily aligned with image encoders of T2I models, greatly enlarging their functionalities at a low cost. Importantly, this can be achieved without requiring retraining from scratch or even finetuning, maintaining the representation alignment between the text and image encoders. The proposed method follows an encoder-decoder structure. The encoder of GlueNet first aligns the representation space of the new condition encoder with that of the T2I model's image generator, minimizing both element-wise and global-wise discrepancy. Then, the decoder of GlueNet maps the aligned condition representations back to the original representation space of the new condition encoder by minimizing the reconstruction loss, preserving rich semantics captured by the pre-trained model during alignment training. Align existing models would inevitably decrease feature discriminability [7, 11], which makes the feature decoder necessary. The entire training of GlueNet requires only a parallel corpus with the same content but different modalities or languages. At inference time, only the encoder of GlueNet is applied on top of the new condition encoder for representation alignment.

To verify the effectiveness of the proposed framework, we conducted three major experiments ranging from single- and multi-modal encoders. Firstly, we upgraded the existing text encoders of the Latent Diffusion Model [45] using a stronger language model, T5-3B [41]. Our model showed competitive improvements in FID score and user study ranking compared to the baselines but it still required finetuning for the overall performance boost. Secondly, we aligned a multilingual language model, XLM-Roberta-L [9], using our approach, enabling multilingual text-to-image generation. It achieved competitive results of translation-based models under a significantly lower training cost. Finally, we demonstrated GlueNet's capability to bring new functionalities beyond text signals into existing T2I models. The alignment of the AudioClip [20] encoder enables sound-to-image generation without requiring any

---

[2]The cost of training a Stable Diffusion model is around 600K USD.

parameter finetuning of the image generator. This new capability allows the existing Stable Diffusion model to generate high-quality images that correspond to sound signals such as dogs barking and street music. This new capability goes beyond the traditional T2I generation and opens up new possibilities for creating multimedia content towards X-to-Image (X2I) generation.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to consider the problem of efficiently aligning a pre-trained audio model with a pre-trained T2I diffusion model for sound-to-image generation.

- Extensive experiments on text-to-image generation benchmarks demonstrate the superiority of our model over the baseline LDM method on both image quality and language controllability.

- Our framework also enables text-to-image generation beyond English prompts without the need of multilingual image-text pairs for retraining.

2. Related Works

2.1. Text-to-Image Models

Generative Adversarial Nets (GANs) [18] is one of the major frameworks in text-to-image generation. The method in [44] was an early stage approach to use a GAN to train a text-to-image generation network. Since then, other GAN based methods, *e.g.*, Stack-GAN [72], Attn-GAN [66] and SD-GAN [69], have obtained promising results. In addition, recent works showed more improvements on the generation quality. DM-GAN [75] improved text-to-image performance by introducing a dynamic memory component. DF-GAN [57] designed a fusion module to fuse text and image features. LAFITE [74] took advantage of CLIP's model to construct pseudo image-text pairs, and proposed a GAN model to learn text-image pairs.

Auto-regressive transformer is another major framework in text-to-image generation. DALL-E [43] and CogView [13] adopted an autoregressive transformer [63] to train the correspondences between visual tokens and text tokens. Parti [70] used a powerful visual encoder, ViT-VQGAN [71], to improve results upon DALL-E. The method in [16] is similar to CogView and DALL-E, while introducing additional controlling elements to improve the tokenization process. NÜWA [65] and NUWA-infinity [64] trained autoregressive visual synthesis model to support both text-to-image and text-to-video generation tasks.

Concurrently, diffusion models [54, 55, 56] became a main research focus. In such methods, noise is added to an image, and a score network is trained to denoise and recover the input image. GLIDE [36] performed guided
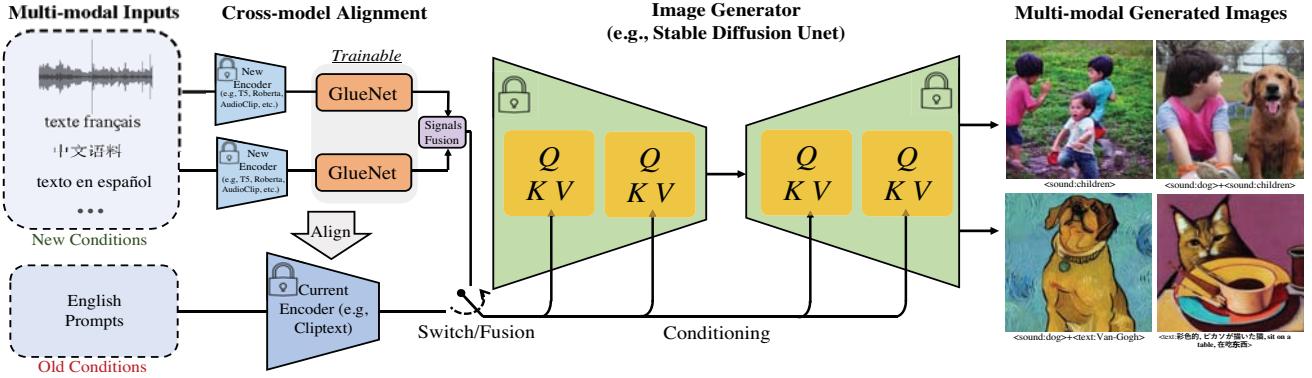
Figure 2. Illustration of our desired GlueGen framework. With the proposed GlueNet model of the GlueGen framework, the pre-trained image generator (*i.e.* UNet) can be bridged to off-the-shelf single- or multi-modal encoders to expand their functionalities, *i.e.*, multilingual/sound-to-image generation, within a limited budget. GlueNet is trained offline and does not require back-propagation of UNet and image-text pairs for training. Therefore, GlueGen is flexible and efficient to achieve.

