Learning Visual Generative Priors without Text

Shuailei Ma[1*], Kecheng Zheng[2*], Ying Wei[1✉], Wei Wu[2], Fan Lu[2],
Yifei Zhang[3], Chen-Wei Xie[4], Biao Gong[2], Jiapeng Zhu[5], Yujun Shen[2✉]

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[2] Ant Group   [3] Shanghai Jiao Tong University   [4] Alibaba Group   [5] HKUST

https://ant-research.github.io/lumos

(a) Text-to-Image Generation

(b) Novel View Synthesis

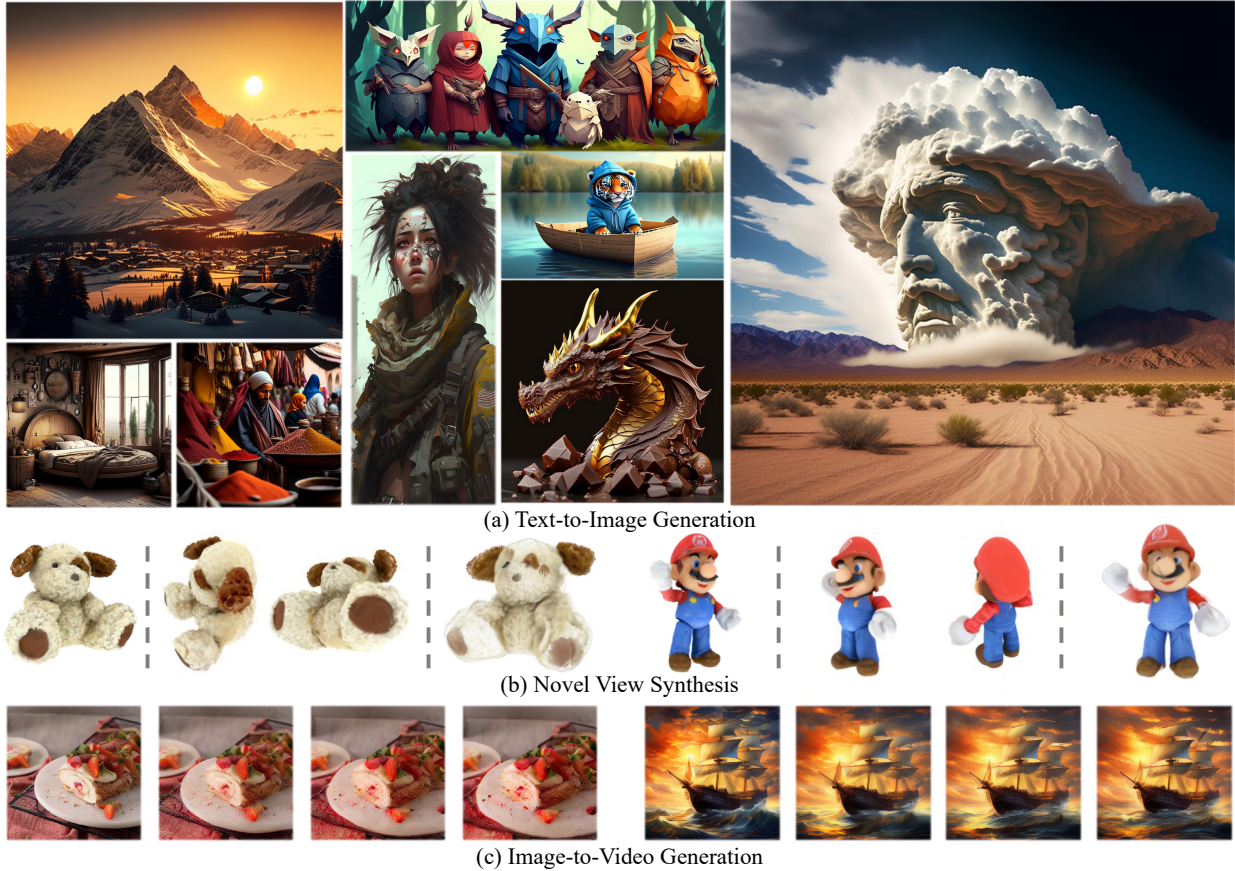
(c) Image-to-Video Generation

Figure 1. **Diverse downstream tasks** of `Lumos` including (a) text-to-image generation, (b) novel view synthesis (left: input view, middle: random novel views, right: reconstruction Gaussian) and (c) image-to-video generation.

Abstract

*Although text-to-image (T2I) models have recently thrived as visual generative priors, their reliance on high-quality text-image pairs makes scaling up expensive. We argue that grasping the cross-modality alignment is **not** a necessity for a sound visual generative prior, whose focus should be on texture modeling. Such a philosophy inspires us to study image-to-image (I2I) generation, where models can learn from in-the-wild images in a self-supervised manner. We first develop a pure vision-based training framework, `Lumos`, and confirm the feasibility and the scalability of learning I2I models. We then find that, as an upstream task of T2I, our I2I model serves as a more foundational visual prior and achieves on-par or better performance than existing T2I models using only 1/10 text-image pairs for fine-tuning. We further demonstrate the superiority of I2I priors over T2I priors on some text-irrelevant visual generative tasks, like image-to-3D and image-to-video.*

---

* : Equal contribution. ✉ : Corresponding author.

1. Introduction

Text-to-Image (T2I) generative models [2, 17, 40, 41, 43], which showcase the stunning ability to generate high-fidelity images from a given text prompt, have made a remarkable leap in the evolution of visual generation. These well-trained T2I generative models are usually regarded as visual generative priors for downstream visual synthesis tasks, such as image-to-video (I2V) [4] and novel view synthesis (NVS) [30], endowing the downstream models with rich semantic information.

