Discriminative Probing and Tuning for Text-to-Image Generation

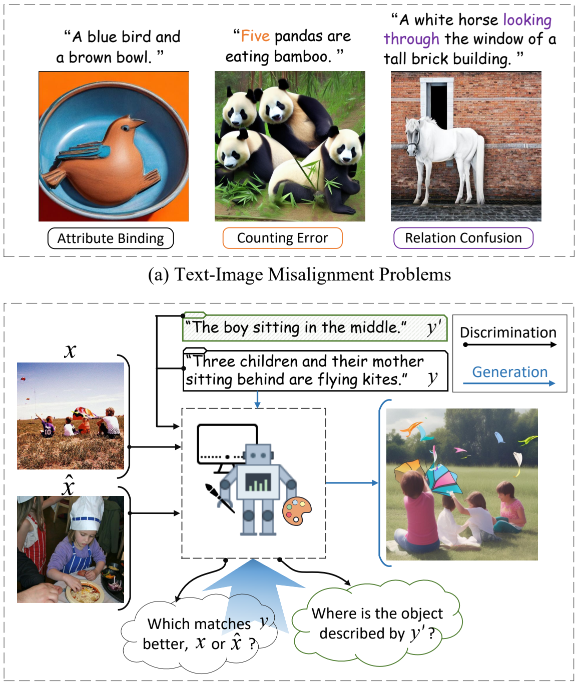
Leigang Qu[1], Wenjie Wang[1]*, Yongqi Li[2], Hanwang Zhang[3,4], Liqiang Nie[5], Tat-Seng Chua[1]
[1]National University of Singapore, [2]Hong Kong Polytechnic University, [3]Nanyang Technological University,
[4]Skywork AI, [5]Harbin Institute of Technology (Shenzhen)
leigangqu@gmail.com, wenjiewang96@gmail.com, liyongqi0@gmail.com
hanwangzhang@ntu.edu.sg, nieliqiang@gmail.com, dcscts@nus.edu.sg

Abstract

Despite advancements in text-to-image generation (T2I), prior methods often face text-image misalignment problems such as relation confusion in generated images. Existing solutions involve cross-attention manipulation for better compositional understanding or integrating large language models for improved layout planning. However, the inherent alignment capabilities of T2I models are still inadequate. By reviewing the link between generative and discriminative modeling, we posit that T2I models' discriminative abilities may reflect their text-image alignment proficiency during generation. In this light, we advocate bolstering the discriminative abilities of T2I models to achieve more precise text-to-image alignment for generation. We present a discriminative adapter built on T2I models to probe their discriminative abilities on two representative tasks and leverage discriminative fine-tuning to improve their text-image alignment. As a bonus of the discriminative adapter, a self-correction mechanism can leverage discriminative gradients to better align generated images to text prompts during inference. Comprehensive evaluations across three benchmark datasets, including both in-distribution and out-of-distribution scenarios, demonstrate our method's superior generation performance. Meanwhile, it achieves state-of-the-art discriminative performance on the two discriminative tasks compared to other generative models. The code is available at https://dpt-t2i.github.io/.

1. Introduction

Text-to-image generation (T2I) aims to synthesize high-quality and semantically-relevant images to a given free-form text prompt. In recent years, the rapid development of diffusion models [22, 47] has ignited the research enthusiasm for content generation, leading to a significant leap in T2I [40, 43, 45]. However, due to the weak compositional reasoning capabilities, current T2I models still suffer from the **Text-Image Misalignment** problem [28], such as attribute binding [15], counting error [37], and relation confusion [37] (see Fig. 1), especially in complicated multi-object generation scenes.


"A blue bird and a brown bowl."
"Five pandas are eating bamboo."
"A white horse looking through the window of a tall brick building."
Attribute Binding
Counting Error
Relation Confusion
x
$\hat{x}$
"The boy sitting in the middle." y'
"Three children and their mother sitting behind are flying kites." y
Discrimination
Generation
Which matches y better, x or $\hat{x}$?
Where is the object described by y' ?

(a) Text-Image Misalignment Problems

(b) Improving Generation with Discrimination

Figure 1. Illustration of the (a) text-image misalignment problem and (b) our motivation by enhancing discriminative abilities of T2I models to promote generative abilities. We list three wrong generation results generated by SD-v2.1 [43] with regard to attribute binding, counting error, and relation confusion in (a).

Two lines of work have made remarkable progress in improving text-image alignment for T2I models. The first line proposes to intervene in cross-modal attention activations guided by linguistic structures [15] or test time optimization [4]. However, they heavily rely on the induc-

*Corresponding author. This work is supported by NExT++ Research Center and National Research Foundation and Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-022).

tive bias for manipulating attention structures, often necessitating expertise in vision-language interaction. This expertise is not easily acquired and lacks flexibility. In contrast, another research line [16, 37] borrows LLM's linguistic comprehension and compositional abilities for layout planning, and then incorporates layout-to-image models (*e.g*., GLIGEN [31]) for controllable generation. Although these methods mitigate misalignment issues like counting error, they heavily rely on intermediate states, *e.g*., bounding boxes, for layout representation. The intermediate states may not adequately capture fine-grained visual attributes, and can also accumulate errors in this two-stage paradigm. Furthermore, the intrinsic compositional reasoning abilities of T2I models are still inadequate.

To tackle these issues, we aim to promote text-image alignment by directly catalyzing the intrinsic compositional reasoning of T2I models, without depending on the inductive bias for attention manipulation or intermediate states. Richard Feynman famously stated, "What I cannot create, I do not understand," underscoring the significance of understanding in the process of creation. This motivates us to consider enhancing the understanding abilities of T2I models to facilitate their text-to-image generation. As illustrated in Fig. 1, T2I models are more likely to generate an image with correct semantics if they can distinguish the alignment difference between the text prompt and the two images with minor semantic variations.

In light of this, we propose to examine the understanding abilities of T2I models by two discriminative tasks. First, we probe the discriminative *global matching* ability[1] of T2I models on Image-text Matching (ITM) [17, 36], a representative task to evaluate fundamental text-image alignment. The second discriminative task inspects the *local grounding* ability of T2I models. One representative task is Referring Expression Comprehension (REC) [58], which examines the fine-grained expression-object alignment within an image. Based on the two tasks, we aim to 1) probe the discriminative abilities of T2I models, especially the compositional semantic alignment, and 2) further improve their discriminative abilities for better text-to-image generation.