inference with and without a classifier network to generate high-fidelity images. DALL-E 2 [42] and Imagen [47] set new state-of-the-art results in the text-to-image generation. In DALL-E 2, a prior to produce CLIP image embeddings conditioned on text was learned, and a diffusion model was used to decode the image embeddings to an image. In Imagen, a large-scale frozen text encoder T5-XXL [41] was adopted to generate embeddings, and a cascade of conditional diffusion models was used to map these embeddings to images of increasing resolutions. Latent diffusion model and stable diffusion model [45] are state-of-the-art methods that apply the diffusion model on the latent embedding space as in [53, 61]. These methods are computationally friendly while achieving impressive results. To incorporate more conditions for controllable content generation, [73, 35, 4] introduced additional parameters with new data to inject external knowledge within the current pre-trained model. [8] tried to alter the language model for more capacities. However, all these models are designed to adapt text and image (or similar 2d grid) conditions without the attempt for sequential signals such as audio.

2.2. Sound-to-Image Generation

Sound and image are two distinct modalities. [26] attempts to address sound-guided semantic image manipulation by aligning sound features with a GAN model. CoDI [6] take a two-stage strategy which pre-trains individual diffusion model with aligned prompt encoder in the beginning. Then it aligns latent to project multimodal features into a shared space. AAI [68] introduces sound-guided image generation, editing and stylization into one model.

3. Preliminary

3.1. Latent Diffusion Model

The Latent Diffusion Model (LDM) [45] and its extension, *i.e.*, Stable Diffusion[3], which are variants of Denoising Diffusion Probabilistic Model (DDPM) [23] family, are selected as our baseline models due to their excellent balance in efficiency and visual quality. In general, LDM is composed of two stages. Firstly, there is an auto-encoder (AE) pre-trained by enormous images with the regularization in either KL-divergence or vector quantization [62, 1]. An encoder network, *i.e.*, E, is applied for latent feature extraction as $z = E(x)$, which can be mapped back to image space by a decoder network. The input images x and reconstructed images $\hat{x}$ are almost identical $x \approx \hat{x}$.

Secondly, LDM trains a diffusion model in the latent space [53, 61]. It follows the standard DDPM [23] with denoising loss and uses U-net [46] as the image decoder as in [55]. To enable generative controllability, LDM has applied multiple conditioning signals (y) such as text, mask, or layout, encoded aside and injected into the U-net, with the help of cross-attention layers. This can be formulated as:

$$\mathcal{L}_{LDM} := \mathbb{E}_{z \sim E(x), \varepsilon \sim N(0,1), t, y} \left[\| \varepsilon - \varepsilon_\theta(z_t, t, c_\phi(y)) \| \right],$$
(1)

where t represents the time step, z_t is the noise corrupted latent tensor at time step t, and $z_0 = E(x)$. ε is the unscaled Gaussian noise, c_ϕ is the conditioning network parameterized by ϕ and ε_θ is the Unet-like denoising network (*a.k.a*, image decoder). The parameters of both conditioning and denoising networks, *i.e.*, θ, ϕ, are trained by the LDM loss. During inference, clean images can be gener-

---

[3]https://github.com/CompVis/stable-diffusion

ated via classifier-free guidance [24] as:

$$\hat{\varepsilon}_\theta(z_t|y) = \varepsilon_\theta(z_t) + s \cdot (\varepsilon_\theta(z_t, c_\phi(y)) - \varepsilon_\theta(z_t)), \quad (2)$$

where s is the guidance weight to balance text controllability and image fidelity.

The basic T2I LDM model is trained on Laion-400M [50] dataset. The Stable Diffusion Model (SDM) was trained with more epochs and training data, *i.e.*, Laion2B-en and Laion-high-resolution [49].

3.2. Condition Encoder

LDM uses Bert-like [12] encoder as the text encoder jointly trained with the image decoder using the DDPM loss. SDM uses a pre-trained CLIP [40] text encoder frozen during training. It has been found that increasing the size of the text encoder is very helpful for performance [47]. In order to upgrade existing models, our goal is to be able to efficiently plug in more advanced language condition encoders such as T5-3B [41], AudioClip [20], or XLM-Roberta [12].