However, recent research [7, 50] has demonstrated that a well-trained T2I generative model intensely depends on high-quality image-text pairs. As shown in Figure 2a, with the noise ratio of text-image pairs increasing from 10% to 90%, the performance of the T2I model degrades around 1.0 CLIP score. It indicates that the performance of T2I model is sensitive to the quality of text-image pairs, because T2I models simultaneously focus on two difficult issues: learning *texture modeling* and *text-image alignment*. Noisy texts of images may be harmful to *text-image alignment* and further exacerbate the difficulty of *texture modeling* in generative models. Scaling up the well-aligned image-text paired data can address this issue but is quite expensive. As a result, it is a necessary step towards unleashing the power of large-scale unannotated data. A commonly adopted strategy is to utilize unconditional generative methods [28, 41] to model texture, which can learn the probabilistic distributions of in-the-wild data without requiring paired data. With this observation, we argue that grasping the cross-modality alignment is **not** a necessity for a sound visual generative prior, whose focus should be on texture modeling.

Such a philosophy inspires us to study image-to-image (I2I) generation, where models can learn from in-the-wild images in a self-supervised manner. We first develop a pure vision-based framework, termed as Lumos, to generate images conditioned on image features extracted by a pretrained vision encoder (*e.g.*, DINO). Then, we fine-tune Lumos to various downstream visual synthesis tasks (*e.g.*, T2I). As shown in Figure 2a, thanks to I2I prior, our T2I model shows less reliance on high-quality data, exhibiting 0.5 less degradation in CLIP score compared with *w/o* I2I prior. Moreover, as illustrated in Figure 2b, with the data of I2I scaling up from 10M to 200M, our model achieves better performance, confirming the feasibility and the scalability of learning I2I priors. We hypothesize that the performance gap arises because the I2I priors enhance texture modeling, thereby simplifying the T2I generative process. Another interesting question lies in the influence of image encoders on I2I model. For a T2I model, text-image alignment is crucial. Therefore, a straightforward approach would be to use a text-image-aligned CLIP in the I2I model, which may enhance the T2I model's performance compared to DINO.

(a) Noise data for T2I.   (b) Training curve of scaling.

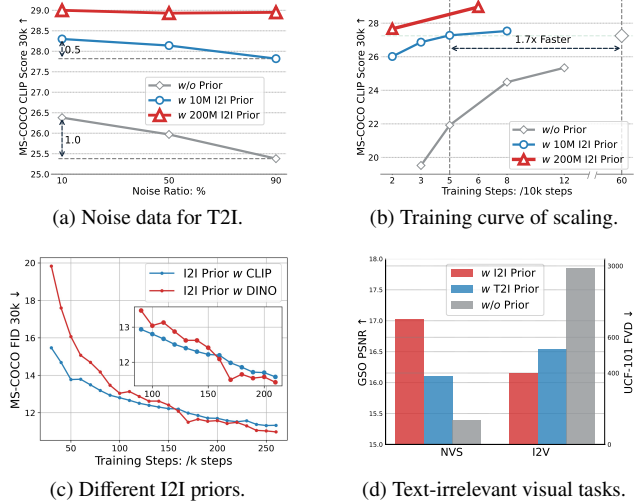(c) Different I2I priors.   (d) Text-irrelevant visual tasks.

Figure 2. **Various image generation tasks can be improved by our image-to-image priors.** I2I prior enables the downstream T2I model to decrease dependence on high-quality data, and with data scaling up in I2I, it enjoys a larger performance improvement. We also adopt a pure vision-based I2I generation (*i.e.*, I2I prior with DINO) that is a late bloomer for T2I generation. We further demonstrate the superiority of I2I priors over T2I priors on some text-irrelevant vision tasks, like I2V and NVS.

However, as shown in Figure 2c, although CLIP shows greater advantages in the early stages of learning I2I priors, DINO, as a late bloomer, achieves better COCO-30K FID than CLIP in the final steps. It demonstrates the feasibility of the pure vision-based training framework. Based on these, we find that, as an upstream task of T2I, our pure-vision I2I model serves as a more foundational visual prior.

We further demonstrate the superiority of I2I priors over T2I priors on some text-irrelevant vision tasks, like image-to-3D and image-to-video. As shown in Figure 2d, we observe that I2I priors bring better performance than T2I priors. On these text-irrelevant vision tasks, it is hard to design a suitable text prompt. However, for I2I prior, it naturally does not require text input and can be efficiently transferred directly to these tasks. Thus, our I2I prior can achieve better performance than T2I prior, which can serve as a more foundational visual prior. We hope our method, results, and analysis will encourage future research on pure visual generative priors.

2. Related Work

2.1. Unsupervised Representation Learning

Unsupervised pre-training methods [6, 9, 10, 15, 20, 37] are undeniably one of the key contributors to the success of deep learning nowadays, as they break the shackles of expensively supervised data on the deep learning model. For Natural Language Processing, masked language modeling and its autoregressive counterparts, *e.g.*, BERT [15] and

GPT [37], first demonstrated the potential of unsupervised pre-training. Similarly, exploring unsupervised pre-training for visual understanding tasks has never stopped. Self-supervised learning methods, *e.g.*, MoCo [10, 20], sim-CLR [9], and DINO [6], are pre-trained in the form of self-distillation without labels, proving that they can learn to extract richer features from images than the supervised counterparts. Mask image modeling, *e.g.*, MAE [21] and SimMIM [51] ingeniously adapt the concept of masked modeling from the NLP field to image processing, leading to remarkably improved results in understanding pre-training. However, the exploration on unsupervised learning in visual generation remains limited. In this work, we investigate developing an unsupervised visual generative learning framework, that provides foundational visual priors for various visual synthesis tasks (*e.g.*, text-to-image).

