ECLIPSE: A Resource-Efficient Text-to-Image Prior for Image Generations

Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, Yezhou Yang
Arizona State University
{maitreya.patel, kch, scheng53, chitta, yz.yang}@asu.edu

Abstract

Text-to-image (T2I) diffusion models, notably the unCLIP models (e.g., DALL-E-2), achieve state-of-the-art (SOTA) performance on various compositional T2I benchmarks, at the cost of significant computational resources. The unCLIP stack comprises T2I prior and diffusion image decoder. The T2I prior model alone adds a billion parameters compared to the Latent Diffusion Models, which increases the computational and high-quality data requirements. We introduce ECLIPSE[1]*, a novel contrastive learning method that is both parameter and data-efficient. ECLIPSE leverages pre-trained vision-language models (e.g., CLIP) to distill the knowledge into the prior model. We demonstrate that the ECLIPSE trained prior, with only 3.3% of the parameters and trained on a mere 2.8% of the data, surpasses the baseline T2I priors with an average of 71.6% preference score under resource-limited setting. It also attains performance on par with SOTA big models, achieving an average of 63.36% preference score in terms of the ability to follow the text compositions. Extensive experiments on two unCLIP diffusion image decoders, Karlo and Kandinsky, affirm that ECLIPSE priors consistently deliver high performance while significantly reducing resource dependency. Project page: https://eclipse-t2i.vercel.app/*

1. Introduction

Diffusion models [13, 36, 38, 43] have demonstrated remarkable success in generating high-quality images conditioned on text prompts. This Text-to-Image (T2I) generation paradigm has been effectively applied to various downstream tasks such as subject/segmentation/depth-driven image generation [4, 6, 10, 21, 30]. Central to these advancements are two predominant text-conditioned diffusion models: Latent Diffusion Models (LDM) [38], also known as Stable Diffusion, and unCLIP models [36]. The LDM, notable for its open-source availability, has gained widespread popularity within the research community. On the other hand, unCLIP models have remained under-studied. Both model types fundamentally focus on training the diffusion models conditioned on text prompts. The LDM contains a singular text-to-image diffusion model, while unCLIP models have a text-to-image prior, and a diffusion image decoder. Both model families work within the vector quantized latent space of the image [44]. In this paper, we focus on unCLIP models because they consistently outperform other SOTA models in various composition benchmarks such as T2I-CompBench [14] and HRS-Benchmark [2].

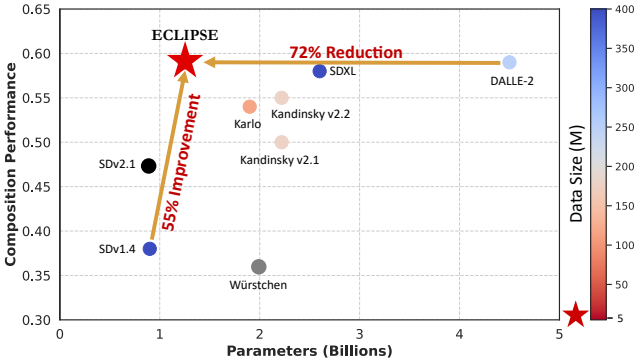


Figure 1. Comparison between SOTA text-to-image models with respect to their total number of parameters and the average performance on the three composition tasks (color, shape, and texture). *ECLIPSE* achieves better results with less number of parameters without requiring a large amount of training data. The shown *ECLIPSE* trains a T2I prior model (having only 33M parameters) using only 5M image-text pairs with Kandinsky decoder.

These T2I models, typically large in parameter count, require massive amounts of high-quality image-text pairs for training. unCLIP models like DALL-E-2 [36], Karlo [8], and Kandinsky [37], feature prior module containing approximately 1 billion parameters, resulting in a significant increase in overall model size ($\geq$ 2B) compared to LDMs. These unCLIP models are trained on 250M, 115M, and 177M image-text pairs, respectively. Therefore, two critical

[1]Our strategy, *ECLIPSE*, draws an analogy from the way a smaller prior model, akin to a celestial entity, offers a glimpse of the grandeur within the larger pre-trained vision-language model, mirroring how an eclipse reveals the vastness of the cosmos.

questions remain: *1) Does the incorporation of a text-to-image prior contribute to SOTA performance on text compositions? 2) Or is scaling up model size the key factor?* In this study, we aim to deepen the understanding of T2I priors and propose substantial enhancements to existing formulations by improving parameter and data efficiency.

As proposed by Ramesh et al. [36], T2I priors are also diffusion models, which are designed to directly estimate the noiseless image embedding at any timestep of the diffusion process. We perform an empirical study to analyze this prior diffusion process. We find that this diffusion process has a negligible impact on generating accurate images and having the diffusion process slightly hurts the performance. Moreover, diffusion models require substantial GPU hours for training due to the slower convergence. Therefore, in this work, we use the non-diffusion model as an alternative. While this approach may reduce the compositional capabilities due to the absence of classifier-free guidance [12], it significantly enhances parameter efficiency and decreases the dependencies on the data.

To overcome the above limitations, in this work, we introduce *ECLIPSE*, a novel contrastive learning strategy to improve the T2I non-diffusion prior. We improve upon the traditional method of maximizing the Evidence Lower Bound (ELBO) for generating the image embedding from the given text embedding. We propose to utilize the semantic alignment (between the text and image) property of the pre-trained vision-language models to supervise the prior training. Utilizing *ECLIPSE*, we train compact (97% smaller) non-diffusion prior models (having 33 million parameters) using a very small portion of the image-text pairs (0.34% - 8.69%). We train *ECLIPSE* priors for two unCLIP diffusion image decoder variants (Karlo and Kandinsky). The *ECLIPSE*-trained priors significantly surpass baseline prior learning strategies and rival the performance of 1 billion parameter counterparts. Our results indicate a promising direction for T2I generative models, achieving better compositionality without relying on extensive parameters or data. As illustrated in Fig. 1, by simply improving the T2I prior for unCLIP families, their overall parameter and data requirements drastically reduce and achieve the SOTA performance against similar parameter models.