4. GlueNet

The text encoder is one of the key components of T2I models, as generation requires precise and fine-grained text embedding for guidance. The overall performance can be greatly boosted by increasing the size of the text encoder [47], however, upgrading an existing text model to a more powerful one is challenging. Because existing T2I models are not modular, directly replacing (or augmenting with) another condition encoder does not work. The key technical challenge is the mis-alignment between the new condition encoder and the old image decoder. Moreover, joint finetuning also falls short due to catastrophic forgetting occuring when updating well trained parameters.

Considering these two different condition encoders as source and target domains, we apply a neural network to learn to align. As shown in Fig. 2, this paper has presented a general framework called GlueGen. Within the framework, GlueNet works as an additional module to bridge the new language model and the old image decoder.

4.1. Objectives

Given sets of parallel corpus $\mathcal{X}_{src} = \{x_{ij}^s | i = 1, ..., L^s, j = 1, ..., N^s\}$ and $\mathcal{X}_{tar} = \{x_{ij}^t | i = 1, ..., L^t, j = 1, ..., N^t\}$ with L as length of tokens and N as total quantity, the source model F^s and a target encoder F^t have mapped the raw data into embeddings denoted as $\mathcal{S} = \{s_{ij} | i = 1, ..., L^s, j = 1, ..., N\}$ where $s_{ij} = F^s(x_{ij}^s) \in \mathbb{R}^{d_s}$ and $\mathcal{T} = \{t_{ij} | i = 1, .., L^t, j = 1, ..., N\}$ where $t_{ij} = F^t(x_{ij}^t) \in \mathbb{R}^{d_t}$. Due to the different models, there is severe distribution mismatch between the source and target features $p(s) \neq p(t)$. The GlueNet is an autoencoder whose encoder $M(\cdot, \Theta_M)$ learns to map the new features from a

source language model, *i.e.*, $s \in \mathcal{S}$, to align with the current target encoder, *i.e.*, $t \in \mathcal{T}$, where $p(M(s, \Theta_M)) \approx p(t)$. Therefore, this translation (*i.e.*, $F^s \xrightarrow{\Theta_M} F^t$) enables the image generator to understand the new features coming from the new condition model without finetuning. To achieve this aim, we consider it as a domain adaptation [17, 39, 31] problem and apply both element-wise and distribution-wise alignment losses, which includes the minimization of the mean square error (MSE) loss, *i.e.*, $\mathcal{L}_{mse}$, and the adversarial loss, *i.e.*, $\mathcal{L}_{adv}$, measured over a discriminator network, *i.e.*, $D(\cdot, \Theta_D)$:

$$\mathcal{L}_{mse}(\Theta_M) := \mathbb{E}_{x_t \sim X_t, x_s \sim X_s}[||F^t(x_t) -$$
$$M(F^s(x_s), \Theta_M)||_2^2], \quad (3)$$
$$\mathcal{L}_{adv}(\Theta_D, \Theta_M) := \mathbb{E}_{x_t \sim X_t}[\log D(F^t(x_t), \Theta_D)]$$
$$+ \mathbb{E}_{x_s \sim X_s}[\log 1 - D(M(F^s(x_s), \Theta_M), \Theta_D)], \quad (4)$$

where X_s and X_t denote the source and target inputs respectively and they can be the same prompts in English (*e.g.*, LDM → T5-3B [41]) or in bilingual but parallel content for multilingual T2I or audio-label pairs.

Some insightful investigations in transfer learning reveal that the cross-domain alignment will inevitably decrease feature discrimination [7, 11]. To project the rich semantics necessary for model upgrading, we further apply a decoder network, *i.e.*, $N(\cdot, \Theta_N)$, for feature reconstruction:

$$\mathcal{L}_{rec}(\Theta_M, \Theta_N) :=$$
$$\mathbb{E}_{x_s \sim X_s}[||x_s - N(M(F^s(x_s), \Theta_M), \Theta_N)||_2^2]. \quad (5)$$

GlueNet is trained in an end-to-end manner and the details of which can be referred to in Appendix and Sec. 5. [34] introduced that a universal learning engine should seek a compressive closed-loop transcription with good property in structure-preserving. Our cross-model mapping also requires a similar functionality. GlueNet desires stability after a loop whose input and output should keep almost equivalent, *i.e.*, $x \approx \hat{x} \approx \hat{\hat{x}} \approx ...$, due to the constraint of reconstruction loss. Therefore, inputting the reconstructed data $\hat{x}$ to GlueNet is expected to be consistent with the previous loop. This is helpful for accelerating GlueNet training for T5-to-LDM adaptation.