2.2. Generative Diffusion Models

Visual generative models have undergone long-term development, targeting learning the probabilistic distributions of data from the images. Some fundamental methods, such as GAN [19], VAE [25], and Flow [42], show impressive performance in modeling simple image distributions such as animal face and indoor scene. In the complex data, generative diffusion models [2, 17, 43] have achieved superior results (*e.g.*, text-to-image generation), whose contents are high-quality. However, current research [2, 7] shows that the success of generative diffusion models is inseparable from high-quality text-image pairs, which greatly limits their scalability. To address this problem, PixArt-α [7] mentions that class-to-image pre-training on ImageNet [14] can serve as a pre-trained model for text-to-image generation, accelerating the convergence speed of the model. But still, human annotations are expensive and hard to scale up. Inspired by unconditional generation, DALLE-2 [41] and RCG [28] proposed to use the image-to-image model as an intermediate bridge between the downstream models (*i.e.*, T2I) and the unconditional generation diffusion model. In this paper, we explore a novel I2I generative method that captures pure visual priors from in-the-wild images, and confirm its feasibility and scalability.

3. Method

In this section, we firstly introduce the image-to-image generation to learn from in the wild images in the self-supervised manner. Then, we transfer the well-trained I2I model, as the image-to-image prior, to some downstream vision tasks.

3.1. Image-to-Image Generative Model

Given an image $x \in \mathbb{R}^{h \times w \times 3}$, we leverage the pre-trained autoencoder [43] $\mathcal{E}$ to get its latent representation $z = \mathcal{E}(x)$. Then, we utilize the off-the-shelf vision encoder (*e.g.*,

DINO) [6, 10, 20, 38] τ^{image} to extract its visual semantic feature $\tau^{\text{img}}(x) \in \mathbb{R}^{M \times d}$. In this way, the pre-trained vision encoders, achieving state-of-the-art performance in representation learning, can provide rich semantic content, which is crucial for guiding image generation. Meanwhile, it is simple enough to scale up. As shown in Fig. 3 (a), condition on the resulting semantic features, we then learn the image-to-image latent diffusion model $\epsilon_{\theta_{I2I}}$ via

$$\mathcal{L}_{\theta_{I2I}} = \mathbb{E}_{\mathcal{E}(x), x, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta \left(z_t, t, \tau^{\text{img}}(x) \right) \right\|_2^2 \right], \quad (1)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and both of $\mathcal{E}$ and τ^{img} remain frozen throughout the pre-training process.

We use DiT [35] as the vision backbone. We conduct ablation experiments to explore different types of vision encoders (*e.g.*, pure-vision pre-training *vs.* language-image pre-training), and visual semantic condition types (global feature *vs.* local features). We briefly summarize our observations as following:

- *Pure-vision visual encoder is better than the language image pre-training encoder for I2I generation.* Compared to CLIP [38], self-supervised pure-vision visual encoders (*e.g.*, DINO [6] and MoCo [10, 20]) exhibit a faster convergence speed and better performance in terms of the FID metric for image-to-image generation.
- *Scaling up the number of images can improve the image generation for I2I generation.* As the data of I2I scales up from 10M to 200M, I2I generative model achieves better performance in FID.
- *Using local features as condition decreases the difficulty of I2I generation.* Local features from pre-trained vision encoders, including more detailed content from images, show a significant advantage for I2I generation than global feature.

The ablation study is presented in Sec. 4.3. Based on these observations, I2I generation can utilize the vast abundance of unannotated data to model rich texture.

3.2. Generative Transfer Learning

After training a I2I generation model, a fine-tuning stage transfers it to various downstream visual synthesis tasks. In this section, we regard widely-used text-to-image generation, novel view synthesis, and image-to-video tasks as our downstream tasks.

Text-to-Image Generation. Following LDM [43] and PixArt-α [7], we use the cross-attention mechanism [49] to attach the condition information from text prompts y (see details in Fig. 3 (a) and (b)). After loading the I2I pre-trained weight θ_{I2I}, we tune the diffusion model $\epsilon_{\theta_{T2I}}$ via

$$\mathcal{L}_{\theta_{T2I}} = \mathbb{E}_{\mathcal{E}(x), y, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta \left(z_t, t, \tau^{\text{txt}}(y) \right) \right\|_2^2 \right], \quad (2)$$