Toward this end, we propose a Discriminative Probing and Tuning (**DPT**) paradigm to examine and improve text-image alignment of T2I models in a two-stage process. 1) To probe the discriminative abilities, DPT incorporates a Discriminative Adapter to do the ITM and REC tasks based on the semantic representations [27] of T2I models. For example, DPT may take the feature maps from U-Net of diffusion models [43] as semantic representations. And 2) in the second stage, DPT further improves the text-image alignment by means of parameter-efficient fine-tuning, *e.g*., LoRA [23]. In addition to the adapter, DPT fine-tunes the foundation T2I models to strengthen its intrinsic compositional reasoning abilities for both discriminative and generative tasks. As an extension, we present a self-correction mechanism to guide T2I models for better alignment by gradient-based guidance signals from the discriminative adapter. We conduct extensive experiments on three alignment-oriented text-to-image generation benchmarks and four ITM and REC benchmarks under in-distribution and out-of-distribution settings, validating the effectiveness of DPT in enhancing both generative and discriminative abilities of T2I models. The main contributions of this work are threefold.

- We retrospect the relations between generative and discriminative modeling, and propose a simple yet effective paradigm called DPT to probe and improve the basic discriminative abilities of T2I models for better text-to-image generation.
- We present a discriminative adapter to achieve efficient probing and tuning in DPT. Besides, we extend T2I models with a self-correction mechanism guided by the discriminative adapter for alignment-oriented generation.
- We conduct extensive experiments on three text-to-image generation datasets and four discriminative datasets, significantly enhancing the generative and discriminative abilities of representative T2I models.

2. Related Work

• **Text-to-Image Generation**. Over the past decades, great efforts on Variational Autoencoders [55], Generative Adversarial Networks [54, 59], and Auto-regression Models [9, 39, 57] have been dedicated to generating high-quality images with text conditions. Recently, there has been a flurry of interest in Diffusion Probabilistic Models (DMs) [22, 47] due to their stability and scalability. To further improve the generation quality, large-scale models such as DALL·E 2 [40], Imagen [45], and GLIDE [35], emerged to synthesize photorealistic images. This work mainly focuses on diffusion models and especially takes the open-sourced Stable Diffusion (SD) [43] as the base model.

• **Improving Text-Image Alignment**. Despite the thrilling success, current T2I models still suffer from Text-Image Misalignment issues [1, 8, 18], especially in complex scenes requiring compositional reasoning [34]. Several pioneering efforts were made to introduce guidance to intervene in internal features of SD to stimulate the high-alignment generation. For example, StructureDiffusion [15] parses prompts into tree structures and incorporates them with cross-attention representations to promote compositional generation. Attend-and-Excite [4] manipulates cross-attention units to attend to all textual subject tokens and enhance the activations in attention maps. Despite the notable

[1] Here we inspect the understanding ability of models with discriminative tasks by considering the taxonomy of discriminative and generative learning in Machine Learning.

momentum, they are limited to tackling problems including missing objects and incorrect attributes, and ignore relation enhancement. Another thread of work, *e.g.*, LayoutLLM-T2I [37] and LayoutGPT [16], resorts to two-stage coarse-to-fine frameworks [14, 19, 41], in which they first induce explicit intermediate bounding box-based layout, and then synthesize images. However, such an intermediate layout may not be sufficient to represent complex scenes and they almost abandon the intrinsic reasoning abilities of pre-trained T2I models. In this work, we propose a discriminative tuning paradigm by stimulating discriminative abilities of pre-trained T2I models for high-alignment generation.

• **Generative and Discriminative Modeling**. The thrilling progress of LLMs enables generative models to complete discriminative tasks, which motivates researchers to exploit understanding abilities [33] with foundation visual generative models in Image Classification [5, 7, 29, 56], Segmentation [2, 53, 60], and Image-Text Matching [26]. Besides, DreamLLM [10] unifies generation and discrimination in a multimodal auto-regressive framework and reveals the potential synergy. On the contrary, a recent work [51] discusses the generative AI paradox and showed LLMs may not indeed understand what they have generated. To the best of our knowledge, we are the first to study discriminative tuning to promote alignment in T2I.

3. Method

In this section, we introduce the DPT paradigm to probe and enhance the discriminative abilities of foundation T2I models. As shown in Fig. 2, DPT consists of two stages, *i.e.*, Discrimination Probing and Discrimination Tuning, as well as a self-correction mechanism in Sec. 3.3.

3.1. Stage 1 – Discriminative Probing

In the first stage, we aim to develop a probing method to explore *"How powerful are discriminative abilities of recent T2I models?"*. To this end, we first select representative T2I models and semantic representations, and then consider adapting the T2I models to do discriminative tasks.

• **Stable Diffusion for Discriminative Probing.** Considering SD is open-sourced and one of the most powerful and popular T2I models, we select its different versions (see Sec. 4.2) as representative models to probe the discriminative abilities. To make generative diffusion models semantically focused and efficient, SD [43] performs denoising in a latent low-dimensional space. It includes VAE [25], Text Encoder of CLIP [38], and U-Net [44]. The U-Net serves as a neural backbone for denoising score matching in the latent space, composed of three parts, *i.e.*, down blocks, mid blocks, and up blocks. During training, given a positive image-text pair (x, y), SD first encodes image x with the VAE encoder and adds noise $\epsilon \sim \mathcal{N}(0, 1)$ to obtain the latent $\mathbf{z}_t = h(x, t)$ at timestep t. Thereafter, SD employs U-Net to predict the added noise and optimizes the model parameters by minimizing the L2 loss between the ground-truth noise and the predicted one.

• **Semantic Representations.** It is non-trivial to leverage T2I models such as SD to do discriminative tasks. Fortunately, recent work [27] demonstrates that diffusion models have a meaningful semantic latent space although they were originally designed for denoising [22] or score estimation [48]. Besides, a series of pioneering work [2, 7, 29, 53] shows the validity and even superiority of representations extracted from U-Net of SD to be qualified to discriminative tasks. Inspired by these studies, we consider utilizing semantic representations from the U-Net of SD to do discriminative tasks via a discriminative adapter.