4.2. Model Architecture

The input data to alignment can be represented as the sequence of tokens: $X_{i,*} = \{x_{i,1}, x_{i,2}, ..., x_{i,L}\}$ where $X_{i,*} \in \mathcal{X}$ and $x_{i,j} \in \mathbb{R}^C$, with sequence length L and token dim C. Inspired by the methods of image super-resolution [28, 2] to densely regress target data under different input-output dimensions, the translator network stacks three sub-nets in sequence, as seen in Fig. 3 (c). The head net is only used for simple feature transformation. The body
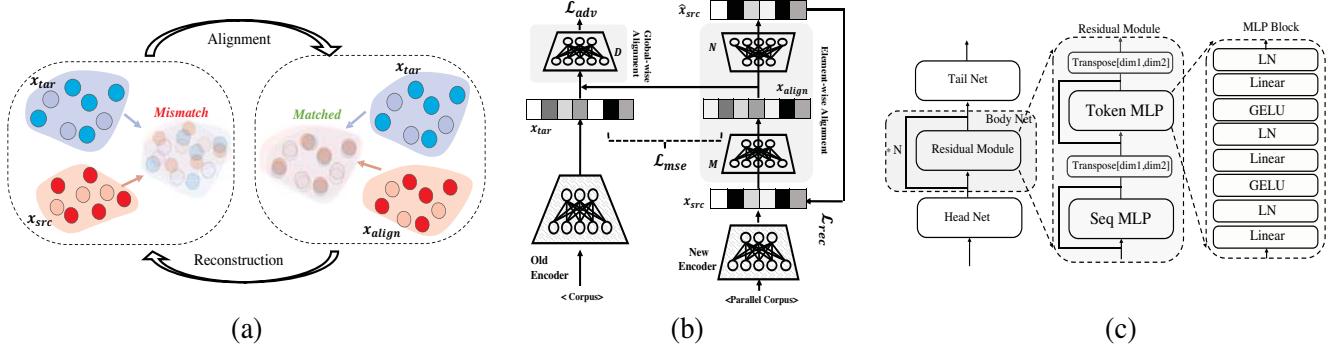
Figure 3. (a) Illustration of features transformation throughout the model translation/alignment. (b) The general pipeline and learning objectives of our proposed GlueNet. (c) Detailed architecture of GlueNet Encoder/Decoder.
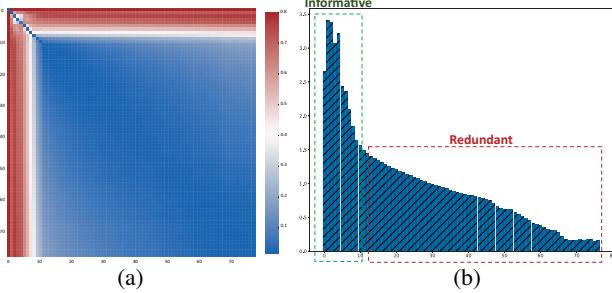

Figure 4. Signal analysis of CLIPText encoder on 77 tokens (used for Stable Diffusion) averaging over 10,000 randomly sampled prompts. (a) Average dis-similarity map (1-cosine) of CLIPText tokens. (b) The distance between each token and the last token, i.e., $dist(s_{\cdot,j}, s_{\cdot,L})$ with L tokens. We can conclude that only the top K tokens are informative.

net has many residual modules for fine-grained representation learning. The tail network is appended and is mainly employed for dimension conversion to match the target tensors. Every residual module is implemented based on the MLP-based token mixer [60, 33] with high efficiency and strong ability in representation learning. Within the MLP-mixer block, it consists of a Token Mixer and a Sequence Mixer for orthometric purposes. The Token Mixer learns the representation of each token with shared MLP, and the Sequence Mixer is designed for learning the same channel in sequential tokens:

$$U_{\cdot,i} = X_{\cdot,i} + W_2\sigma(W_1 * LN_{\gamma_1,\beta_1}(X)_{\cdot,i}), i = 1, ..., C \quad (6)$$
$$X_{j,\cdot} = X_{j,\cdot} + W_4\sigma(W_3 * LN_{\gamma_2,\beta_2}(U)_{j,\cdot}), j = 1, ..., L \quad (7)$$

where $X_{\cdot,\cdot}$ is the feature, $LN_{\gamma,\beta}$ denotes layer normalization parameterized by γ and β, and σ indicates the non-linear operator such as GELU [21]. The $U_{\cdot,i}$ is the outputs of Seq MLP and W represents the weight matrix within the MLP mixer.

4.3. Cross-modality Alignment

GlueNet provides a way to align the cross-modal representations between the text encoder, CLIPText [40], with

the AudioClip [20] model that maps sound signals into embedding spaces. This alignment is useful for extending the functionality of diffusion models like Stable Diffusion, which only takes text signals as input. However, achieving this alignment is challenging because the length of the tokens in each model is different. CLIPText's embeddings require 77 tokens in a sequence, while AudioClip maps sound signals to a single token with a sequence length of 1. Using the standard GlueNet directly fails in this case.

During our exploration, we found that condition information is not uniformly distributed across all 77 tokens of CLIPText outputs. The top K tokens contain the majority of the information. We computed an average dissimilarity map over the CLIPText tokens, visualized in Fig. 4 (a), and observed a prominent gap around the eighth or ninth token, with the later ones being highly similar. We also found that corrupting the last token embedding with some random noise would not degrade the generative results. Therefore, it can be concluded that the contribution of different tokens decreases progressively and could be estimated by the feature distance with the last token, $w_{\cdot,j} = dist(s_{\cdot,j}, s_{\cdot,M})$, as shown in Fig. 4 (b). We take the l2 distance as $dist(\cdot, \cdot)$ and use this value to reweight the GlueNet objective over different tokens for the sound-to-image generation.