where τ^{txt} represents the pre-trained text encoder. Both τ^{txt} and $\mathcal{E}$ are kept frozen during training.
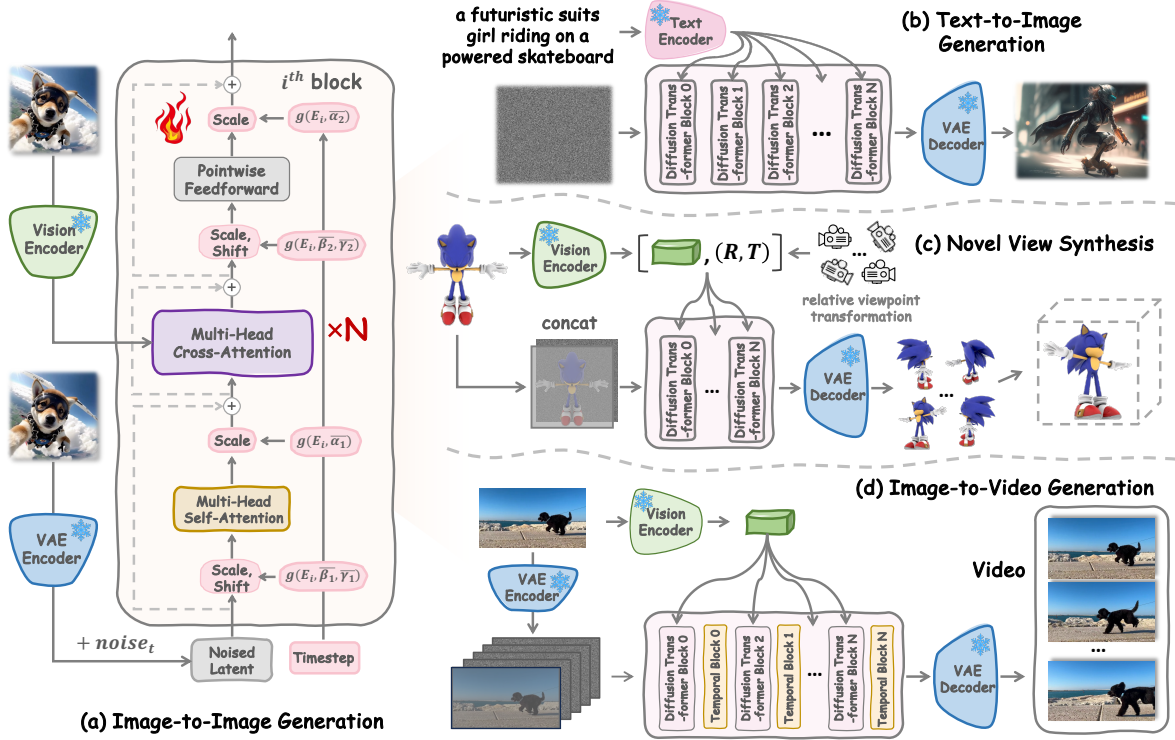
Figure 3. **Overall architecture of our framework.** (a) Image-to-Image Generation, (b) Text-to-Image Generation, (c) Novel View Synthesis and (d) Image-to-Video Generation.

T2I generation is a challenge task that allows model simultaneously focus on two difficult issues: learning *texture modeling* and *text-image alignment*. Our I2I model can provide the visual generative priors to help texture modeling, and let T2I model pay more attention to text-image alignment. We briefly summarize our observations as following:

- *The performance of upstream and downstream is inconsistency.* The performance of downstream tasks is not significantly influenced by the performance (*e.g.*, FID) of I2I model. For example, when using local features as the condition of I2I model, the FID of downstream T2I model is worse than one of global feature.

- *Scaling up I2I generation can improve the performance of downstream tasks.* With the data of I2I scaling up from 10M to 200M, our T2I model achieves better performance in terms of FID, confirming the feasibility and the scalability of learning I2I priors.

- *Pure-vision visual encoder is better than language-image pre-training models for T2I generation.* Although CLIP shows greater advantages in the early stages of learning I2I priors, DINO, as a late bloomer, achieves better FID than CLIP in the final steps.

Based on these observation, we find that, as an upstream task of T2I, our pure-vision I2I model serves as a more foundational visual prior. For the detailed comparison, see the ablation experiments in Sec. 4.3.

Text-irrelevant Visual Generative Tasks. We further demonstrate the superiority of I2I priors over T2I priors on some text-irrelevant vision tasks, like image-to-3D and image-to-video. Specifically, following Zero-1-to-3 [30], we tune our I2I model to the novel view synthesis task. As shown in Fig. 3 (c), the task is to synthesize an image of an object from a new camera viewpoint. Moreover, We also finetune our pre-trained I2I model for the image-to-video generation task, where the video model receives a still input image as the condition as shown in Fig. 3 (d). On these text-irrelevant vision tasks, previous methods adopt pre-trained T2I model as visual generative prior, where it is hard to design a suitable text prompt. However, for I2I prior, it naturally does not require text input and can be efficiently transferred directly to these tasks. Our I2I prior can achieve better performance than T2I prior, which can serve as a more foundational visual prior.

4. Experiments

4.1. Image-to-Image Generation

Implementation Details & Datasets. Since our image-to-image generation does not require labels, we can easily filter totaling 190 million images via some criteria of image quality from existing open-source datasets (*i.e.*, LAION-5B [45], COYO-700M [5], SAM [26], JourneyDB [34], and ImageNet-1K [14]). More details can refer to *Supplemen-*

tary Material. We use DINO-B as the vision encoder of the image-to-image model and DiT-XL-2 [35] as the diffusion backbone. We evaluate the image-to-image model via Fréchet Inception Distance (FID) [22] on MSCOCO [29] and ImageNet-1K [14].

Comparison with SOTA I2I model. We compare our model with the open-source state-of-the-art I2I model, *i.e.*, RCG [28], which is pre-trained on ImageNet-1K [14]. For

Table 1. **Comparison with RCG in I2I generation.**

Method	MS-COCO FID 30K ↓	ImageNet FID 50K ↓
RCG [28]	12.70	4.89
Lumos-I2I	**4.82**	**2.60**

the RCG model, we use the open-source version, namely pre-trained DiT-XL/2 (the same as our diffusion model structure) conditioned on Moco-v3 ViT-B [20]. As shown in Table 1, Lumos significantly outperforms RCG in terms of FID on ImageNet-1K and MSCOCO, demonstrating the excellent scalability of Image-to-Image generation on large-scale datasets.

4.2. Downstream Visual Generative Tasks

Subsequently, Lumos-I2I is used as visual generative prior and transferred to various downstream tasks, including Text-to-Image Generation, Novel View Synthesis, and Image-to-Video Generation.

4.2.1 Text-to-Image Generation

We first transfer Lumos-I2I to the text-to-image generation task and compare Lumos-T2I with the state-of-the-art T2I models in terms of the quality of the generated images and alignment assessment.