Contributions. 1) We introduce *ECLIPSE*, the first attempt to employ contrastive learning for text-to-image priors in the unCLIP framework. 2) Through extensive experimentation, we demonstrate *ECLIPSE*'s superiority over baseline priors in resource-constrained environments. 3) Remarkably, *ECLIPSE* priors achieve comparable performance to larger models using only 2.8% of the training data and 3.3% of the model parameters. 4) We analyze and offer empirical insights on the shortcomings of T2I diffusion priors.

2. Related Works

Text-to-Image Generative Models. Advancements in vector quantization and diffusion modeling have notably enhanced text-to-image generation capabilities. Notable works like DALL-E [35] have leveraged transformer models trained on quantized latent spaces. Contemporary state-of-the-art models, including GLIDE [27], Latent Diffusion Model (LDM) [38], DALL-E-2 [36], and Imagen [39], have significantly improved over earlier approaches like StackGAN [48] and TReCS [20]. As these models achieve remarkable photorealism, several works focus on making T2I models more secure [9, 17, 18, 28]. LDM models primarily focus on unified text-to-image diffusion models that incorporate the cross-attention layers [38]. Additionally, several studies aim at refining Stable Diffusion models during inference through targeted post-processing strategies [4, 6, 33]. In contrast, unCLIP models, exemplified by DALL-E-2 [16], Karlo [8], and Kandinsky [37], incorporate a two-step process of text-to-image diffusion transformer prior model and diffusion image decoder having the same model architecture as LDMs. Recent benchmarks have highlighted the superior compositional capabilities of DALL-E-2 over LDM methods [2, 14]. Our work examines and enhances existing prior learning strategies in open-source pre-trained unCLIP models, Karlo and Kandinsky.

Efficient Text-to-Image Models. The current generation of T2I models is characterized by extensive parameter sizes and demanding training requirements, often necessitating thousands of GPU days. Research efforts have primarily centered on model refinement through knowledge distillation, step distillation, and architectural optimization [22, 26, 40]. Würstchen [32] presents an efficient unCLIP stack requiring less training time. Concurrently, Pixart-α [5] leverages pre-trained Diffusion-Transformers (DiT) [31] as base diffusion models, further reducing training time. Distinctively, *ECLIPSE* focuses on refining text-to-image priors within the unCLIP framework using a mere 3.3% of the original model parameters, thereby significantly reducing the training duration to approximately 50 GPU hours. Our work falls orthogonal to the existing efficient T2I methodologies that mainly focus on knowledge and step distillation, and/or architectural compression. When integrated with these model compression strategies, *ECLIPSE* can position the unCLIP family models as a compact yet highly accurate and efficient methodology.

Contrastive Learning in Generative Models. Contrastive learning, traditionally applied in visual discriminative tasks, has seen utilization in image-text alignment models like CLIP [34], LiT [46], and SigLIP [47]. However, its application in generative models, particularly in Generative Adversarial Networks (GANs), remains limited [7, 23, 49]. For instance, Lafite [49] employs a contrastive approach

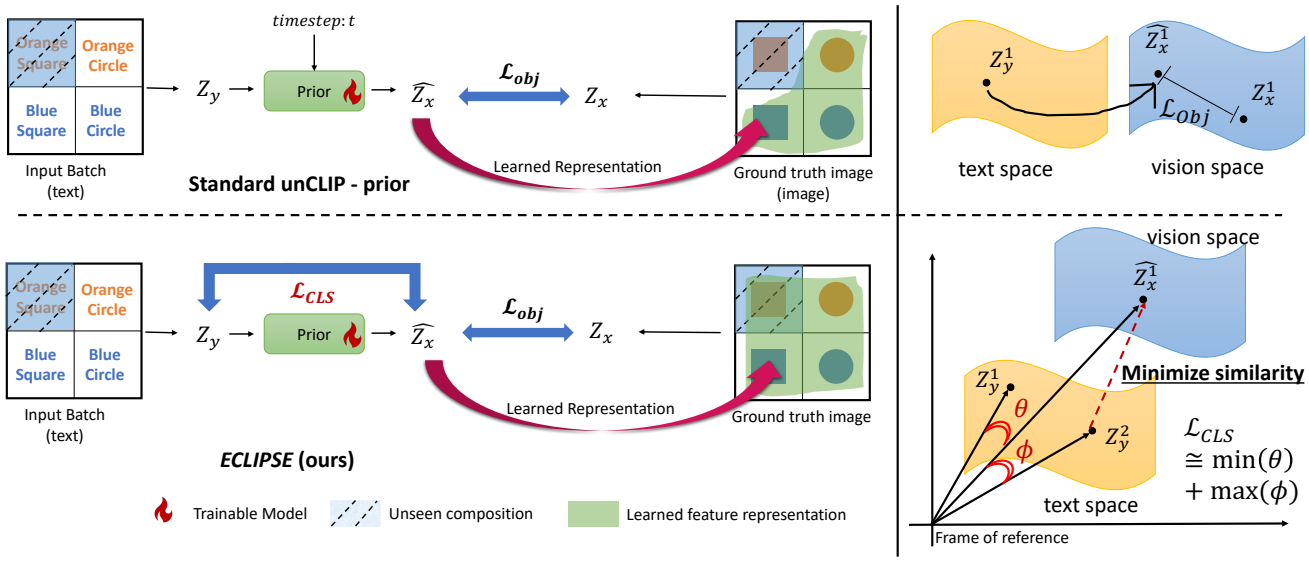


Figure 2. Standard T2I prior learning strategies (top) minimizes the mean squared error between the predicted vision embedding $\hat{z}_x$ w.r.t. the ground truth embedding z_x with or without time-conditioning. This methodology cannot be generalized very well to the outside training distribution (such as Orange Square). The proposed *ECLIPSE* training methodology (bottom) utilizes the semantic alignment property between z_x and z_y with the use of contrastive learning, which improves the text-to-image prior generalization.