Cross-modal signals fusion is another contribution. To do this, we select the top K tokens of each modality output from GlueNet. Then, the average of the remaining later tokens (excluding last K tokens) of two modality features is concatenated with the top ones. This is a non-parametric operation without any training. K is empirically chosen from 4 to 8 (or higher). For more details, please refer to Appendix.

5. Experiments

5.1. Experiments Setup

Upon the shoulders of giants, this paper applies the Latent-Diffusion Model (LDM) and Stable Diffusion Model (SDM) [45] as the baselines with a similar model but dif-

| Chinese: 下午的花园的印象派绘画 | French: Peinture impressionniste d'un jardin d'après-midi | Spanish: Pintura impresionista de un jardín de tarde | Japanese: 午後の庭の印象派絵画 | Italian: Pittura impressionista di un giardino pomeridiano |

Figure 5. Multilingual generation results in resolution 512×512 of XLM-Roberta + Glue-Net + SDM decoder (sd-v1-4) with the same caption, *i.e.*, "afternoon garden oil painting painted by impressionists". With the help of different Glue-Nets and multilingual text encoder, the SDM decoder can support different languages including Japanese, Italian, Chinese, French and Spanish. The guidance weight is assigned as 7.5 and PLMS [30] sampling steps are 50.

ferent text encoders and scalabilities. LDM uses a Bert-like [12] model as the text encoder that is jointly updated with the image decoder based on DDPM loss, while SDM uses a frozen pre-trained CLIP text encoder during training. The LDM is designed for the images in 256×256 and SDM targets 512×512 images. We implement the multilingual T2I and sound-to-image generation upon the SDM.

Implementation. The GlueNet is implemented as the Fig. 3 with an encoder M and a decoder N in symmetrical architecture consisting $\sim$51M parameters. For simplicity in denotation, we refer GlueNet encoder as GlueNet in

Figure 6. Percentage of user votes for text controllability with the prompts from Drawbench [47] and Winoground [59] for T5-to-LDM.

the following experiments, whose body net consists of 5 residual modules as default. We take the AdamW [32] as the optimizer based on PyTorch [38]. The learning rate is assigned as 1×10^{-4} for GlueNet training and 1×10^{-6} for LDM-T5-GlueNet finetuning. The GlueNet is trained on Nvidia-A100, requiring 0.5 to 5 GPU days for different uses.

Condition Encoders. We apply the T5-3B [41] to upgrade the text encoder of LDM Bert. T5-3B has 6x more parameters than LDM Bert and is trained by an enormous text corpus beyond the image captions. Furthermore, to enable multilingual T2I generation, we choose the XLM-Roberta-L [12] as the new condition encoder. In sound-to-image generation, we take the AudioCLIP [20] audio encoder.

Datasets. We have collected the text data for GlueNet training. In monolingual T2I, only English texts are required, and we extract 18M captions from Laion-400M as training corpus, and more data will be more helpful.
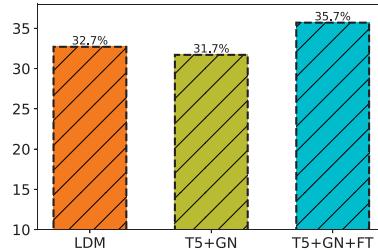
Multilingual GlueNets require parallel texts in different languages. These data are collected by Wikimarix [51] and Googletrans with $\sim$2M sentences for each. To train GlueNet for sound-to-image generation, we use the Urban-sound8k dataset [48][4], which consists of more than 8,000 audio samples collected from city environments in ten different classes, such as dog barking, children playing, car horn, siren, etc.

5.2. Monolingual Text-to-Image Generation

Monolingual T2I generation focuses on the comparison with LDM given the English text prompts. We have decomposed our pipeline into two stages: alignment and finetuning. GlueNet helps to align the new language model whose potential can be better exploited by further finetuning of image text pairs. This section verifies the benefits of our vanilla and finetuned models in both quantitative and qualitative evaluations. **T5+GlueNet** refers to the LDM model with the stacking of T5 and GlueNet as the text encoder where the LDM image decoder is unchanged. **T5+FT** represents finetuning of the LDM image decoder with DDPM loss based on text conditions from a fixed T5. **T5+GlueNet+FT** is the final version, including both GlueNet and joint finetuning with the image decoder. The finetuning takes $\sim$ 100 GPU days of 116M image-text pairs filtered from Laion-400M by BLIP score [27]. Unet finetuning is optional and not needed for the remaining experiments in Sec. 5.3 and Sec. 5.4.

5.2.1 Quantitative Evaluation

Tab. 1 reported the zero-shot FID score [22] on COCO dataset [29] based on the pyorch-fid [52] package. Our re-implementation score of LDM is slightly higher than the paper's one [45], which may because of different implementations. The table shows the minor drop of T5+GlueNet from LDM since our current GlueNet model still leaves some in-

correct alignments, which could be fixed by either joint fine-tuning or collecting more text corpus to train GlueNet. With the ~100 GPU-days finetuning of a subset of the Laion dataset (*i.e.*, 116M image-text pairs), the overall performance has been considerably improved, even outperforming the LDM by certain margins.