Implementation Details & Datasets. We construct a dataset of 30 million text-image pairs, all of which are sourced from easily accessible open datasets. This includes 10M and 5M images selected from LAION-5B and COYO-700M, respectively. Additionally, 10M from SAM [26], 4M from JourneyDB [34], and 1M from Imagenet-1K [14] are selected. Following DALLE-3 [2], we incorporate raw captions and detailed captions generated by InternVL [11] into training. Besides FID, we evaluate T2I models on GenEval [18] and DPG-Bench [23]. We use the T5 [39] (specifically 4.3B Flan-T5-XXL) as the text encoder of our T2I model. More details are in *Supplementary Material*.

Quantitative Comparison for FID score. In Table 2, we report the evaluation results on the widely used MS-COCO dataset in terms of FID score and compare Lumos with existing SOTA T2I models. It is worth mentioning that we evaluate the open-sourced model of PixArt-α [7] that is trained on all the datasets proposed in their paper. Thanks to the visual generative prior from the pre-trained image-to-image model, Lumos-T2I model achieves on-par or better performance than existing T2I models, requiring only a

Table 2. **Comparison to the recent text-to-image models on COCO FID-30k.** ' T&I Pairs ' refers to the number of the training text-image pairs. ⋆ refers to the result of ' PixArt-α-256 '. † indicates evaluating with the long captions. Both + and - in the table denote the unknown internal dataset size and training step.

Method	Type	T&I Pairs	Steps	FID-30K↓
DALL·E [40]	Diff	250M	-	27.50
GLIDE [33]	Diff	250M	2500k	12.24
LDM [43]	Diff	400M	-	12.64
DALL·E2 [41]	Diff	650M	2400k	10.39
SDv1.5 [43]	Diff	2000M	1026k	9.62
GigaGAN [24]	GAN	2700M	1350k	9.09
Imagen [44]	Diff	860M	5000k	7.27
RAPHAEL [54]	Diff	5000M+	-	6.61
PixArt-α [7]	Diff	24M	240k	7.32
PixArt-α⋆ [7]	Diff	24M	240k	23.67
PixArt-α⋆† [7]	Diff	24M	240k	21.35
Flux.1 [3]	Diff	-	-	22.76
Kolors [47]	Diff	-	-	23.15
Qihoo-T2I [50]	Diff	50M	100k	15.70
Lumos-T2I	Diff	30M	65k	12.20
Lumos-T2I†	Diff	30M	65k	**6.44**

minimal amount of training data and steps. Moreover, Lumos-T2I surpasses all current text-to-image models in terms of the FID metric with the help of long captions. The experimental results demonstrate the significant benefits of the I2I prior for T2I models.
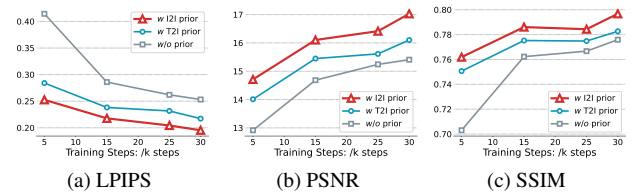

(a) LPIPS (b) PSNR (c) SSIM

Figure 5. **Comparison with different priors for novel view synthesis.** I2I prior shows better metrics consistently from the start of fine-tuning.

Quantitative Comparison for Image-text Alignment Assessment. Beyond the image fidelity evaluation, we evaluate the alignment between the generated images and text prompts through GenEval [18] and DPG-Bench [23]. GenEval [18] is designed for evaluating T2I models from multiple perspectives, including the number of objects, color, attributes, and position. DPG-Bench [23] is a more challenging benchmark, which evaluates the generating performance of the dense prompts and the relation between the entities. Despite using only a small amount of training data and steps, the ability of Lumos-T2I's text-image alignment is on par with or exceeds existing models with a similar parameter scale, as shown in the comparison results of Table. 3. And on the more challenging DPG-Bench [23], Lumos-T2I's performance is still outstanding. This further

Table 3. **Text-image alignment comparison with the state-of-the-art text-to-image models on GenEval [18] and DPG-Bench [23].** Our model is decently trained on publicly available datasets without using any self-collected data. We highlight the **best** and *second best*.

Model	Param	Object		Counting	Colors	Color Attribution	Position	Overall↑	Global	Entity	Attribute	Relation	Other	Average↑
		Single	Two											
								GenEval [18]	DPG-Bench [23]					
LUMINA-Next [55]	2.0B	0.92	0.46	0.48	0.70	0.13	0.09	0.46	82.8	88.7	86.4	80.5	81.8	74.6
SDXL [36]	2.6B	0.98	0.74	0.39	0.85	0.23	0.15	0.55	83.3	82.4	80.9	86.8	80.4	74.7
Playground v2.5 [27]	2.6B	0.98	0.77	0.52	0.84	0.17	0.11	0.56	83.1	82.6	81.2	84.1	83.5	75.5
SD3-8B [17]	8.0B	0.99	0.94	0.72	0.89	0.60	0.33	0.74	-	-	-	-	-	-
FLUX-dev [3]	12.0B	0.99	0.81	0.79	0.74	0.47	0.20	0.67	82.1	89.5	88.7	91.1	89.4	84.0
FLUX-schnell [3]	12.0B	0.99	0.92	0.73	0.78	0.54	0.28	0.71	91.2	91.3	89.7	86.5	87.0	84.8
SDv1.5 [43]	0.9B	0.97	0.38	0.35	0.76	0.06	0.04	0.43	74.6	74.2	75.4	73.5	67.8	63.2
SDv2.1 [43]	0.9B	0.98	0.51	0.44	**0.85**	0.17	0.07	0.50	77.7	78.1	74.9	80.7	80.7	68.1
PixArt-α [7]	0.6B	0.98	0.50	0.44	0.80	0.07	0.08	0.48	81.7	80.1	80.4	81.7	76.5	71.6
PixArt-Σ [8]	0.6B	0.98	0.59	0.50	0.80	0.15	0.10	0.52	**87.5**	87.1	**86.5**	84.0	86.1	79.5
SD3-1B [17]	1.0B	0.97	**0.72**	**0.52**	0.78	**0.34**	**0.16**	**0.58**	-	-	-	-	-	-
Lumos-T2I	0.8B	**0.99**	0.64	**0.52**	0.84	0.30	0.15	0.57	87.4	**87.4**	86.5	**87.2**	**88.1**	**79.9**

proves the feasibility of I2I pre-training followed by T2I alignment training framework in the text-to-image task.

