This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

IMG: Calibrating Diffusion Models via Implicit Multimodal Guidance

Jiayi Guo[1,2*] Chuanhao Yan[1,2*] Xingqian Xu[1] Yulin Wang[2] Kai Wang[1]
Gao Huang[2†] Humphrey Shi[1†]
[1]SHI Labs @ Georgia Tech [2]Tsinghua University

https://github.com/SHI-Labs/IMG-Multimodal-Diffusion-Alignment

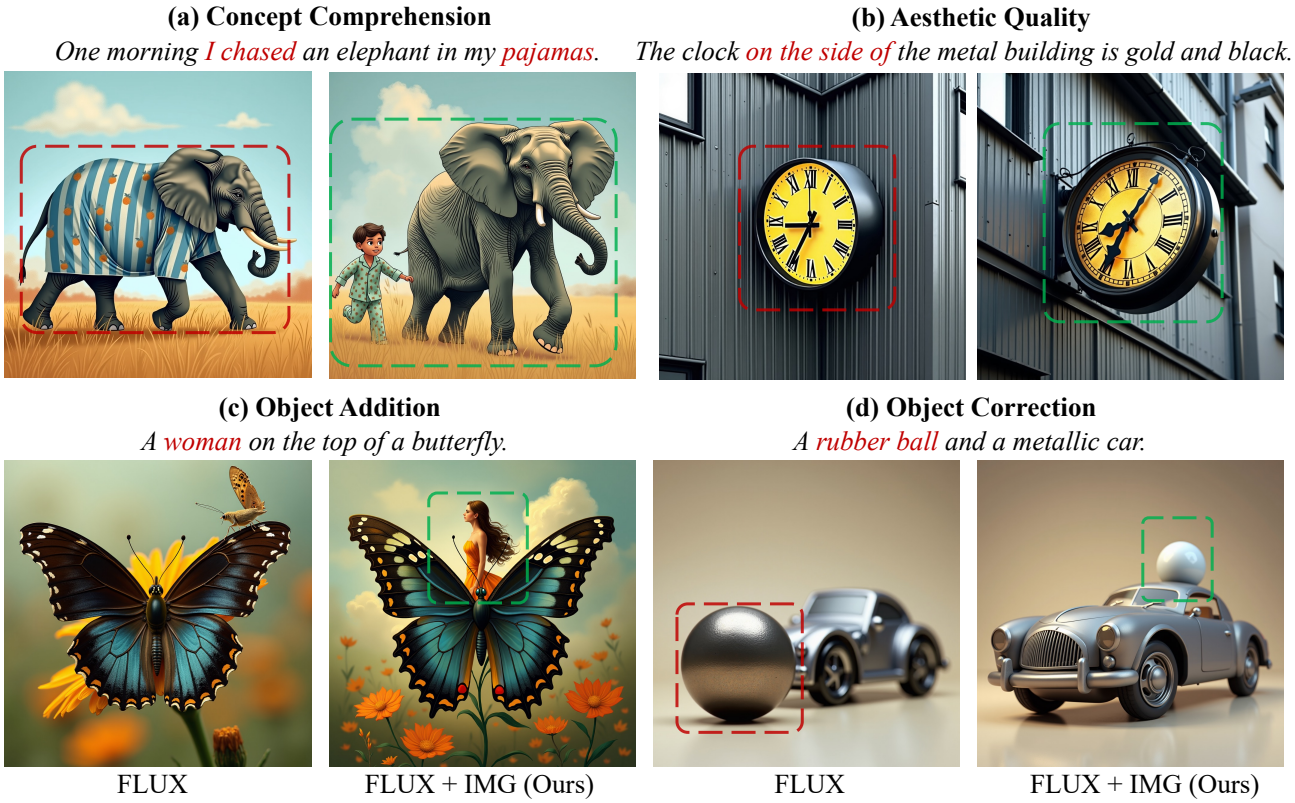


Figure 1. **The multimodal misalignment issue.** Even the latest state-of-the-art diffusion model, FLUX.1 [dev] (FLUX) [35], may overlook or misinterpret concepts in prompts. Assisted with our proposed Implicit Multimodal Guidance (IMG) framework, the prompt-image misalignment issues are significantly mitigated in various aspects such as concept comprehension, aesthetic quality, object addition, and correction. In each case, both images are generated with the same random seed for fair comparison.