We have also conducted a user study to compare the text controllability in Fig. 6 by Amazon Turk [5]. There are ~1K prompts collected from Drawbench [47] and Winoground [59], which comprehensively cover various scenarios and numerous difficult cases.

Table 1. FID on COCO. * or ** indicate our implemented results with 100 or 200 GPU days. ZS means zero-shot.

Method	FID↓	ZS
CogView [13]	27.10	✗
LAFTTE [74]	26.94	✗
GLIDE [36]	12.24	✗
Make-A-Secne [16]	**11.84**	✗
LDM [45]	12.63	✓
LDM*	13.55	✓
T5+FT*	23.30	✓
T5+FT**	12.41	✓
T5+GlueNet	14.32	✓
T5+GlueNet+FT*	12.05	✓

We report the percentages of users' votes for the three methods in Fig. 6. In the user study of controllability, ours outperformed the LDM by 3%, which is a non-trivial improvement. Moreover, the p-value is computed as 0.04 with three repeats of user study that is statistically significant according to the criteria of <0.05.

5.2.2 Visual Comparison

In Fig. 8, we have exhibited some generated images of different methods. By comparing LDM and T5+GlueNet, we can conclude that the proposed GlueNet helps to achieve comparable high-quality images even replacing conditions. Moreover, the T5+GlueNet+FT shows stronger controllability in many cases compared with LDM, such as "a cube made of brick", whereas the LDM's results fail to display the concept of "cube".

5.3. Multilingual Text-to-Image Generation

Our proposed framework is very general and generic to many language models, including the multilingual backbones. We choose the SDM as our base model with the pretrained clip-vit-large-patch14[5] as the text encoder to be upgraded by the multilingual one, *i.e.*, XLM-Roberta-L. The GlueNet helps to align such two backbones with the parallel corpus as training data. Thus, the old SDM image decoder can successfully understand the text embeddings of other languages encoded by GlueNet and XLM-Roberta-L without any additional parameters updating except GlueNet.

---
[5]https://huggingface.co/openai/clip-vit-large-patch14

Table 2. Comparison with AudioClip model for sound-to-image generation. The CLIP↑ score over generated results on Urban-Sound8K benchmark is reported here. We have evaluated three different settings including 1) single sound, 2) sound-sound mix and 3) sound-text mix.

	Sound	Mixed Sounds	Sound-Text
AudioCLIP+SDM	16.07	13.25	17.22
GlueNet+SDM	**20.85**	**20.50**	**22.02**

Table 3. Comparison with translation model, *i.e.*, M2M100-418M [15], for multilingual generation. The Multilingual-CLIP↑ score over Crossmodal [58] benchmark is reported here. GlueNetR denotes the re-weighted objective as described in Sec. 4.3.

	M2M100 [15]	GlueNet	GlueNetR
Chinese	**24.50**	22.01	23.17
French	**25.08**	23.09	23.91
Spanish	23.83	22.91	**24.18**
Japanese	23.73	23.99	**24.23**
Italian	21.08	22.40	**22.88**
Average	23.64	22.88	**23.67**

5.3.1 Comparison

Fig. 5 shows the example of multilingual generation results with the prompt "afternoon garden oil painting painted by impressionists" in many languages. Most of the results reveal the content of "afternoon garden". The "impressionists" is a challenging notion that rarely appears in our training data. Thus, more high-quality parallel data will be greatly helpful. Compared with those finetuned over multilingual image text pairs, our solution is far more efficient and cheaper. For quantitative evaluation, we have compared it with a popular multilingual translation model M2M100-418M [15] (418M #params) trained by 7.5B parallel sentences. In contrast, our GlueNet uses far fewer training data (2M sentences for each pair). The results are listed in Tab. 3 where we reported the Multilingual-CLIP score calculated over the 100 randomly selected multilingual image-caption pairs from Crossmodal dataset [58].The re-weighted version achieved competitive results of translation-based models under a significantly lower training cost.

5.3.2 Hybrid Multilingual Generation

The hybrid multilingual generation is more challenging that both the original SDM and third-party translation toolbox cannot address. Our proposed GlueNet has surprising capability in this problem. According to the results in Fig. 7, the multilingual backbone bonded with GlueNet gives the precise guide to the SDM decoder even with the prompts in three different languages and randomly mixed.

"彩色的, ピカソが描いた猫, sit on a table, 在吃东西"

Stable Diffusion   XLM-Roberta+GlueNet

(a)

"一辆白色的, セダン, crash into a building"

Stable Diffusion   XLM-Roberta+GlueNet

(b)

Figure 7. Hybrid multilingual generation in resolution of 512×512. There are three-different-language texts in the input caption including Chinese, Japanese and English. The caption of (a) is "colorful, a cat painted by Picasso, sit on a table, is eating food" and the caption of (b) is "a white, sedan, crash into a building". With our GlueNet infused ahead, the XLM-Roberta can guide SDM decoder to generate reasonable results where the original SDM fails to work.