Qualitative Results. We provide more qualitative generated images as shown in Figure 6 to validate the alignment between the generated images and prompts.

4.2.2 Novel View Synthesis

We further validate the advantages of the I2I prior on 3D tasks through the novel view synthesis task.

Implementation Details & Datasets. We transfer Lumos-I2I to novel view synthesis through a subset (750k) of the released Objaverse [13] dataset. Following Zero123 [30], we randomly sample 32 camera extrinsic matrices $\mathcal{M}_1$ which are oriented towards the center of the object, followed by rendering 32 views using a ray tracing engine. Lumos-NVS is evaluated on Google Scanned Objects (GSO) [16], which is a dataset of high-quality scanned household objects. We evaluate the models extensively with three metrics covering different aspects of image similarity: PSNR, SSIM, and LPIPS. More details are in *Supplementary Material*.

I2I *v.s.* **T2I Prior on NVS Task.** To validate the effectiveness of I2I prior on the novel view synthesis task, we compare with training from scratch, I2I prior, and T2I prior on the novel view synthesis task. For fair comparison, we use the SOTA method, *i.e.*, PixArt-α [7] model, as T2I prior, whose structure is similar to our model. As shown in Figure 5, the model transferred from I2I prior is better than from no prior and T2I prior, which demonstrates the superiority of I2I prior over T2I prior.

Comparison with SOTA NVS models. We compare Lumos-NVS with the existing state-of-the-art models in Table 4. We uniformly render photos at 16 angles around the object at two elevations of 0° and 30° for testing. The comparison results show that benefiting from the I2I

prior, Lumos-NVS outperforms the current state-of-the-art models.

Table 4. **Comparison with SOTA models on novel view synthesis task.**

Method	Elevation 0°			Elevation 30°		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Zero123 [30]	17.73	0.8115	0.1763	18.12	0.8099	0.1584
Zero123-XL [12]	17.68	0.7984	0.1988	19.20	0.8210	0.1521
SyncDreamer [31]	-	-	-	18.98	0.8284	0.1535
Lumos-NVS	**19.63**	**0.8439**	**0.1526**	**20.28**	**0.8419**	**0.1317**

Qualitative Results. Figure 7 illustrates a qualitative comparison of novel view generation results for a GSO test object. As can be seen, the novel view images Lumos-NVS generated are more notably consistent and realistic.

4.2.3 Image-to-Video Generation

Following the experimental setup in Stable Video Diffusion [4], we transfer Lumos-I2I to Lumos-I2V for the image-to-video (I2V) generation task. The image-to-video task takes a still-input image as a conditioning input to generate subsequent continuous video frames.

Implementation Details & Datasets. We use the common open-source video datasets Webvid-10M [1] and a subset of Openvid-1M [32], which contains approximately 0.4M videos. For evaluating the performance of I2V models, we report the Fréche Video Distance (FVD) [48], the Kernel Video Distance (KVD) [48], and the Perceptual Input Conformity (PIC) [52] on UCF-101 [46] and MSR-VTT [53]. More details are in *Supplementary Material*.

I2I *v.s.* **T2I Prior on I2V Task.** We compare I2I and T2I prior for the image-to-video generation task through zero-shot evaluating on UCF-101 [46] and MSR-VTT [53] datasets. For the T2I prior, the experiment setting follows
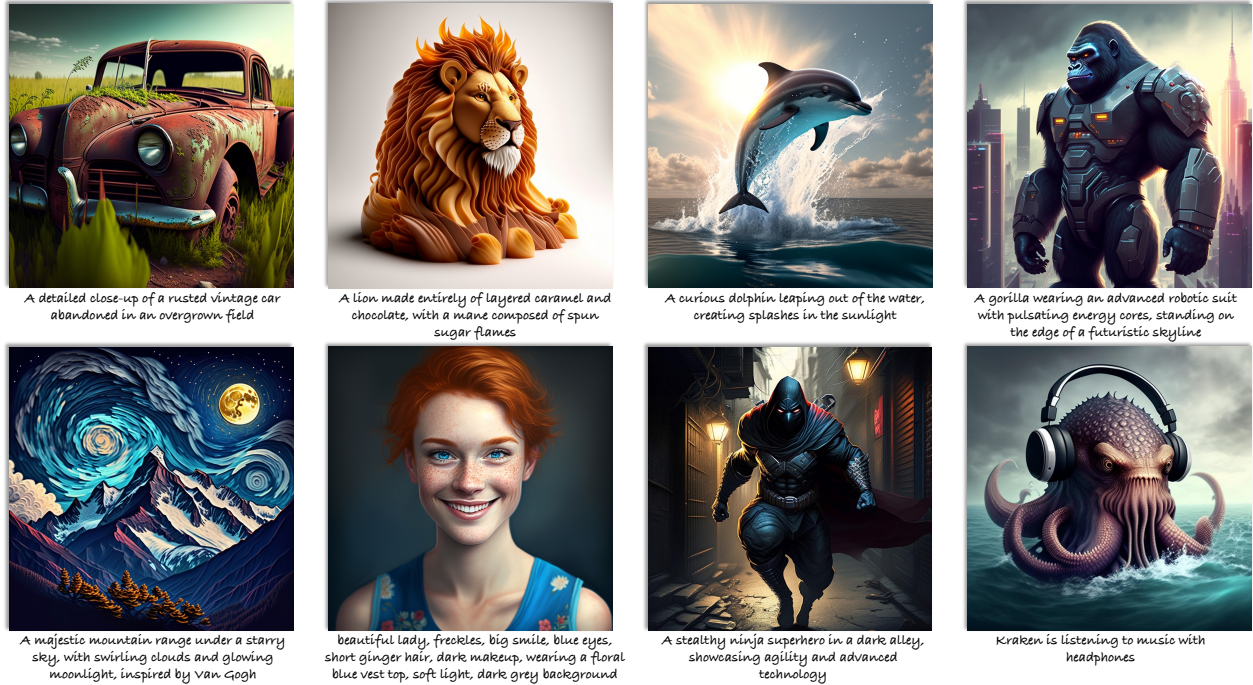
Figure 6. Samples produced by Lumos-T2I exhibit exceptional quality, characterized by a remarkable level of fidelity and precision in adhering to the provided textual prompts.