for image-to-text prior training in language-free T2I GANs. StyleT2I [23] attempts to learn the latent edit direction for StyleGAN [15], which is supervised via spatial masks on the images making the method not scalable. ACTIG [7] introduces an attribute-centric contrastive loss to enhance discriminator performance. These methods are constrained by their domain-specific knowledge requirements and inability to be directly applied to diffusion models [7, 23]. In contrast, *ECLIPSE* applies CLIP-based contrastive learning to train more effective T2I prior models in diffusion-based T2I systems. This strategy is not only resource-efficient but significantly enhances the traditional text-to-image diffusion priors by exploiting the semantic latent space of pre-trained vision-language models.

3. Methodology

This section elaborates on the Text-to-Image (T2I) methodologies, beginning with an overview of unCLIP, followed by the formal problem statement. We then delve into our proposed training strategy, *ECLIPSE*, for T2I prior in detail. Figure 2 provides the overview of baselines and *ECLIPSE* training strategies.

3.1. Preliminaries

Without the loss of generality, let's assume that $y \in Y$ denotes the raw text and $x \in X$ denotes the raw image. z_x and z_y denote the image and text latent embeddings extracted using the pre-trained vision and text encoders ($z_x = C_{vision}(x)$; $z_y = C_{text}(y)$). Ideally, these C_{text} and C_{vision} can be any model (e.g., T5-XXL, ViT, and CLIP). Both model families (LDM and unCLIP) fundamentally focus on learning a mapping function $f_\theta : Y \rightarrow X$. The LDMs contain a singular text-to-image decoder model (f_θ), while unCLIP framework ($f_\theta = h_\theta \circ g_\phi$) contains two primary modules:

- **Text-to-Image Prior** ($g_\phi : z_y \rightarrow z_x$): This module maps the text embeddings to the corresponding vision embeddings. Ramesh et al. [36] showed that the diffusion model as T2I prior leads to slightly better performance than the autoregressive models. For each timestep t and a noised image embedding $z_x^{(t)} \sim q(t, z_x)$ (here, q is a forward diffusion process), the diffusion prior directly estimates noiseless z_x rather than estimating Gaussian noise distribution $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ as:

$$\mathcal{L}_{prior} = \mathop{\mathbb{E}}_{\substack{t\sim[0,T],\\ z_x^{(t)}\sim q(t,z_x)}} \left[||z_x - g_\phi(z_x^{(t)}, t, z_y)||_2^2 \right]. \quad (1)$$

- **Diffusion Image Decoder** ($h_\theta : (z_x, z_y) \rightarrow x$): This module generates the final image conditioned on the z_x and the input text features z_y. This diffusion decoder follows the standard diffusion training procedure by estimating $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ after [13]:

$$\mathcal{L}_{decoder} = \mathop{\mathbb{E}}_{\substack{\epsilon\sim N(0,I)\\ t\sim[0,T],\\ (z_x,\, z_y)}} \left[||\epsilon - h_\theta(x^{(t)}, t, z_x, z_y)||_2^2 \right]. \quad (2)$$

Different versions of the unCLIP decoder (i.e., Kandinsky and Karlo) vary in whether they include text conditioning (z_y) in the diffusion image decoder. Both approaches yield comparable results, provided that image conditioning (z_x) is accurate. The training objectives, $\mathcal{L}_{prior}$ and

$\mathcal{L}_{decoder}$, integrate Classifier-Free Guidance (CFG) [12], enhancing the model's generative capabilities.

3.2. Problem Formulation

Given the pivotal role of the T2I prior module in image generation from text, in this paper, our focus is on enhancing g_ϕ, while keeping the pre-trained h_θ frozen. Let's consider a training distribution P_{XY}, comprising input pairs of image and text (x, y). Maximizing the Evidence Lower Bound (ELBO) on the training distribution P_{XY} facilitates this mapping of $z_y \rightarrow z_x$. However, such a strategy does not inherently assure generalization, especially when the input text prompt (y) deviates from the assumed independently and identically distributed (i.i.d.) pattern of P_{XY} [45]. Therefore, attaining a more diverse and representative P_{XY} becomes crucial for improving the performance. While a diffusion prior combined with CFG has been shown to bolster generalization, especially with diverse training data and extensive training iterations [29], it is computationally expensive and is not always reliable (especially, in low resource constraint settings) as shown in Section 4.2. Given these constraints, our goal is to develop an alternative prior learning methodology that improves parameter efficiency (97% reduction) and mitigates the need for large-scale high-quality data ($\leq 5\%$) while maintaining the performance.

3.3. Proposed Method: *ECLIPSE*

This section elaborates on *ECLIPSE*, our model training strategy to learn text-to-image prior (g_ϕ). We focus on enhancing non-diffusion prior models through the effective distillation of pre-trained vision-language models, such as CLIP, while preserving the semantic alignment between the input text embedding z_y and corresponding estimated vision embeddings $\hat{z}_x$ by using the contrastive loss.

Base Prior Model. T2I diffusion prior deviates from the standard diffusion objective (such as Eq. 2). Unlike the standard ϵ prediction diffusion objective, the T2I diffusion prior objective instead estimates the z_x which is noiseless. Despite its convertibility, the empirical analysis (Section 5) shows that having more diffusion prior steps does not benefit the overall text-to-image generation abilities. During the inference, for diffusion priors, we still adhere to the conventional denoising process, introducing additional noise $(\sigma_t \epsilon)$ at each step, except for the final step according to Ho et al. [13]. This degradation in performance is likely due to the compressed latent space of the CLIP models. Moreover, if we repeat this for T timesteps, it can lead to the accumulation of errors, which is undesirable.