Figure 8. Monolingual generation in resolution 256×256 with guidance weight 7.5 and DDIM steps 200. The details of each method can be referred to in Sec. 5.1.

5.4. Sound-to-Image

Beyond the text signals, the proposed GlueNet also achieves sound-to-image generation by aligning the CLIP text encoder with AudioCLIP [20] audio encoders. This is a challenging task due to the modalities gap and the mismatch of token lengths (77 for CLIP and 1 for AudioCLIP). To address these challenges, we apply a reweighted objective with a high focus on the top and informative signals, as described in Sec. 4.3. The input for GlueNet is the pairing data $< sound, label >$ encoded by AudioCLIP-audio-encoder (ESResNeXt [19]) and the CLIP-text-encoder respectively. Then, GlueNet is trained to align the sound embedding with the text embedding according to the objectives described in Sec. 4. To evaluate the effectiveness of GlueNet, we conducted both quantitative and visual comparisons over the testing set of Urbansound8k. Our baseline for comparison is the deployment of AudioCLIP audio and text models to retrieve the labels with audio input. The retrieved labels are then fed into the Stable Diffusion to generate results. The

CLIP score in Tab. 2 strongly reveals the superiority of our GlueNet over the baseline, particularly over mixed sounds, which are more challenging than single-source sounds and can be difficult for humans to distinguish sometimes. We also introduce cross-modal signals fusion as another contribution of this paper. Our analysis of the sound-text fusion in Tab.2 and Fig. 10 demonstrate that our proposed GlueNet achieves the best results.

5.5. Alignment Analysis

To understand how GlueNet maps the cross-model features, we have visualized its trajectory with the generative results in different iterations in Fig. 9. In the beginning, the old image decoder could not understand the new text embeddings at all. Then, certain patterns such as "pencil drawing" appear after 300 iterations of training with a significant decrease in losses. Such a gap can finally be sealed at the end where both $\mathcal{L}_{mse}$ and $\mathcal{L}_{rec}$ converged to small values.

5.6. Comparison with PEFT methods

GlueNet belongs to adapter family but it is designed for offline alignment. Popular methods like ControlNet [73] or T2I-adapter [35] require new condition-image pairs to train the adapter models. Our comparison is with Lora [25], one of the most representative Parameters-Efficient-Finetuning (PEFT) methods. We tested Lora+T5 and GlueNet+T5 for image generation on LDM (following the setting of Tab. 1, +FT

Method	FID ↓
Lora	16.72
GlueNet	**14.32**
Lora+FT	14.80
GlueNet+FT	**13.19**

indicates 10-gpu-day fine-tuning). We trained Lora+T5 to align with LDM-bert and then fine-tuned Lora+T5 with LDM-Unet. Evidently, our GlueNet outperforms both Lora and Lora+FT in quantitative terms.
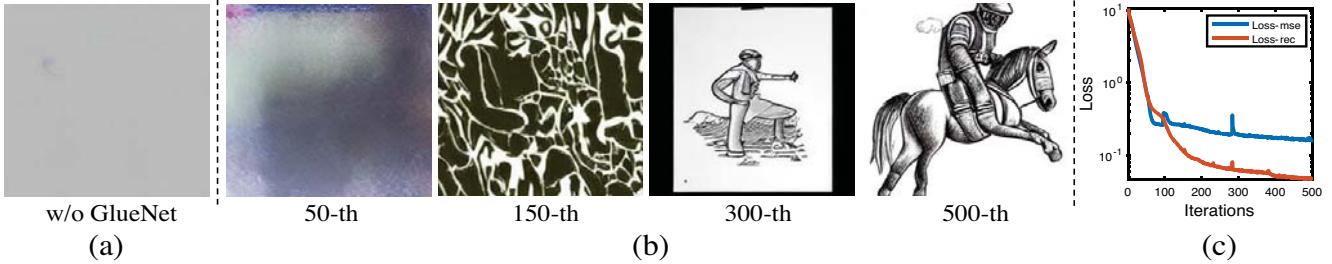
Figure 9. Alignment analysis. (a): Initial result of direct replacement of a new text encoder, *i.e.*, LDM-Bert $\rightarrow$ T5-3B. (b): Results of GlueNet in different iterations with the new text backbone. The prompt is "an astronaut riding a horse as a pencil drawing". (c): Curve of $\mathcal{L}_{mse}$ and $\mathcal{L}_{rec}$ in training.
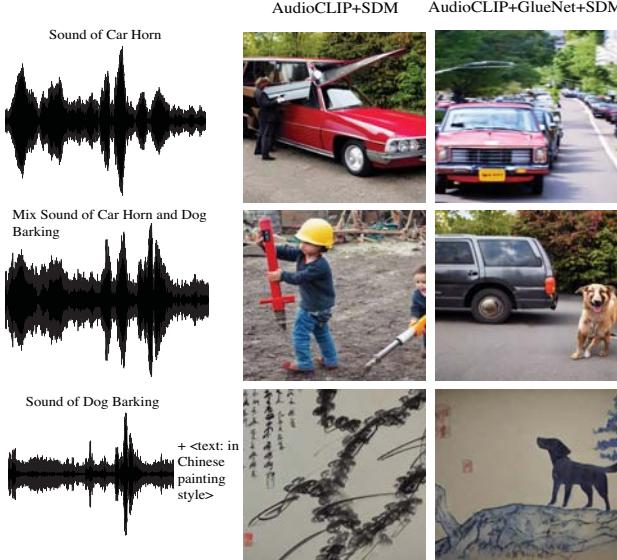