Abstract

*Ensuring precise multimodal alignment between diffusion-generated images and input prompts has been a long-standing challenge. Earlier works finetune diffusion weight using high-quality preference data, which tends to be limited and difficult to scale up. Recent editing-based methods further refine local regions of generated images but may compromise overall image quality. In this work, we propose **I**mplicit **M**ultimodal **G**uidance (**IMG**), a novel re-generation-based multimodal alignment framework that requires no extra data or editing operations. Specifically, given a generated image and its prompt, IMG a) utilizes a multimodal large language model (MLLM) to identify misalignments; b) introduces an Implicit Aligner that manipulates diffusion conditioning features to reduce misalignments and enable re-generation; and c) formulates the re-alignment goal into a trainable objective, namely Iteratively Updated Preference Objective. Extensive qualitative and quantitative evaluations on SDXL, SDXL-DPO, and FLUX show that IMG outperforms existing alignment methods.*

*Equal contribution. †Corresponding authors.

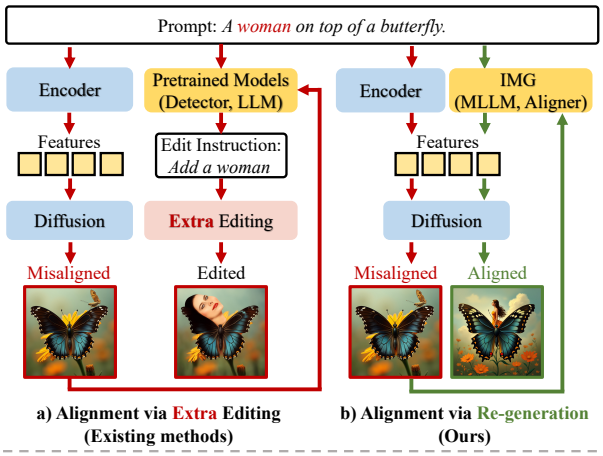
Furthermore, IMG acts as a flexible plug-and-play adapter, seamlessly enhancing prior finetuning-based alignment methods. Our code is available at https://github.com/SHI-Labs/IMG-Multimodal-Diffusion-Alignment.

1. Introduction

Recently, diffusion models have become powerful text-to-image (T2I) generation tools capable of producing diverse and realistic images. However, most methods still face input-output misalignment challenges based on human inspection. More specifically, these models may overlook or misinterpret a few aspects of prompts, thus creating undesired visual results with misinterpretations. Fig. 1 shows evidence that even the latest state-of-the-art diffusion model, FLUX [35], generates images that require further refinement in terms of prompt awareness and adherence.

Due to the aforementioned challenges, exploring methods that improve prompt-image alignment has become an emerging and critical area of research. Early works in this field primarily employ the *preference-based weight finetuning*. Methods such as [11, 52] finetune diffusion models on high-quality prompt-image pairs. Later studies [5, 10, 13, 62, 75] apply reinforcement learning from human feedback (RLHF), exploring rewarding algorithms and datasets that are proven to be effective for large language models (LLMs) [1, 12]. However, these methods are constrained by the limited availability of high-quality finetuning data, which is difficult to scale up further. In contrast, the recent *LLM-driven image editing* method [65] eliminates the need for fine-tuning model weights. It combines an open-set detector [46] with an LLM [1] to identify misalignments in generated images and produces language instructions for image editing (see Fig. 2a). Leveraging powerful LLMs for verification and reflection on generation results is a promising research direction [18, 48, 49]. However, the current editing pipeline primarily focuses on improving alignment in locally edited regions, and we empirically find that the overall image quality often fails to maintain the pre-editing level. Additionally, its detector sometimes omits critical misalignments, resulting in inaccurate editing instructions from the LLM.