Therefore, to mitigate this unnecessary computing, we use non-diffusion T2I prior, making the prior model both parameter-efficient and less demanding in terms of computational resources. This non-diffusion architecture forms our base model, and we introduce the training objective that leverages pre-trained vision-language models trained on extensive datasets to improve generalization outside the P_{XY}.

Projection Objective. Despite vision-language models aligning the semantic distributions across modalities, each modality may exhibit unique distributions. Therefore, our approach involves projecting the text embedding onto the vision embedding. This is achieved using a mean squared error objective between the predicted vision embedding $(\hat{z}_x)$ and the ground truth vision embedding (z_x):

$$\mathcal{L}_{proj} = \mathop{\mathbb{E}}_{\substack{\epsilon \sim \mathcal{N}(0,I) \\ z_y, z_x}} \left[||z_x - g_\phi(\epsilon, z_y)||_2^2 \right], \tag{3}$$

where ϵ is the Gaussian input noise. Notably, as discussed previously, this is an approximation of the diffusion prior objective (Eq. 1) with $t = T$ and without CFG. $\mathcal{L}_{proj}$ learns latent posterior distribution with the *i.i.d.* data assumption. However, this model, fine-tuned on P_{XY}, may not generalize well beyond its distribution. The optimal solution would be to train on a dataset that encapsulates all potential distributions to cover all possible scenarios, which is an impractical and resource-consuming task.

CLIP Contrastive Learning. To address these limitations, we propose utilizing the CLIP more effectively, which contains the semantic alignment between image and language. Specifically, we apply the CLIP Contrastive Loss after [34] to train the T2I priors. For a given input batch $\{(z_x^i, z_y^i)\}_{i=1}^N$ from the P_{XY} distribution, we calculate the text-conditioned image contrastive loss for the i^{th} image embedding prediction relative to the all input ground truth text embeddings as:

$$\mathcal{L}_{CLS;\, y \rightarrow x} = -\frac{1}{N} \sum_{i=0}^{N} \log \frac{\exp(\langle \hat{z}_x^i, z_y^i \rangle / \tau)}{\sum_{j \in [N]} \exp(\langle \hat{z}_x^i, z_y^j \rangle / \tau)}, \tag{4}$$

where τ is the temperature parameter, $\langle , \rangle$ denotes the cosine similarity, and N is the batch size. This loss encourages the model to understand and follow the input text better, effectively reducing overfitting to the P_{XY}, as illustrated in Figure 2. Consequently, the final objective function is:

$$\mathcal{L}_{ECLIPSE} = \mathcal{L}_{proj} + \lambda * \mathcal{L}_{CLS;\, y \rightarrow x}, \tag{5}$$

where λ is the hyperparameter balancing the regularizer's effect. Overall, the final objective function aims to map the text latent distribution to the image latent distribution via $\mathcal{L}_{proj}$ and such that it preserves the image-text alignment using $\mathcal{L}_{CLS;\, y \rightarrow x}$. This makes the prior model generalize beyond the given training distribution P_{XY} such that it can follow the semantic alignment constraint. Importantly, we cannot use $\mathcal{L}_{CLS;\, y \rightarrow x}$ alone or with a high value of λ as

Table 1. The comparison (in terms of FID and compositions) of the baselines and state-of-the-art methods with respect to the *ECLIPSE*. * indicates the official reported ZS-FID. Ψ denotes the FID performance of a model trained on MSCOCO. The best performing *ECLIPSE* variant (with respect to its big counterpart) is highlighted by green. *ECLIPSE* consistently outperforms the SOTA big models despite being trained on a smaller subset of dataset and parameters.

Methods	Model Type	Training Params [M]*	Total Params [B]	Data Size [M]	ZS-FID (↓)	T2I-CompBench Color (↑)	Shape (↑)	Texture (↑)	Spatial (↑)	Non-Spatial (↑)
Stable Diffusion v1.4	LDM	900	0.9	400	16.31*	0.3765	0.3576	0.4156	0.1246	0.3076
Stable Diffusion v2.1	LDM	900	0.9	2000	14.51*	0.5065	0.4221	0.4922	0.1342	0.3096
Würstchen	unCLIP	1000	2.0	1420	23.60*	0.3216	0.3821	0.3889	0.0696	0.2949
Kandinsky v2.1	unCLIP	1000	2.22	177	18.09	0.4647	0.4725	0.5613	0.1219	0.3117
DALL-E-2	unCLIP	1000	4.5	250	10.65*	0.5750	0.5464	0.6374	0.1283	0.3043
Karlo	unCLIP	1000	1.9	115	20.64	0.5127	0.5277	0.5887	0.1337	0.3112
***ECLIPSE* (ours)**	Karlo	**33**	**0.93**	**0.6**$_{\text{MSCOCO}}$	23.67$^{\Psi}$	0.5965	0.5063	0.6136	0.1574	0.3235
		33	**0.93**	**2.5**$_{\text{CC3M}}$	26.73	0.5421	0.5090	0.5881	0.1478	0.3213
		33	**0.93**	**10.0**$_{\text{CC12M}}$	26.98	0.5660	0.5234	0.5941	0.1625	0.3196
Kandinsky v2.2	unCLIP	1000	2.22	177	20.48	0.5768	0.4999	0.5760	0.1912	0.3132
***ECLIPSE* (ours)**	Kandinsky v2.2	**34**	**1.26**	**0.6**$_{\text{MSCOCO}}$	16.53$^{\Psi}$	0.5785	0.4951	0.6173	0.1794	0.3204
		34	**1.26**	**5.0**$_{\text{HighRes}}$	19.16	0.6119	0.5429	0.6165	0.1903	0.3139

the prior model will converge outside the vision latent distribution that optimizes the contrastive loss (such input text latent space itself). And keeping λ to a very low value cannot do knowledge distillation well enough. Empirical studies suggest setting $\lambda = 0.2$ for optimal performance, balancing knowledge distillation, and maintaining alignment within the vision latent distribution.