Figure 7. **Comparison of qualitative results between Lumos-NVS with the SOTA models.**

the ones of the NVS experiments. As shown in Table 5, T2I and I2I prior exhibit a notable initialization for the I2V task. Meanwhile, the comparative results indicate that the I2I prior holds distinct advantages in this task.

Table 5. **Comparison with different priors for I2V.**

Method	UCF-101 [46]			MSR-VTT [53]		
	FVD↓	KVD↓	PIC↑	FVD↓	KVD↓	PIC↑
w/o Prior	2892.09	170.81	0.3795	1490.64	72.97	0.3999
w T2I Prior	532.77	52.53	0.7321	313.50	**19.61**	0.7321
w I2I Prior	**399.59**	**39.17**	**0.7930**	**271.00**	21.16	**0.7869**

4.3. Ablation Study

Finally, we conduct a comprehensive series of experiments to thoroughly analyze and evaluate the effectiveness of
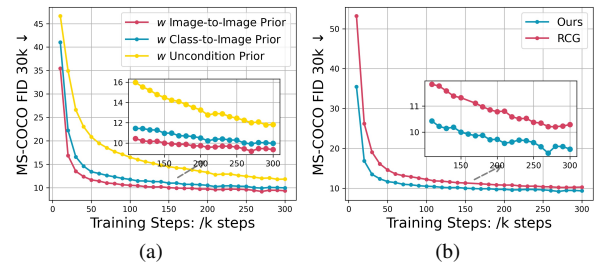
various generative priors and model components.

(a)   (b)

Figure 8. (a) Comparison with different priors of I2I generation on T2I task, (b) Comparison our I2I model with RCG on T2I task.

Ablating different generative priors. We compare Lumos-I2I prior with the class-to-image prior and uncondition prior. We also conduct a comparison with I2I prior from RCG [28], which uses adaLN-Zero block for injecting conditions. To ensure fair experiments, all frameworks are pre-trained on ImageNet-1K [14] and fine-tuned on the same image-text dataset. As shown in Figure 8 (a) and (b), Lumos-I2I provides better prior knowledge for downstream the T2I generation task.

Ablating different features for I2I generative prior. To explore which types of visual features (global semantic or fine-grained local information) are more suitable for the I2I framework, we set up three comparative experiments, *i.e.*, a single global token ([CLS] token), local patch tokens, and all vision tokens (single global token + local patch tokens). As shown in Figure 9 (a) and (b), sufficient fine-grained local information significantly accelerates the convergence speed

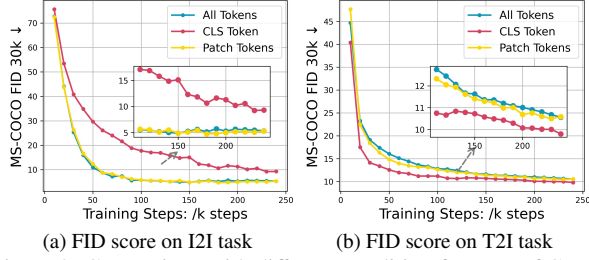(a) FID score on I2I task   (b) FID score on T2I task

Figure 9. Comparison with different condition features of CLIP.

during the I2I training process. However, it is interesting that fine-grained local features lead to a high dependence of the pre-trained model on the conditions, which is not conducive to the rapid transfer of the pre-trained model to downstream tasks.
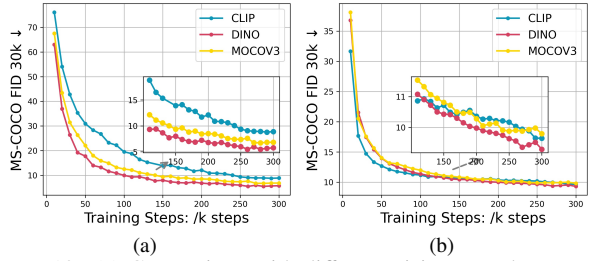

(a)   (b)

Figure 10. (a) Comparison with different vision encoders on I2I task, (b) Comparison with different vision encoders on downstream T2I task.

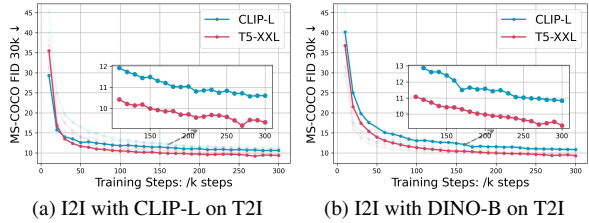(a) I2I with CLIP-L on T2I   (b) I2I with DINO-B on T2I

Figure 11. Comparison between T5 and CLIP text encoder.