• **Discriminative Adapter**. We propose a lightweight discriminative adapter, which relies on the semantic representations of SD to handle discriminative tasks. Inspired by DETR [3], we implement the discriminative adapter with the Transformer [50] structure, including a Transformer encoder and a Transformer decoder. Besides, we adopt a fixed number of randomly initialized and learnable queries to adapt the framework to specific discriminative tasks.

Concretely, given a noisy latent $\mathbf{z}_t$ at a sampled timestep t and a prompt y, we first feed them into U-Net and extract a 2D feature map $\mathbf{F}_t \in \mathbb{R}^{h \times w \times d}$ from one of the intermediate blocks[2], where h, w, and d denote the height, width, and dimension, respectively. Formally, we extract $\mathbf{F}_t$ via

$$\mathbf{F}_t = \mathrm{UNet}_l(\mathbf{z}_t, \mathrm{CLIP}(y), t), \tag{1}$$

where UNet_l refers to the operation of extracting the feature maps in the l-th block of U-Net. Afterward, we combine $\mathbf{F}_t$ with learnable position embeddings [11] and timestep embeddings [43] of t via additive fusion, and then flatten it into the semantic representation $\tilde{\mathbf{F}}_t \in \mathbb{R}^{hw \times d}$. For simplicity, we will omit the subscript t in the following.

To probe the discriminative abilities, we feed $\tilde{\mathbf{F}}$ into the Transformer encoder $\mathrm{Enc}(\cdot)$, and then perform interaction between the encoder output and some learnable queries $\mathbf{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_N\}$ with $q_i \in \mathbb{R}^d$ in the Transformer decoder $\mathrm{Dec}(\cdot, \cdot)$. The whole process is formulated as

$$\mathbf{Q}^* = f(\tilde{\mathbf{F}}; \mathcal{W}_a, \mathbf{Q}) = \mathrm{Dec}(\mathrm{Enc}(\tilde{\mathbf{F}}), \mathbf{Q}) \tag{2}$$

where $f(\cdot)$ abstracts to the discriminative adapter with parameters $\mathcal{W}_a$ and $\mathbf{Q}$. $\mathcal{W}_a$ includes the parameters in Enc and Dec. The queries $\mathbf{Q}$ serve as a bridge between visual representations and downstream discriminative tasks, which attends the encoded semantic representation $\tilde{\mathbf{F}}_t$ via cross-attention [50] of the decoder for downstream tasks. Thanks to multiple queries in $\mathbf{Q}$, the query representations

[2] We select the medium block by default, and also delve into the influence of different blocks in Sec. 4.3.

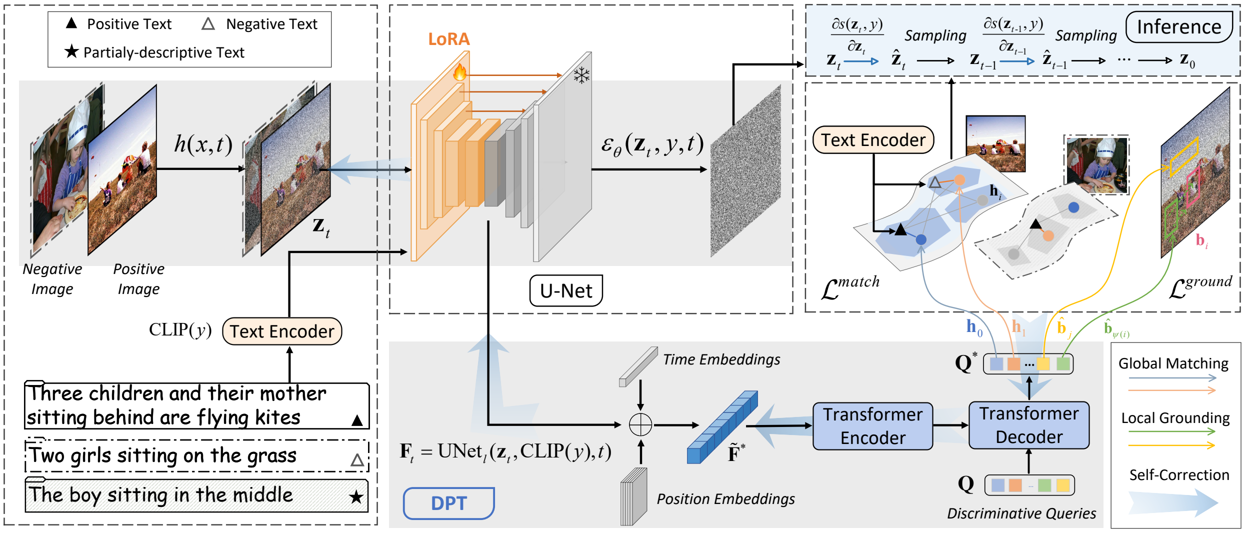

▲ Positive Text
△ Negative Text
★ Partialy-descriptive Text
$h(x,t)$
$\mathbf{z}_t$
Negative Image
Positive Image
CLIP(y)
Text Encoder
Three children and their mother sitting behind are flying kites ▲
Two girls sitting on the grass △
The boy sitting in the middle ★
LoRA
U-Net
$\varepsilon_\theta(\mathbf{z}_t, y, t)$
$\frac{\partial s(\mathbf{z}_t, y)}{\partial \mathbf{z}_t}$
Sampling
$\frac{\partial s(\mathbf{z}_{t-1}, y)}{\partial \mathbf{z}_{t-1}}$
Sampling
Inference
$\mathbf{z}_t \rightarrow \hat{\mathbf{z}}_t \rightarrow \mathbf{z}_{t-1} \rightarrow \hat{\mathbf{z}}_{t-1} \rightarrow \cdots \rightarrow \mathbf{z}_0$
Text Encoder
$\mathbf{h}_i$
$\mathbf{b}_i$
$\mathcal{L}^{match}$
$\mathcal{L}^{ground}$
$\mathbf{h}_0$
$\mathbf{h}_1$
$\hat{\mathbf{b}}_j$
$\hat{\mathbf{b}}_{\sigma(i)}$
$\mathbf{Q}^*$
Time Embeddings
$\mathbf{F}_t = \text{UNet}_l(\mathbf{z}_t, \text{CLIP}(y), t)$
$\tilde{\mathbf{F}}^*$
Transformer Encoder
Transformer Decoder
$\mathbf{Q}$
Position Embeddings
Discriminative Queries
DPT
Global Matching
Local Grounding
Self-Correction

Figure 2. Schematic illustration of the proposed discriminative probing and tuning (DPT) framework. We first extract semantic representations from the frozen SD and then propose a discriminative adapter to conduct discriminative probing to investigate the global matching and local grounding abilities of SD. Afterward, we perform parameter-efficient discriminative tuning by introducing LoRA parameters. During inference, we present the self-correction mechanism to guide the denoising-based text-to-image generation.