4. Experiments & Results

This section introduces the datasets, training specifications, comparative baselines, and evaluation metrics utilized in our experiments. We conduct an extensive assessment of our proposed *ECLIPSE* methodology and its variants.

4.1. Experimental Setup

Dataset. Our experiments span four datasets of varying sizes: MSCOCO [24], CC3M [42], CC12M [3], and LAION-HighResolution[2] [41]. MSCOCO comprises approximately 0.6 million image-text pairs, while CC3M and CC12M contain around 2.5 and 10 million pairs, respectively [3]. We select a very small subset of 5 million (2.8%) image-text pairs from the LAION-HighRes dataset (175M). We perform Karlo diffusion image decoder-related experiments on MSCOCO, CC3M, and CC12M as these datasets are subsets of the data used to train the Karlo diffusion image decoder. Similarly, we use MSCOCO and LAION-HighRes for the Kandinsky decoder.

Baselines. *ECLIPSE* variants are compared against leading T2I models, including Stable Diffusion, Würstchen, Karlo, Kandinsky, and DALL-E-2. Additionally, we introduce two more baselines along with *ECLIPSE* to evaluate the impact of our training strategy in a resource-constrained environment: 1) Projection: A non-diffusion prior model trained with $\mathcal{L}_{proj}$ (Eq. 3). 2) Diffusion-Baseline: A diffusion prior model trained with $\mathcal{L}_{prior}$ (Eq. 1) – the traditional T2I prior, and 3) *ECLIPSE*: A non-diffusion prior model trained with our proposed methodology $\mathcal{L}_{ECLIPSE}$ (Eq. 5).

Training and inference details. We evaluate *ECLIPSE* using two pre-trained image decoders: Karlo-v1-alpha and Kandinsky v2.2, trained on distinct CLIP vision encoders. Our prior architecture is based on the standard PriorTransformer model [36], modified to be time-independent. The detailed architecture is outlined in the appendix. We configure prior models with 33 and 34 million parameters for Karlo and Kandinsky, respectively. This contrasts with larger models in the field, which often use up to 1 billion parameters (as summarized in Table 1). The Projection, Diffusion-Baseline, and *ECLIPSE* priors are trained for both diffusion image decoders, maintaining consistent hyperparameters (including total number of parameters) across all models. Training on CC12M, CC3M, and LAION-HighRes is performed on 4 x RTX A6000 GPUs with a 256 per-GPU batch size, a learning rate of 0.00005, and the CosineAnnealingWarmRestarts scheduler [25]. Each model undergoes approximately 60,000 iterations, totaling around 200 GPU hours. For MSCOCO, training takes about 100 GPU hours. This can be further reduced to $\leq$ 50 GPU hours if image-text pairs are preprocessed beforehand. The diffusion prior is trained with a linear scheduler and 1000 DDPM timesteps. Inferences utilize 25 DDPM steps with 4.0 classifier-free guidance, while Projection and *ECLIPSE* models do not require diffusion sampling. Image diffusion decoders are set to 50 DDIM steps and 7.5 classifier-free guidance.

Evaluation setup. Our evaluation framework encompasses various metrics. We employ MS-COCO 30k to assess