Ablating vision encoders for I2I generative prior. We set up experiments to probe the impact of different types of vision encoders (*e.g.*, unimodal self-supervised *v.s.* multimodal alignment) on our I2I frameworks. We adopt three types of vision encoders: CLIP, DINO, and MoCoV3. The comparison results in Figure 10 (a) demonstrate that the self-supervised pre-trained vision encoder (*i.e.*, DINO and MoCoV3) has a significantly faster convergence speed in the image-to-image task. We then validate the transfer performance on the downstream task that relies most on text-image alignment (text-to-image generation). The comparison results in Figure 10 (b) show that the image-text alignment models have significant advantages during the initialization phase of transfer learning. However, this advantage quickly diminishes during the fine-tuning process. Thus, I2I generation framework has a robustness for different vision encoder types.

Ablating different text Encoders for T2I task. We compare the performance of T5 and CLIP text encoder

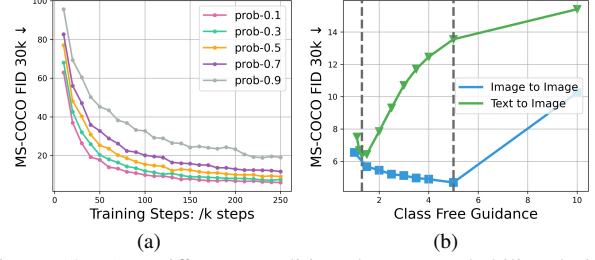(a)   (b)

Figure 12. (a) Different condition dropout probability during training, (b) Different class-free guidance factor during inference.

on different I2I generative priors, as shown in Figure 11 (a) and (b). For the I2I generative priors with unimodal vision encoders (*e.g.*, DINO), T5 outperformed the CLIP text encoder in both convergence speed and performance in terms of FID. For priors with multimodal vision encoders (*e.g.*, CLIP-L), the corresponding CLIP text encoder shows greater advantages in the early transferring stage. However, T5 finally achieves better performance.

Ablating class-free guidance for I2I model. We ablate condition dropout probability during the training process and the class-free guidance factor in the inference phrase in Figure 12 (a) and (b). We investigate that the condition dropout probability during training exhibits similar properties in image-to-image tasks as in text-to-image tasks. However, for the class-free guidance factor, we observe that image generation tasks have a broader applicability and a significant robustness advantage.

5. Conclusion

In this paper, we probe the image-to-image (I2I) generation, where models can learn from in-the-wild images in a self-supervised manner. First, we develop a purely vision-based training framework, `Lumos`, and validate the feasibility and scalability of learning I2I models. Our findings reveal that, as an upstream task of text-to-image (T2I) generation, our I2I model provides a more fundamental visual prior, achieving comparable or superior performance to current T2I models. It is worth mentioning that we only utilize 1/10 of the text-image pairs for fine-tuning of T2I models. Furthermore, we demonstrate the advantages of I2I priors over T2I priors in text-irrelevant visual generative tasks, such as image-to-3D and image-to-video generation. We believe this approach has the potential to liberate image generation from the constraints of image-text pairs, allowing it to learn a more foundational pure-visual prior.

Acknowledgment

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Int. Conf. Comput. Vis.*, pages 1728–1738, 2021. 6

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2, 3, 5

[3] BlackForestlabs AI. Flux. https://blackforestlabs.ai/#get-flux, 2024. 5, 6

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 6

[5] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 4

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, pages 9650–9660, 2021. 2, 3

[7] 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, 3, 5, 6

[8] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*, pages 1597–1607, 2020. 2, 3

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24185–24198, 2024. 5

[12] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Adv. Neural Inform. Process. Syst.*, 36, 2023. 6

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13142–13153, 2023. 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 3, 4, 5, 7

[15] 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. 2

[16] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *International Conference on Robotics and Automation*, pages 2553–2560. IEEE, 2022. 6

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Int. Conf. Mach. Learn.*, 2024. 2, 3, 6

[18] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 5, 6

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.*, 27, 2014. 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020. 2, 3, 5

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16000–16009, 2022. 3

[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. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 5

[23] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 5, 6

[24] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10124–10134, 2023. 5

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *arxiv*, 2013. 3

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 4, 5

[27] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 6

[28] Tianhong Li, Dina Katabi, and Kaiming He. Return of Unconditional Generation: A self-supervised representation generation method. *arXiv preprint arXiv:2312.03701*, 2023. 2, 3, 5, 7

[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 *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 5

[30] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Int. Conf. Comput. Vis.*, pages 9298–9309, 2023. 2, 4, 6

[31] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 6

[32] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 6

[33] 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. 5

[34] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In *arxiv*, 2023. 4, 5

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, pages 4195–4205, 2023. 3, 5

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6

[37] Alec Radford. Improving language understanding by generative pre-training. 2018. 2, 3

[38] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 3

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 5

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Int. Conf. Mach. Learn.*, pages 8821–8831. Pmlr, 2021. 2, 5

[41] 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. 2, 3, 5

[42] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Int. Conf. Mach. Learn.*, 2015. 3

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 2, 3, 5, 6

[44] 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. *Adv. Neural Inform. Process. Syst.*, 35:36479–36494, 2022. 5

[45] 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. *Adv. Neural Inform. Process. Syst.*, 35:25278–25294, 2022. 4

[46] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6, 7

[47] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. 5

[48] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6

[49] A Vaswani. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 2017. 3

[50] Jing Wang, Ao Ma, Jiasong Feng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Qihoo-t2x: An efficiency-focused diffusion transformer via proxy tokens for text-to-any-task. *arXiv preprint arXiv:2409.04005*, 2024. 2, 5

[51] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9653–9663, 2022. 3

[52] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *Eur. Conf. Comput. Vis.*, pages 399–417. Springer, 2025. 6

[53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5288–5296, 2016. 6, 7

[54] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 5

[55] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024. 6