$\mathbf{Q}^*$ capture multiple aspects of the semantic representation $\tilde{\mathbf{F}}$. Thereafter, $\mathbf{Q}^*$ can be used to do various downstream tasks, possibly with a classier or regressor.

In the following, we will introduce two probing tasks, *i.e.*, ITM and REC, and train the discriminative adapter on them to investigate the global matching and local grounding abilities of T2I models, respectively.

• **Global Matching**. From the view of discriminative modeling, a model with strong text-image alignment should be able to identify subtle alignment differences between various images and a text prompt. In light of this, We utilize the task of Image-Text Matching [17] to probe the discriminative global matching ability. This task is defined to achieve bidirectional matching or retrieval, including text-to-image ($T \rightarrow I$) and image-to-text ($I \rightarrow T$).

To achieve this, we first collect the first $M(M < N)$ query representations $\{\mathbf{q}_1^*, ..., \mathbf{q}_M^*\}$ from $\mathbf{Q}^*$, and then project each of them into a matching space with the same dimension as CLIP and obtain $\mathbf{h}_i = g(\mathbf{q}_i^*; \mathcal{W}_m)$. Intuitively, different query representations may capture different aspects to understand the same image. Inspired by this, we calculate the cross-modal semantic similarities between x and y by comparing the CLIP textual embedding of y and the most matched projected query representations via $s(y, \mathbf{z}) = \max_{i \in \{1,...,M\}} \cos(\text{CLIP}(y), \mathbf{h}_i)$. Based on pairwise similarities, we optimize the discriminative adapter $f(\cdot; \mathcal{W}_a, \mathbf{Q})$ and the projection layer $g(\cdot; \mathcal{W}_m)$ using contrastive learning loss $\mathcal{L}^{match} = \mathcal{L}^{T \rightarrow I} + \mathcal{L}^{I \rightarrow T}$. The first term optimizes the model to distinguish the correct image matched with a given text from all samples in a batch, *i.e.*,

$$\mathcal{L}^{T \rightarrow I} = -\log \frac{\exp(s(\mathbf{z}, y)/\tau)}{\sum_{j=1}^{B} \exp s((\mathbf{z}_j, y)/\tau)}, \tag{3}$$

where B denotes the min-batch size, and τ is a learnable temperature factor. Similarly, the opposite direction from image to text is computed by

$$\mathcal{L}^{I \rightarrow T} = -\log \frac{\exp(s(\mathbf{z}, y)/\tau)}{\sum_{j=1}^{B} \exp s((\mathbf{z}, y_j)/\tau)}. \tag{4}$$

With $\mathcal{L}^{match}$ as the optimization objective, the discriminative adapter and the projection layers are enforced to discover discriminative information from the semantic representations for matching, implying the global matching ability of a T2I model.

• **Local Grounding**. Local grounding requires a model to recognize the referred object from others in an image given a partially descriptive text. We adapt SD to the REC [58] task to evaluate its discriminative local grounding ability.

Formally, given a textual expression y' referring to a specific object with index i in an image x, REC aims to predict the coordinate and the size, *i.e.*, the bounding box $\mathbf{b}_i$, of the ground-truth object. To achieve it, we share the same discriminative adapter and employ the other $(N - M)$ learnable queries as object prior queries and obtain the corresponding query representations from the transformer decoder as $\{\mathbf{q}_j^*\}_{j \in \{M+1,...,N\}}$. We then project each $\mathbf{q}_j^*$ into three spaces separately by three different project layers $g(\cdot)$: 1) the grounding space to get the probability of predicting the correct object, *i.e.*, $p_j = g(\mathbf{q}_j^*; \mathcal{W}_p) \in \mathbb{R}^1$; 2) the box space to estimate the bounding box parameters, *i.e.*, $\hat{\mathbf{b}}_j = g(\mathbf{q}_j^*; \mathcal{W}_b) \in \mathbb{R}^4$; and 3) the semantic space to bridge the semantic gap between queries and the text, *i.e.*, $\mathbf{o}_j = g(\mathbf{q}_j^*; \mathcal{W}_s) \in \mathbb{R}^d$.

After projection, we perform maximum matching to discover the most matched query with index $\psi(i)$. The cost used for matching includes using grounding probability, L1, and GIoU [42] losses between the prediction and the ground-truth box as costs. It is formulated as

$$\psi(i) = \underset{j \in \{M+1,\ldots,N\}}{\arg\min} \; -p_j + \mathrm{L1}(\hat{\mathbf{b}}_j, \mathbf{b}_i) + \mathrm{GIoU}(\hat{\mathbf{b}}_j, \mathbf{b}_i) \quad (5)$$

Besides, we adopt a text-to-object contrastive loss to further drive the model to distinguish the positive object from others at the semantic level:

$$\mathcal{L}^{T \to O} = -\log \frac{\exp(\cos(\mathbf{o}_{\psi(i)}, \mathrm{CLIP}(y'))/\tau)}{\sum_{j=1}^{K_x} \exp(\cos(\mathbf{o}_j, \mathrm{CLIP}(y'))/\tau)}, \quad (6)$$

We combine all the losses and obtain the grounding loss as

$$\begin{aligned}\mathcal{L}^{ground} = &-\lambda_0 p_{\psi(i)} + \lambda_1 \mathrm{L1}(\hat{\mathbf{b}}_{\psi(i)}, \mathbf{b}_i) \\ &+ \lambda_2 \mathrm{GIoU}(\hat{\mathbf{b}}_{\psi(i)}, \mathbf{b}_i) + \lambda_3 \mathcal{L}^{T \to O},\end{aligned} \quad (7)$$

where $\{\lambda_k\}_{k \in \{0,1,2,3\}}$ serve as trade-off factors.