In this work, we explore enhancing alignment performance without additional data required by finetuning-based methods or extra editing operations in editing-based methods. To this end, we propose the *Implicit Multimodal Guidance* (IMG), a novel diffusion alignment framework that improves prompt adherence through a simple yet effective re-generation process. Specifically, IMG involves a multimodal large language model (MLLM) and a newly introduced Implicit Aligner. The MLLM first detects misalignment between the prompts and the images generated by diffusion models. Subsequently, the Implicit Aligner takes as input both features of MLLM and misaligned images, producing better-aligned diffusion conditioning features to reduce misalignment and enable re-generation. To train the Implicit Aligner, we introduce an Iteratively Updated Preference Objective, which combines direct preference optimization [55] and self-play finetuning [8].



	a) Existing methods	b) Ours
Pipeline Complexity	High (Extra Editing)	Low (Re-generation)
Reuse Diffusion Process	✗	✓
Performance (HPS score↑)	28.14	29.48

Figure 2. **Comparison between our Implicit Multimodal Guidance (IMG) and existing editing-based alignment methods.** a) Existing methods require additional editing operations that improve alignment in local regions but may compromise overall image quality. b) In contrast, IMG employs a re-generation-based alignment framework by manipulating diffusion conditioning features, ensuring pipeline simplicity and high-quality outputs.

To conclude, IMG distinguishes itself from prior approaches by: a) serving as a flexible plug-and-play adapter that seamlessly integrates with both base diffusion models and their finetuned versions once trained, b) leveraging the native diffusion generation process to maintain pipeline simplicity and high-quality outputs, and c) eliminating the need for additional editing operations, making it a more straightforward and efficient solution. Built on these principles, we conduct extensive qualitative and quantitative evaluations across popular T2I models, including SDXL [52], SDXL-DPO [62], and FLUX [35], as well as both finetuning-based and editing-based alignment methods. Results consistently demonstrate that IMG effectively re-aligns image outputs to improved versions and outperforms existing alignment methods.

2. Related Work

Diffusion model alignment is a research area focusing on improving the prompt adherence of diffusion models in terms of human perception. Early efforts [11, 52, 59, 66, 69]

directly finetune diffusion models on high-quality datasets to produce visually appealing results. Later studies conduct rewarding algorithms based on reinforcement learning [5, 10, 13, 29, 68] and preference learning [62, 71, 75] using datasets labeled with human preferences [19, 34]. For example, DDPO [5] utilizes black-box reward functions to optimize diffusion models within specific prompt sets. Diffusion-DPO [62] re-formulates the direct preference optimization [55] into a differentiable diffusion objective via the evidence lower bound. Inspired by the rapid advances of large language models (LLMs) [1, 12, 40, 56], recent works explore integrating LLMs inside the diffusion system to enhance prompt comprehension and representation [14, 22, 39, 65, 72]. For instance, LMD [39] performs layout-grounded image generation using captioned bounding boxes generated from prompts by LLMs [1]. SLD [65] introduces an LLM-driven image editing framework that utilizes open-set detectors [46] and LLMs [1] to identify misalignments and edit initial generation results. Additionally, recent studies explore the use of MLLMs to aid in image editing [15, 17, 28, 45, 50] and customization [25, 38, 51, 61, 76, 79]. However, these methods primarily focus on executing human-provided instructions, rather than automatically correcting misalignment.

Diffusion adapters [20, 21, 27, 47, 57, 64, 70, 73, 77, 78] aim to extend the capability of diffusion models by incorporating additional input conditions beyond text prompts. Pioneering works like ControlNet [77] and T2I-Adapter [47] introduce structure-conditioned adapters to enable controllable image synthesis. On top of these baselines, works such as Composer and Uni-ControlNet [27, 53, 78] propose a unified adapter capable of handling various structural conditions. Recently, diffusion adapters enabling visual-encoding [9, 30, 32, 42, 70, 73] have continuously gained attention in variation, editing and video tasks. For instance, Prompt-free Diffusion [70] introduces the SeeCoder as an image context encoder, replacing the original text encoder to support image prompts. IP-Adapter [73] integrates CLIP image encoder [54] with decoupled cross-attention layers to create an effective image-prompt adapter.

3. Methodology

In this section, we first go through preliminaries of diffusion models with text and image prompts [59, 73] in Sec. 3.1. Then, we provide an in-depth explanation of our method: Implicit Multimodal Guidance (IMG) in Secs. 3.2 to 3.4.