[2] https://huggingface.co/datasets/laion/laion-high-resolution

[3] According to the download date: 08/26/2023

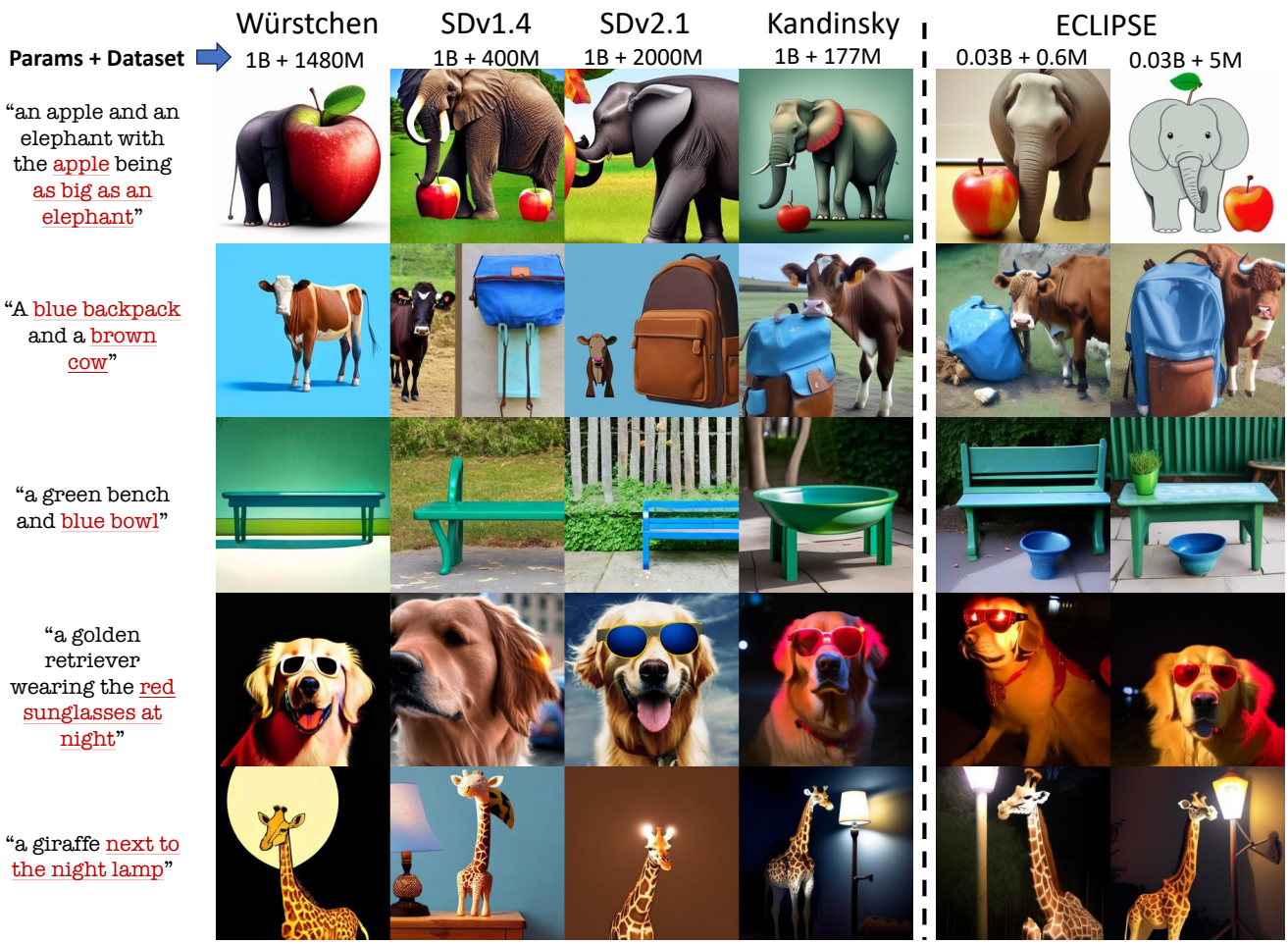


Figure 3. Qualitative result of our text-to-image prior, *ECLIPSE*, comparing with SOTA T2I model. Our prior model reduces the model parameter requirements (from 1 Billion → 33 Million) and data requirements (from 177 Million → 5 Million → 0.6 Million). Given this restrictive setting, *ECLIPSE* performs close to its huge counterpart (i.e., Kandinsky v2.2) and even outperforms models trained on huge datasets (i.e., Würstchen, SDv1.4, and SDv2.1) in terms of compositions.

FID [11] and T2I-CompBench [14] for evaluating composition abilities in color, shape, texture, spatial, and non-spatial compositions. Given the impracticality of large-scale human studies, we approximate human preferences using PickScore [19], reporting results on the T2I-CompBench validation set comprising about 1500 unique prompts.

4.2. Quantitative Evaluations

In Table 1, we present a performance comparison between *ECLIPSE* variants and leading T2I models. Our evaluation metrics include zero-shot Fréchet Inception Distance (FID) on MS-COCO 30k for image quality assessment and T2I-CompBench [14] for evaluating compositionality. *ECLIPSE* priors, trained with both types of diffusion image decoders, demonstrate notable improvements. *ECLIPSE* consistently surpasses various baselines in terms of compositionality, irrespective of the dataset size. Its performance is comparable to that of DALL-E-2 and other SOTA models, a significant improvement considering *ECLIPSE*'s parameter efficiency. Standard T2I priors usually incorporate 1 billion parameters, while *ECLIPSE* operates with only 3.3% of these parameters, maintaining competitive performance levels. When combined with corresponding diffusion image decoders, the total parameter count of *ECLIPSE* is close to that of Stable Diffusion models, yet it outperforms them, especially considering that the latter are trained on a massive set of image-text pairs. A noticeable decline in zero-shot FID (ZS-FID) is observed in comparison to the original Karlo. We attribute this variation to the image quality differences in the training dataset, suggesting a potential area for further investigation and improvement. At the same time, if we utilize the smaller subset of high-resolution datasets then we can still maintain better FID and improve the compositions, as shown in the

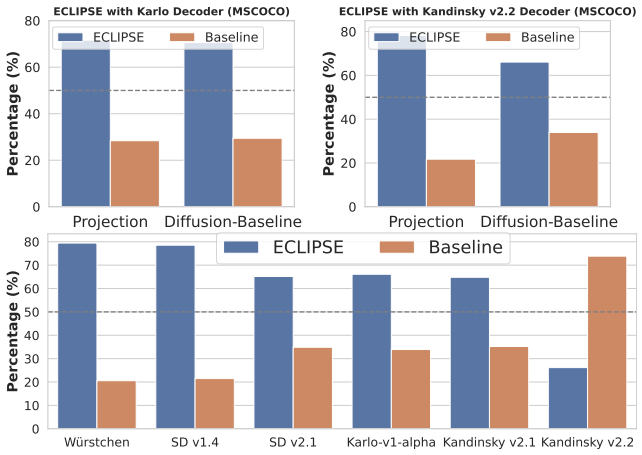


Figure 4. Qualitative evaluations by human preferences approximated by the PickScore [19]. The top two figures compare *ECLIPSE* to Projection and Diffusion Baselines trained with the same amount of data and model size for both Karlo and Kandinsky decoders. In the bottom figure, we compare *ECLIPSE* with the Kandinsky v2.2 decoder trained on the LAION-HighRes dataset against SOTA models.

last row of Table 1. *ECLIPSE* prior with Kandinsky v2.2 decoder trained on LAION-HighRes subset achieves similar FID to other original Kandinsky v2.2 unCLIP model and at the same time outperforming in terms of compositions.

Table 2 provides a comparison of various baseline training strategies for small prior models, using identical datasets and hyperparameters. *ECLIPSE* exhibits superior performance across all datasets. We also note that diffusion priors benefit from larger datasets, supporting our premise that such priors necessitate extensive training data for optimal results, which is also attributed to the CFG. In contrast, *ECLIPSE* demonstrates the consistent performance on compositions irrespective of the amount of image-text pairs.