Finally, we optimize the parameters of the whole model, including $\mathbf{Q}$ and $\{\mathcal{W}_i\}, i \in \{a, p, b, s\}$, with the following loss function on two tasks:

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t}(\mathcal{L}_t^{match} + \mathcal{L}_t^{ground}) \quad (8)$$

The probing process includes training and inference on the two discriminative tasks. During training, we freeze all parameters of SD, and adopt its semantic representations for matching and grounding by optimizing the discriminative adapter and several projection layers. During inference, we obtain the testing performance on the two discriminative tasks, which reflects the discriminative abilities of SD.

3.2. Stage 2 – Discriminative Tuning

In the second stage, we propose to improve the generative abilities, especially text-image alignment, by optimizing T2I models in a discriminative tuning manner. Most prior work [2, 53] only views SD as a fixed feature extractor for segmentation tasks due to its fine-grained semantic representation power but overlooks the potential back-feeding of discrimination to generation. Besides, though a recent study [26, 52] fine-tunes the SD model using discriminative objectives, it only pays attention to specific downstream tasks (*e.g.*, ITM) and ignores the effect of tuning on generation. The advancement of discrimination may sacrifice the original generative power. In this stage, we mainly focus on enhancing generation, but also investigate the superior limit of discrimination under the premise of priority generation. It may shed new light on giving full play to the versatility of visual generative foundation models. In this vein, we strive to explain *"How can we enhance text-image alignment for T2I models by discriminative tuning?"*

In the previous stage, we freeze SD and probe how informative intermediate activations are in global matching and local grounding. Here, we conduct parameter-efficient fine-tuning using LoRA [23] by injecting trainable layers over cross-attention layers and freezing the parameters of the pre-trained SD. We use the same discriminative objective functions as stage 1 to tune the LoRA, discriminative adapter, and task-specific projection layers. Due to the participation of LoRA, we can flexibly manipulate the intermediate activation of T2I models.

3.3. Self-Correction

Equipping the T2I model with the discriminative adapter enables the whole model to execute discriminative tasks. As a bonus of using the discriminative adapter, we propose a self-correction mechanism to guide high-alignment generation during inference. Formally, we update the latent $\mathbf{z}_t$ aiming to enhance the semantic similarity between $\mathbf{z}_t$ and the prompt y through gradients:

$$\hat{\mathbf{z}}_t = \mathbf{z}_t + \eta \frac{\partial s(\mathbf{z}_t, y)}{\partial \mathbf{z}_t}, \quad (9)$$

where the guidance factor η control the guidance strength. $\frac{\partial s(\mathbf{z}_t, y)}{\partial \mathbf{z}_t}$ represents the gradients from the discriminative adapter to the latent $\mathbf{z}_t$. Afterward, we predict the noise by feeding $\hat{\mathbf{z}}_t$ into U-Net and then obtain $\mathbf{z}_{t-1}$ for generation.

4. Experiments

We conduct extensive experiments to evaluate the generative and discriminative performance of DPT, justify its effectiveness, and conduct an in-depth analysis.

4.1. Experimental Settings

- **Benchmarks**. During training, we adopt the training set of MSCOCO [32] for ITM and three commonly used datasets [58], *i.e.*, RefCOCO, RefCOCO+, and RefCOCOg for REC. To evaluate the text-image alignment, we utilize five benchmarks: COCO-NSS1K [37], CC-500 [15], ABC-6K [15], TIFA [24], and T2I-CompBench [13]. According to the distribution differences of textual prompts between the training set and the test sets, we adopt three settings, *i.e.*, In-Distribution (ID) and Out-of-Distribution (OOD) [49] on COCO-NSS1K and CC-500, respectively, and Mixed Distribution (MD) on ABC-6K, TIFA, and T2I-CompBench. More details can be found in Appendix **??**.
- **Evaluation Metrics**. Following the existing baselines [4, 15, 37], we adopt CLIP score [20] and BLIP score[3] [30]

[3] OpenCLIP (ViT-H-14) [6] and BLIP-2 (pretrain) are used to compute text-image similarities as CLIP and BLIP scores, respectively. We will

Table 1. Performance comparison for *text-to-image generation* on COCO-NSS1K, CC-500, and ABC-6K. ID, OOD, and MD refer to in-distribution, out-of-distribution, and mixed-distribution settings, respectively. According to the version of Stable Diffusion, we split methods into two groups, top and down for v1.4 and v2.1, respectively. SC denotes self-correction.

Method	COCO-NSS1K (ID)					CC-500 (OOD)					ABC-6K (MD)			
	CLIP	BLIP-M	BLIP-C	IS	FID	CLIP	BLIP-M	BLIP-C	GLIP	IS	CLIP	BLIP-M	BLIP-C	IS
Stable Diffusion-v1.4 [CVPR22] [43]	33.27	67.96	39.48	31.32	54.77	34.82	70.95	40.36	31.17	14.28	35.33	72.03	40.82	34.47
LayoutLLM-T2I [ACMMM23] [37]	32.42	67.42	39.46	25.57	59.26	-	-	-	-	-	-	-	-	-
StructureDiffusion [ICLR23] [15]	-	-	-	-	-	33.71	66.71	39.54	31.39	14.14	34.95	69.55	40.69	34.97
HN-DiffusionITM [NeurIPS23] [26]	33.26	70.06	**40.14**	31.53	53.26	34.15	68.77	40.30	31.54	13.99	35.02	72.28	41.12	34.83
DPT (Ours)	**33.85**	**71.84**	40.11	31.65	54.96	**35.97**	**76.74**	**41.15**	**37.07**	13.56	**35.88**	**75.88**	**41.26**	34.46
Stable Diffusion-v2.1 [CVPR22] [43]	34.96	73.32	40.22	30.40	55.35	39.24	85.45	43.36	52.09	11.53	37.53	81.98	41.77	33.31
Attend-and- Excite [TOG23] [4]	34.95	74.68	40.32	30.27	55.16	39.43	90.03	44.08	**53.29**	11.82	37.59	82.64	41.83	32.94
HN-DiffusionITM [NeurIPS23] [26]	35.14	75.64	40.77	30.34	52.73	38.81	85.76	43.22	48.95	12.11	37.58	82.33	42.07	34.14
DPT (Ours)	**35.83**	78.58	41.14	30.83	55.55	40.23	90.72	44.55	**53.29**	11.59	38.39	**86.19**	**42.36**	32.97
DPT + SC (Ours)	35.75	**79.15**	**41.14**	30.50	54.89	**40.25**	**91.33**	**44.69**	**53.29**	11.89	**38.41**	85.63	42.34	33.56

Table 2. Performance comparison for *text-to-image generation* on TIFA [24] and T2I-CompBench [13]. According to the version of Stable Diffusion, we split methods into two groups, top and down for v1.4 and v2.1, respectively. SC denotes self-correction.