Figure 10. Sound-to-Image generation results.
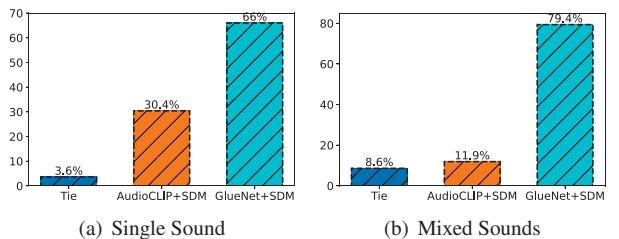

(a) Single Sound

(b) Mixed Sounds

Figure 11. Percentage of user votes of best results in sound-to-image generation with (a) single-source sound and (b) mixed-sources sound as inputs.

5.7. Transformer-based GlueNet

We have implemented a transformer-based GlueNet (GlueNet-Trans) in a multilingual setting following Tab. 2. Our results indicate that GlueNet-MLP marginally outperforms

Language	GNMLP	GNTrans
French	**23.91**	23.87
Spanish	24.18	**24.55**
Italian	**22.88**	21.72
Avg	**23.66**	23.38

GlueNet-Trans. We intend to further investigate the self-attention models in this context.

5.8. Point-Cloud-to-Image

To extend the scope of this work, we have incorporated the 3D-to-image generation. We applied the point-cloud-based pre-training model, *i.e.*, ULIP [67], as the 3D condition encoder and aligned it with CLIPText using our proposed GlueNet. As a result, we have enabled point-cloud-based image generation by substituting the ULIP-Point encoder. We also compared it with the retrieval-based baseline, following the settings of Tab. 2 (where sound is replaced with point cloud objects), and our method demonstrated a significant improvement.

Method	CLIP $\uparrow$
ULIP	17.29
GlueNet	**21.67**

6. Conclusion

Infusing the pre-trained conditional encoder into the existing T2I image generator is an exciting direction toward a more powerful AI system. However, the current encoder cannot be easily upgraded due to the tight match. This paper attempts to break such a strong constraint of the corresponding image-text models towards flexible modularization and efficient upgrading. To address the severe misalignment, we have proposed the GlueNet with both objectives for cross-model alignment and originality protection. Empirically, it is beneficial for the overall performance and enables versatile functionalities for X-to-image generation within a limited budget. We wish this work to be inspirable to the community in large-scale AI system design.

References

[1] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks. *arXiv preprint arXiv:1704.00648*, 3, 2017.

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[5] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 2011.

[6] Ziyi Chang, George A Koulieris, and Hubert PH Shum. On the design fundamentals of diffusion models: A survey. *arXiv preprint arXiv:2306.04542*, 2023.

[7] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 2019.

[8] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022.

[11] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *CVPR*, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 2021.

[14] Benj Edwards. With stable diffusion, you may never believe what you see online again, Sep 2022.

[15] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *JMLR*, 2021.

[16] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014.

[19] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In *IJCNN*, 2021.

[20] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2022.

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[26] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *CVPR*, 2022.

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[30] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[31] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 2018.

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[33] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *ICLR*, 2021.

[34] Yi Ma, Doris Tsao, and Heung-Yeung Shum. On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, 2022.

[35] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[37] Guy Parsons. Dall-e 2 prompt book. `https://dallery.gallery/the-dalle-2-prompt-book/`, 2022.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019.

[39] Can Qin, Haoxuan You, Lichen Wang, C-C Jay Kuo, and Yun Fu. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. *NeurIPS*, 2019.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

[44] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[48] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *ACMMM*, 2014.

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[50] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[51] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*, 2019.

[52] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, August 2020. Version 0.2.1.

[53] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021.

[54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019.

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[57] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*, 2022.

[58] Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*, 2022.

[59] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

[60] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 2021.

[61] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *NeurIPS*, 34:11287–11302, 2021.

[62] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, and Laukasz Kaiser. Attention is all you need. *NeurIPS*, 2017.

[64] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*, 2022.

[65] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜwa: Visual synthesis pretraining for neural visual world creation. In *ECCV*, 2022.

[66] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

[67] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, 2023.

[68] Yue Yang, Kaipeng Zhang, Yuying Ge, Wenqi Shao, Zeyue Xue, Yu Qiao, and Ping Luo. Align, adapt and inject: Sound-guided unified image generation. *arXiv preprint arXiv:2306.11504*, 2023.

[69] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *CVPR*, 2019.

[70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[71] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Baldridge Jason, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *ICLR*, 2022.

[72] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[73] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[74] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.

[75] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dmgan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019.