4.3. Qualitative Evaluations

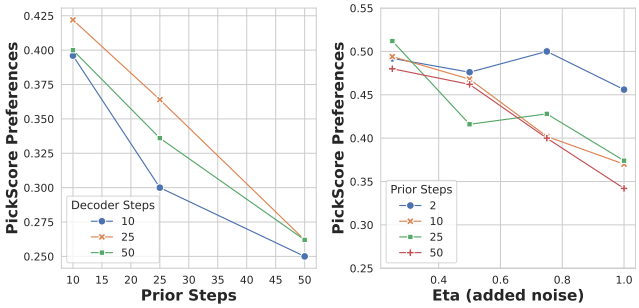
In Figure 3, we display qualitative examples from various methods responding to complex prompts. *ECLIPSE* demonstrates superior performance in comparison to Stable Diffusion v1.4, Stable Diffusion v2.1, and Würstchen, while closely matching the quality of its big counterpart, Kandinsky v2.2. Interestingly, *ECLIPSE* trained on only 0.6 million images maintains the compositions with minor degradation in image quality. These observations align with our previously established quantitative results. Beyond numerical metrics, understanding human preferences is crucial. To this end, we selected 1500 unique validation prompts from T2I-CompBench and assessed PickScore preferences. The results, illustrated in Figure 4, reveal that *ECLIPSE* notably surpasses its baselines in respective restrictive settings with an average score of 71.6%. We can also observe that the best *ECLIPSE* variant (with Kandinsky decoder and trained on LAION-HighRes) consistently outperforms the other big SOTA models achieving an average performance of 63.36%. We observe that in terms of preferences, the original Kandinsky v2.2 diffusion prior (with a 1 billion parameter) trained on LAION-HighRes (175M) performs better than the *ECLIPSE* prior (having 33 million parameters). We hypothesize that this might be due to its use of a large-scale dataset that contains more aesthetically pleasing images. We provide a set of qualitative results in the appendix to show that *ECLIPSE* performs similarly well, if not better, *w.r.t.* semantic understanding of the text.

Table 2. Comparison of *ECLIPSE* with respect to the various baseline prior learning strategies on four categories of composition prompts in the T2I-CompBench. All prior models are of 33 million parameters and trained on the same hyperparameters.

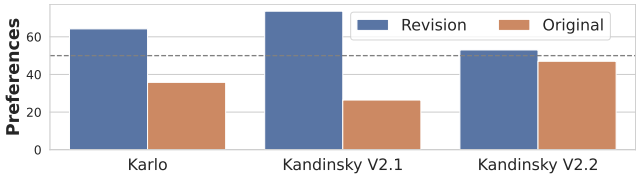
Methods	T2I-CompBench			
	Color (↑)	Shape (↑)	Texture (↑)	Spatial (↑)
MSCOCO with Karlo				
Projection	0.4667	0.4421	0.5051	0.1478
Diffusion-Baseline	0.4678	0.4797	0.4956	0.1240
ECLIPSE	**0.5965**	**0.5063**	**0.6136**	**0.1574**
CC3M with Karlo				
Projection	0.4362	0.4501	0.4948	0.1126
Diffusion-Baseline	**0.5493**	0.4809	0.5462	0.1132
ECLIPSE	0.5421	**0.5091**	**0.5881**	**0.1477**
CC12M with Karlo				
Projection	0.4659	0.4632	0.4995	0.1318
Diffusion-Baseline	0.5390	0.4919	0.5276	0.1426
ECLIPSE	**0.5660**	**0.5234**	**0.5941**	**0.1625**
MSCOCO with Kandinsky v2.2				
Projection	0.4678	0.3736	0.4634	0.1268
Diffusion-Baseline	0.4646	0.4403	0.4834	0.1566
ECLIPSE	**0.5785**	**0.4951**	**0.6173**	**0.1794**
HighRes with Kandinsky v2.2				
Projection	0.5379	0.4983	0.5217	0.1573
Diffusion-Baseline	0.5706	0.5182	0.5067	0.1687
ECLIPSE	**0.6119**	**0.5429**	**0.6165**	**0.1903**

5. Analysis

Analyzing the traditional diffusion priors. To further support our choice of using non-diffusion prior models, we analyze the existing diffusion prior formulation. We conducted two key empirical studies: 1) Evaluating the Impact of Prior Steps: We examined how the number of prior steps influences model performance. 2) Assessing the Influence of Added Noise ($\sigma_t \epsilon$): We focused on understanding how the introduction of noise affects human preferences. For these studies, we utilized PickScore preferences, and the outcomes, depicted in Figure 5, corroborate our hypothesis: both the prior steps and the addition of ($\sigma_t \epsilon$) detrimentally affect performance. Furthermore, as indicated in Table 2, diffusion prior surpasses the projection baseline if provided with more high-quality data. We attribute this enhanced performance to the incorporation of classifier-free guidance, which bolsters the model's generalization capabilities to a



(a) Left: Performance comparison by varying the prior steps and decoder steps *w.r.t.* the fixed prior steps ($t = 2$). Right: Performance comparison by varying the mean η of the added scheduler noise ($\sigma_t \epsilon$) *w.r.t.* the noiseless predictions ($\eta = 0$). Both experiments are on the Kandinsky v2.1.



(b) Overall performance comparisons on various pre-trained unCLIP models before and after reducing the prior steps to two and η to 0.0.

Figure 5. Empirical analysis of the PickScore preferences of diffusion priors with respect to the various hyper-parameters.

certain extent. However, it's worth noting that both baselines are still outperformed by *ECLIPSE*. This observation underscores the effectiveness of our proposed methodology in comparison to traditional approaches in the realm of T2I.