	TIFA	T2I-CompBench					
		Color	Shape	Text.	Sp.	Non-Sp.	Comp.
SD-v1.4 [43]	79.15	36.82	35.94	42.16	10.64	30.45	28.18
HN-DiffusionITM [26]	79.02	36.71	35.48	39.84	11.22	**30.91**	28.05
VPGen [12]	77.33	32.12	32.36	35.85	19.08	30.07	24.39
LayoutGPT [16]	79.31	33.86	36.35	44.07	**35.06**	30.30	26.36
DPT (Ours)	82.04	48.84	38.93	50.10	14.63	30.83	30.05
DPT + SC (Ours)	**82.40**	**51.51**	**39.61**	**49.38**	15.45	30.84	**30.29**
SD-v2.1 [43]	81.35	48.21	40.49	46.83	16.94	30.63	29.96
Attend-and-Excite [4]	81.98	53.72	43.41	48.53	16.30	30.64	30.38
HN-DiffusionITM [26]	82.02	46.45	40.09	49.35	15.01	**30.99**	30.35
DPT (Ours)	84.49	60.59	48.18	**58.24**	20.78	30.95	32.44
DPT + SC (Ours)	**84.63**	**62.59**	**48.44**	57.60	**21.04**	30.76	**32.52**

including BLIP-ITM and BLIP-ITC, and GLIP score [15] based on object detection to evaluate text-image alignment, and IS [46] and FID [21] as quality evaluation metrics. As for TIFA and T2I-CompBench, we follow the recommended VQA accuracy or specifically curated protocols.

4.2. Performance Comparison

• **Text-to-Image Generation.** As shown in Tab. 1 and Tab. 2, we have the following observations and discussions: 1) Compared with the base foundation models, *i.e.*, SD [43], the proposed DPT manages to improve the text-image alignment remarkably, which illustrates that enhancing discriminative abilities could benefit the generative semantic alignment for T2I models. 2) DPT achieves superior performance on CC-500 and ABC-6K under the OOD setting, showing its powerful generalization to other prompt distributions. It also reveals its capability to resist the risk of overfitting when tuning T2I models with discriminative tasks. 3) The consistent improvement on both SD-v1.4 and SD-v2.1 demonstrates that the proposed DPT may be parallel with the generative pre-training based on score matching, reflecting the possibility of activating the intrinsic reasoning abilities of T2I models using DPT. And 4) in all, the proposed method achieves the best generation performance consistently on text-image alignment across comprehensive benchmarks, distribution settings, and evaluation protocols. Besides, the improvement in alignment does not result in a loss of image quality per IS and FID. These results confirm the effectiveness of the proposed paradigm DPT.

• **Discriminative Matching and Grounding**. In Sec. 3.1, we incorporate a discriminative adapter on top of T2I models and probe and improve its understanding abilities based on ITM and REC. In an empirical sense, we carry out experiments by training the adapter in the first stage and introducing the LoRA for tuning in the second stage using ITM and REC data, and then evaluate the matching and grounding performance. We show experimental results of baselines including discriminative and generative models under the zero-shot and fine-tuning settings in Tab. 3. See Appendix **??** for more details on the implementation and settings. From this table, we observe that our method could outperform the existing state-of-the-art generative methods, such as Diffusion Classifier [29] and DiffusionITM [26], by large margins on ITM and REC tasks. Even it could achieve competitive performance in the first probing stage or when selected with a priority generation in the second stage. These results show that the generative representations extracted from the intermediate layers of U-Net convey meaningful semantics, verifying that T2I models have basic discriminative matching and grounding abilities. Besides, it also indicates that such abilities could be further improved by the discriminative tuning introduced in Sec. 3.2.

4.3. In-depth Analysis

To verify the effectiveness of each component in DPT, including discriminative tuning on Global Matching (GM) and Local Grounding (LG) in the 2nd stage, and the Self-

adopt Image-Text Matching (ITM) and Image-Text Contrastive (ITC) as BLIP scores in the following.

Table 3. Performance comparison for *image-text matching* and *referring expression comprehension* to evaluate global matching and local grounding abilities, respectively. Datasets include MSCOCO-HN for ITM, and RefCOCO, RefCOCO+, and RefCOCOg for REC. All the methods are grouped into three parts, in which the upper, middle, and lower groups correspond to zero-shot discriminative, zero-shot generative, and fine-tuning generative methods, respectively. All the generative models are based on Stable Diffusion-v2.1.

Method	MSCOCO-HN			RefCOCO			RefCOCO+			RefCOCOg	
	I-to-T	T-to-I	Overall	val	testA	testB	val	testA	testB	val	test
Random Chance	25.00	25.00	25.00	16.53	13.51	19.20	16.29	13.57	19.60	18.12	19.10
CLIP (ViT-B-32) [ICML21] [38]	47.63	42.82	45.23	44.79	46.12	42.61	49.60	51.07	46.04	58.31	58.42
OpenCLIP (ViT-B-32) [CVPR23] [6]	49.07	47.45	48.26	43.22	43.15	44.65	48.21	48.60	50.64	60.32	60.84
Diffusion Classifier § [ICCV23] [29]	34.59	24.12	29.36	6.23	2.14	12.11	6.07	2.11	12.29	8.68	8.45
DiffusionITM § [NeurIPS23] [26]	34.59	29.83	32.21	28.88	30.16	29.01	29.97	31.17	30.25	38.07	38.91
Local Dinoising	-	-	-	23.83	21.20	24.85	24.07	21.31	25.45	28.66	28.59
Diffusion Classifier †§ [ICCV23] [29]	37.72	24.03	30.88	6.11	2.10	10.91	6.04	2.13	11.48	8.05	7.54
DiffusionITM †§ [NeurIPS23] [26]	37.72	29.88	33.80	34.09	32.70	35.29	35.86	35.42	38.23	49.67	49.05
HN-DiffusionITM †§ [NeurIPS23] [26]	37.55	30.37	33.96	31.43	28.50	35.47	33.47	30.16	37.47	47.98	48.20
Local Dinoising †	-	-	-	23.70	21.55	24.81	24.01	21.52	25.32	28.53	28.77
DPT (Stage1, Ours)	42.29	34.75	38.52	48.79	53.28	43.06	42.56	47.69	36.14	46.56	45.75
DPT (Ours)	42.07	34.97	38.52	52.73	57.84	46.73	45.34	50.12	38.41	48.61	47.45
DPT* (Ours)	**43.12**	**35.25**	**39.18**	**63.45**	**66.70**	**57.90**	**51.56**	**56.81**	**42.73**	**54.96**	**54.80**