3.1. Preliminaries

Diffusion models [35, 59] are a class of generative methods involving forward and reverse processes. The forward process is usually known as a gradual procedure, transforming the data point $\boldsymbol{x}_0$ into Gaussian noise $\boldsymbol{x}_T$ with T steps. For example, a canonical formulation for $\boldsymbol{x}_t$ is defined as such:

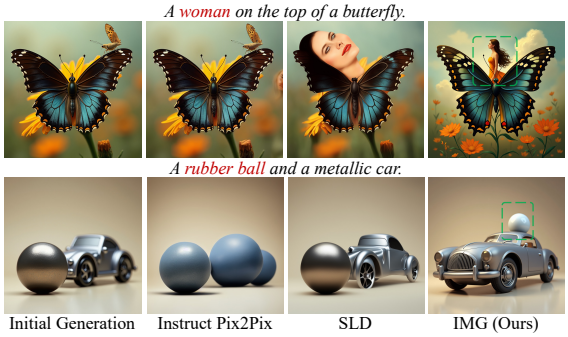


Figure 3. **Comparison with editing-based methods.** We evaluate the performance of Instruct Pix2Pix [6] and SLD [65] with IMG. For Instruct Pix2Pix, the instructions are "add a woman" and "make the ball a rubber ball", generated by our finetuned MLLM.

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \tag{1}$$

where $t \sim \mathcal{U}(0, T)$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I})$ is random Gaussian noise, α_t and σ_t are predefined functions of t. The reverse process iteratively transforms $\boldsymbol{x}_T$ into $\boldsymbol{x}_0$, assessing intermediate $\boldsymbol{x}_t$ through a well-trained deep neural net $\boldsymbol{\epsilon}_\theta$ [23, 41, 44, 60].

Considering the text-to-image generation task, given an image $\boldsymbol{x}_0$ and text condition $\boldsymbol{c}_T$, the training objective of $\boldsymbol{\epsilon}_\theta$ is formulated as:

$$L_{\text{diff}} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}, \boldsymbol{c}_T, t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, t) \right\|_2^2. \tag{2}$$

Diffusion with image prompt [70, 73, 77, 78] has recently gained popularity in the community, in which diffusion models reconstruct images similar to the input image, usually through a frozen T2I model attaching an additional image prompt (IP) encoder that takes visual inputs. Thus, the training objective is extended to include an additional image condition $\boldsymbol{c}_I$, which is the encoded features of $\boldsymbol{x}_0$:

$$L_{\text{diff}} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}, \boldsymbol{c}_T, \boldsymbol{c}_I, t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_T, \boldsymbol{c}_I, t) \right\|_2^2. \tag{3}$$

During inference, users input both $\boldsymbol{c}_I$ and, optionally, $\boldsymbol{c}_T$ to create or enhance a content-consistent variant of $\boldsymbol{x}_0$. IP encoders [67, 73] are widely available for diffusion models such as SD series [59] and FLUX [35].

3.2. MLLM-driven Misalignment Analysis

Diffusion models sometimes misinterpret or overlook parts of a prompt when generating images. To address this issue, our first step is to identify these misalignments via Multimodel Large Language Models (MLLMs) that are capable of visual question answering [3].

While MLLMs [40] can describe image contents, there has been limited effort to customize them as misalignment detectors. In IMG, we address this by introducing a customized MLLM, which we finetune using instruction-based image data [6], enabling the MLLM to analyze and respond to potential misalignments.