Importance of data selection. In our previous analysis (Table 1 and 2), we demonstrated that *ECLIPSE* attains competitive performance on composition benchmarks regardless of dataset size. This achievement is largely due to the integration of the contrastive loss $\mathcal{L}_{CLS}$ (Eq.4). However, the final objective function also incorporates the $\mathcal{L}_{proj}$ (Eq.3), which is pivotal in estimating the vision latent distribution. This estimation is fundamentally dependent on the training distribution (P_{XY}), leading the model to learn spurious correlations within P_{XY}. Consequently, the model's image quality could directly correlate with the overall quality of images in the training set. To further substantiate this, we evaluated the preferences for *ECLIPSE* models trained on MSCOCO, CC3M, and CC12M, in comparison to among themselves and Karlo-v1-alpha. The outcomes, presented in Figure 6, reveal that the *ECLIPSE* model trained on CC12M outperforms those trained on other datasets, exhibiting performance on par with its big counterpart. *ECLIPSE* prior (w Karlo decoder) trained on the CC12M dataset performs comparably to Karlo-v1-alpha while *ECLIPSE* priors trained on other datasets struggle to do so. Furthermore, as illustrated in Figure 6, the *ECLIPSE* model trained on MSCOCO demonstrates a tendency to learn spurious correlations, such as associating the term "young tiger" with the person.

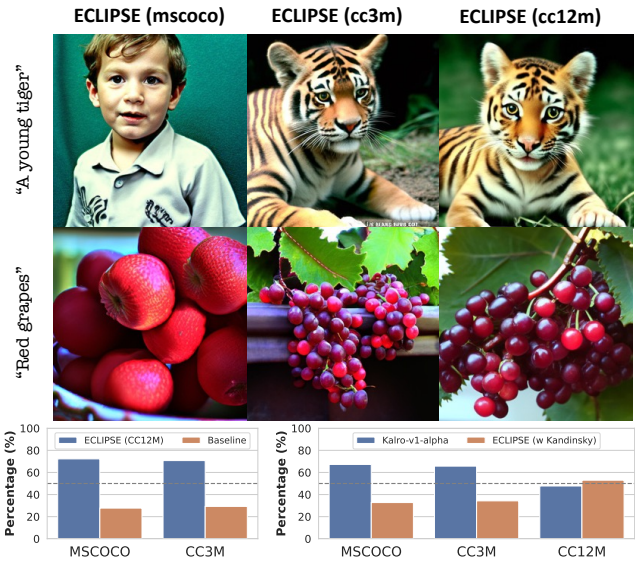


Figure 6. The top figure shows the qualitative examples of the biases learned by the T2I prior models. Bottom figures show the PickScore preferences of the *ECLIPSE* models trained on various datasets with respect to the other datasets (left) and Karlo (right).

6. Conclusion

In this paper, we introduce a novel text-to-image prior learning strategy, named *ECLIPSE*, which leverages pre-trained vision-language models to provide additional supervision for training the prior model through contrastive learning. This approach significantly enhances the training efficiency of prior models in a parameter-efficient way. Through comprehensive quantitative and qualitative evaluations, we assessed *ECLIPSE* priors alongside various diffusion image decoders. The results indicate that *ECLIPSE* surpasses both the baseline projection models and traditional diffusion-prior models. Remarkably, *ECLIPSE* achieves competitive performance alongside larger, state-of-the-art T2I models. It demonstrates that priors can be trained with merely 3.3% of the parameters and 2.8% of image-text pairs typically required, without compromising the performance. This advancement directly leads to at least 43% overall compression of the unCLIP models. Our findings show that pre-trained vision-language can be utilized more effectively; suggesting promising research direction where improving the vision-language models may directly benefit the T2I.

Acknowledgement

This work was supported by NSF RI grants #1750082 and #2132724, and a grant from Meta AI Learning Alliance. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Pranav Aggarwal, Hareesh Ravi, Naveen Marri, Sachin Kelkar, Fengbin Chen, Vinh Khuc, Midhun Harikumar, Ritiz Tambi, Sudharshan Reddy Kakumanu, Purvak Lapsiya, et al. Controlled and conditional text to image generation with diffusion prior. *arXiv preprint arXiv:2302.11710*, 2023. 11

[2] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023. 1, 2

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 2

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 1, 2

[7] Yuren Cong, Martin Renqiang Min, Li Erran Li, Bodo Rosenhahn, and Michael Ying Yang. Attribute-centric compositional text-to-image generation. *arXiv preprint arXiv:2301.01413*, 2023. 2, 3

[8] Lee Donghoon, Kim Jiseob, Choi Jisu, Kim Jongmin, Byeon Minwoo, Baek Woonhyuk, and Kim Saehoon. Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo, 2022. 1, 2

[9] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *arXiv preprint arXiv:2303.15435*, 2023. 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[11] 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. *Advances in neural information processing systems*, 30, 2017. 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 4

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 4

[14] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 1, 2, 6

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[16] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. On architectural compression of text-to-image diffusion models. *arXiv preprint arXiv:2305.15798*, 2023. 2

[17] Changhoon Kim, Yi Ren, and Yezhou Yang. Decentralized attribution of generative models. In *International Conference on Learning Representations*, 2021. 2

[18] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. *arXiv preprint arXiv:2306.04744*, 2023. 2

[19] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 6, 7

[20] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 237–246, 2021. 2

[21] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1

[22] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980*, 2023. 2

[23] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207, 2022. 2, 3

[24] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 5

[26] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[27] 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. 2

[28] Guangyu Nie, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributing image generative models using latent fingerprints. *arXiv preprint arXiv:2304.09752*, 2023. 2

[29] Maya Okawa, Ekdeep Singh Lubana, Robert P Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *arXiv preprint arXiv:2310.09336*, 2023. 4

[30] Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. *arXiv preprint arXiv:2306.04695*, 2023. 1

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[32] Pablo Pernias, Dominic Rampas, and Marc Aubreville. Wuerstchen: Efficient pretraining of text-to-image models. *arXiv preprint arXiv:2306.00637v2*, 2023. 2

[33] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023. 2

[34] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2, 3, 5

[37] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023. 1, 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn 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. 1, 2

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[40] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021. 2

[41] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[45] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 4

[46] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[47] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 2

[48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[49] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 2