†: fine-tuning with the denoising objective;
§: cropping an image into blocks and then matching them with the referring text for REC;
∗: model selection with a priority discriminative task, *i.e.*, ITM or REC

Correction (SC) during inference, we conduct several analytic experiments on COCO-NSS1K and CC-500 under ID and OOD settings. The results are summarized in Tab. 4.

• **Effectiveness of Discriminative Tuning**. From the compared results in Tab. 4 between different variants, we observe that the two tuning objectives, *i.e.*, GM and LG, could consistently promote the alignment performance for T2I according to CLIP and BLIP scores. It verifies the validity of discriminative tuning on ITM and REC tasks. Compared with GM, LG achieves more remarkable improvement over semantic and object detection metrics. It may be attributed to the enhanced grounding ability brought by the prediction of local concepts based on partial descriptions. Furthermore, combining the two objectives to conduct multi-task learning may contribute to a slight improvement in BLIP scores under the OOD setting, but other metrics are slightly compromised. This phenomenon indicates that some contradictions may exist during model optimization, reflecting that unifying multiple tasks is still challenging.

• **Effectiveness of Self-Correction**. In Sec. 3.3, we propose to recycle the discriminative adapter in the inference phase by guiding iterative denoising. Comparing the 3rd and 4th variants in Table 4, we can see that the self-correction scheme could consistently improve the alignment for T2I, attesting to its effectiveness.

• **Impact of Probed U-Net Block**. Due to the hierarchical structure of the U-Net in SD, we could extract multi-level feature maps from its different blocks. Prior work [52] has shown that different blocks may have different discriminative powers in image classification. To further investigate the matching and grounding abilities empowered by various blocks and the trade-off between discrimination and generation, we probe consecutive seven blocks of the U-Net shown in Fig. 2 from left to right and then tune the whole model based on the probed block. The generative and discriminative results are shown in Fig. 3. It can be observed that the T2I performance gets continuously improved with the probed block shifting from bottom to up. The reason may be

Table 4. Ablation study for the influence of two objectives of discriminative tuning including Global Matching (GM) and Local Ground (LG) in the 2nd stage, and the Self-Correction (SC) during inference on alignment-oriented text-to-image generation. The COCO-NSS1K and CC-500 datasets are used to evaluate in-distribution (ID) and out-of-distribution (OOD) generation. All experiments are based on Stable Diffusion-v2.1.

Index	GM	LG	SC	COCO-NSS1K (ID)			CC-500 (OOD)			
				CLIP	BLIP-M	BLIP-C	CLIP	BLIP-M	BLIP-C	GLIP
0				34.96	73.32	40.22	39.24	85.45	43.36	52.09
1	✓			35.14	74.83	40.45	39.28	86.23	43.36	49.55
2		✓		**35.94**	**79.19**	41.11	**40.31**	90.63	44.31	**57.03**
3	✓	✓		<u>35.83</u>	78.58	<u>41.14</u>	40.23	90.72	44.55	53.29
4	✓	✓	✓	35.75	<u>79.15</u>	**41.14**	<u>40.25</u>	**91.33**	**44.69**	<u>53.29</u>

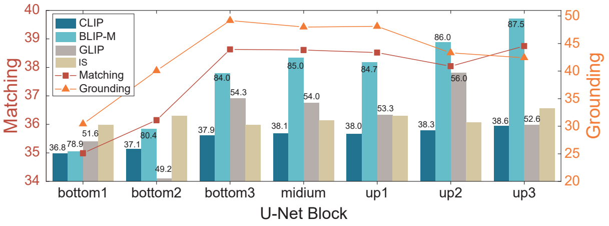
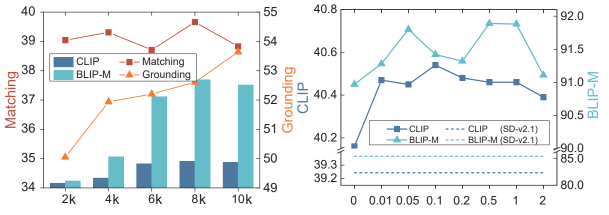


Figure 3. Generative and discriminative results by probing different layers of U-Net in SD-v2.1 and adapting to ITM and REC. We report average CLIP and BLIP-M scores over COCO-NSS1K and CC-500, overall matching performance on MSCOCO-HN, and average grounding performance over all test sets of RefCOCO, RefCOCO+, and RefCOCOg. We conduct model selection based on T2I performance on the validation set of COCO-NSS1K.

that more LoRA parameters would be introduced and more layers would be tuned during back-propagation. In contrast, the discriminative performance regardless of matching and grounding starts to increase and then deteriorates. It may be attributed to two points: 1) the feature maps from those blocks (*e.g.*, up2 and up3) close to final outputs, *i.e.*, predicted noises, are less semantic; 2) the feature sequences flattened from these feature maps may be too long, making it difficult for the discriminative adapter to probe.



(a) Tuning Step

(b) Guidance Factor η

Figure 4. Impact of (a) the variation of generation and discrimination performance with the progress of tuning and (b) the self-correction strength on the performance of T2I on CC-500.

- **Impact of Tuning Step**. To further delve into the durative impact of discriminative tuning on two aspects of performance, we show the dynamics of the performance with the increment of the tuning step in the 2nd stage in Fig. 4a. We can see that the generative performance gets better with tuning and seems to reach the saturation point at the 8k step. In contrast, there is still potential for grounding performance to get higher while the matching performance seems to remain stable in the tuning stage.
- **Impact of Self-Correction Factor**. As shown in Fig. 4b, we study the influence of the guidance factor η in Eqn. (9) on the alignment performance of T2I. The results demonstrate that the proposed self-correction mechanism could alleviate the text-image misalignment issue with a proper range of guidance factor, *i.e.*, $(0.05, 1)$.

4.4. Qualitative Results

To intuitively show the alignment improvement achieved by DPT and SC, we illustrate generated examples with prompts from COCO-NSS1K for object appearance, counting, relation, and compositional reasoning evaluation, as shown in Fig. 5. These cases demonstrate the effectiveness of incorporating discriminative probing and tuning into T2I models.

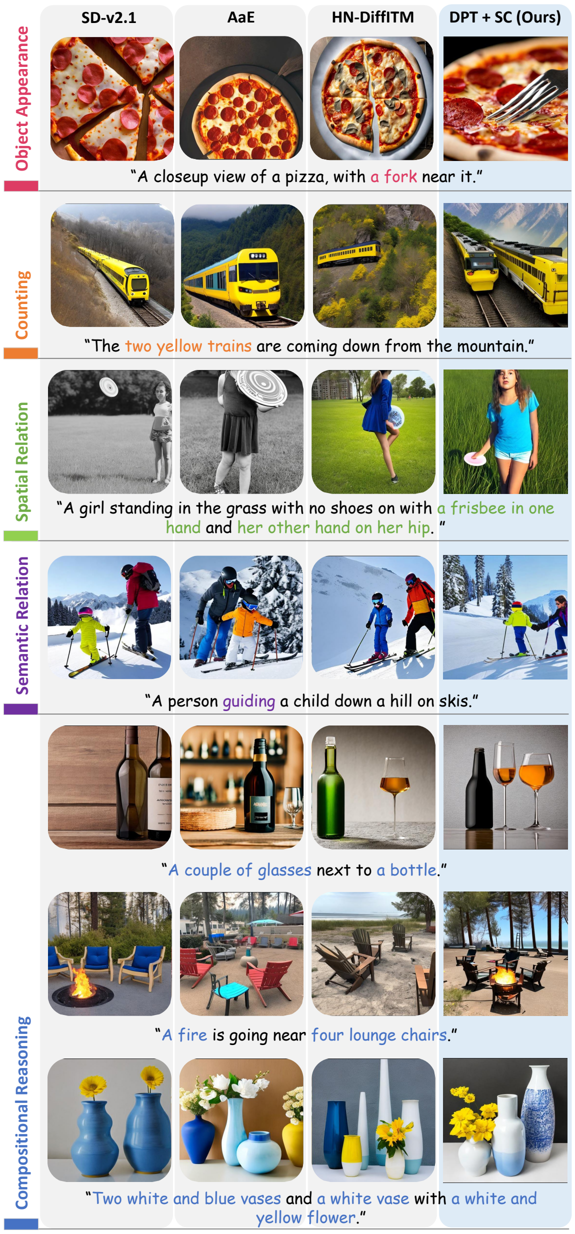

SD-v2.1
AaE
HN-DiffITM
DPT + SC (Ours)
Object Appearance
"A closeup view of a pizza, with a fork near it."
Counting
"The two yellow trains are coming down from the mountain."
Spatial Relation
"A girl standing in the grass with no shoes on with a frisbee in one hand and her other hand on her hip. "
Semantic Relation
"A person guiding a child down a hill on skis."
"A couple of glasses next to a bottle."
"A fire is going near four lounge chairs."
"Two white and blue vases and a white vase with a white and yellow flower."
Compositional Reasoning

Figure 5. Qualitative results on COCO-NSS1K. We compare DPT with SD-v2.1 and two baselines including Attend-and-Excite (AaE) [4] and HN-DiffusionITM (HN-DiffITM) [26] regarding object appearance, counting, spatial relation, semantic relation, and compositional reasoning. Categories and the corresponding keywords in prompts are highlighted with different colors.

5. Conclusion and Future Work

In this work, we tackled the text-image misalignment issue for text-to-image generative models. Toward this end, we retrospected the relations between generative and discriminative modeling and presented a two-stage method named DPT. It introduces a discriminative adapter for probing basic discriminative abilities in the first stage and performs discriminative fine-tuning in the second stage. DPT exhibited effectiveness and generalization across five T2I datasets and four ITM and REC datasets.

In the future, we plan to explore the effect of discriminative probing and tuning to more generative models using more conception and understanding tasks. Besides, it is interesting to discuss more complicated relations between discriminative and generative modeling such as trade-offs and mutual promotion across different tasks.

References

[1] 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. 2

[2] Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S Ryoo. Peekaboo: Text to image diffusion models are zero-shot segmentors. *arXiv preprint arXiv:2211.13224*, 2022. 3, 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[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, 6, 8

[5] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. *arXiv preprint arXiv:2305.15241*, 2023. 3

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 5, 7

[7] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 3

[8] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022. 2

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[10] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Cho et al. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 6

[13] Huang et al. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023. 5, 6

[14] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 579–587, 2023. 3

[15] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1, 2, 5, 6

[16] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023. 2, 3, 6

[17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2, 4

[18] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 2

[19] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1004–1014, 2021. 3

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[21] 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

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5

[24] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 5, 6

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[26] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 5, 6, 7, 8

[27] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2, 3

[28] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1

[29] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 3, 6, 7

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5

[31] 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. 2

[32] 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

[33] Xinyu Lin, Wenjie Wang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. A multi-facet paradigm to bridge large language model and recommendation. *arXiv preprint arXiv:2310.06491*, 2023. 3

[34] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 2

[35] 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

[36] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113, 2021. 2

[37] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023. 1, 2, 3, 5, 6

[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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 7

[39] 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

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

[41] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[42] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[43] 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, 3, 6

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[45] 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. 1, 2

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[47] 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, 2

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[49] Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 15–23, 2022. 5

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[51] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. The generative ai paradox:" what it can create, it may not understand". *arXiv preprint arXiv:2311.00059*, 2023. 3

[52] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. *arXiv preprint arXiv:2303.09769*, 2023. 5, 7

[53] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3, 5

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[55] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 776–791. Springer, 2016. 2

[56] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 3

[57] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[58] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 4, 5

[59] 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

[60] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 3