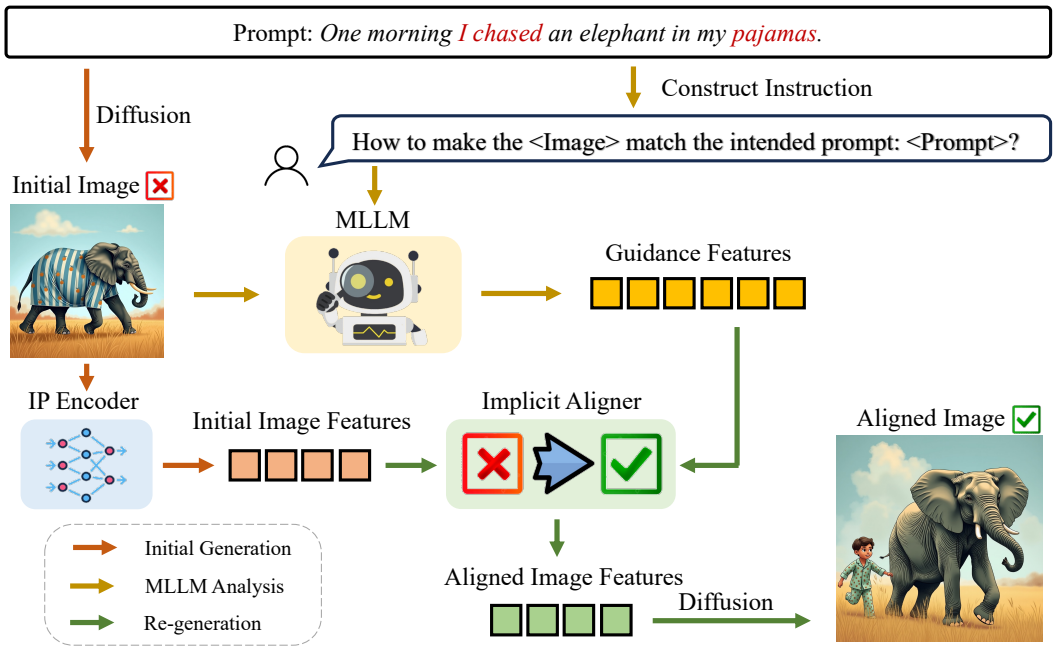


Figure 4. **Overview of the Implicit Multimodal Guidance (IMG) framework.** Given an initial image that exhibits misalignments with its prompt, IMG begins by conducting an MLLM-driven misalignment analysis. Following this, IMG utilizes an Implicit Aligner to translate the initial image features into better-aligned features according to the MLLM's guidance. Finally, these aligned image features are incorporated as new conditions to re-generate images with improved prompt-image alignment.

Specifically, the instruction-based dataset contains original images I_0, edited images I_1, prompts for both images T_0 & T_1, and edit instructions T_E. We utilize the data by taking out I_0, T_1 and T_E as our training triplets. While I_0 and T_1 are fed into the MLLM as inputs, we request the model to describe alignment through question prompts such as "How to make the <Original Image> match the intended prompt: <Edited Prompt>?", then supervise model outputs against T_E. Through experiments in Sec. 4.5, we demonstrate that our specialized finetuning makes MLLM a much more reliable model for misalignment detection. For more details regarding MLLM finetuning, please see our supplementary.

3.3. Implicit Multimodal Guidance

Our next step in IMG is to improve the generated images by removing the previously detected misalignment. A straightforward baseline is to edit these images through existing editing methods like [6, 65]. In Fig. 3, we provide a quick demo using Instruct Pix2Pix [6] and SLD [65], both showing unsatisfactory results. For Instruct Pix2Pix, we use the responses generated by our finetuned MLLM as instructions. However, the result shows that the editing takes place incorrectly. For SLD, although it successfully generates a woman in the first case, the overall aesthetic quality is degraded, and SLD fails to deal with the second case.

The unsuccessful try-out on existing editing methods motivates us to tackle the challenge from a new angle: set up an image re-generation process conditioned on *implicit features*, highlighted as *Implicit Multimodal Guidance* (IMG). A complete diagram of our IMG framework is shown in Fig. 4, in which a) we generate an initial image that may exhibit misalignments via a diffusion model; b) we detect potential misalignments via MLLM using questions same as those in the finetuning process, (*e.g.*, "How to make the <Image> match the intended prompt: <Prompt>?"); and c) we re-generate the image conditioned on aligned features produced by the newly introduced Implicit Aligner.

Generally, our Implicit Aligner functions as an adapter network for the original diffusion model, facilitating the realignment of the initial image to better adhere to its prompt. Instead of explicitly editing the initial image [6, 65], our design implicitly refines the diffusion conditioning features. Structured as a stack of cross-attention layers, the Implicit Aligner takes MLLM hidden states, namely guidance features, as one input and initial image features, embedded via an image prompt (IP) encoder, as the other. The guidance features promote modifications on initial features, producing better-aligned features that are trained to fit the embedded features of a well-aligned image. According to the properties of image prompt diffusion described in Sec. 3.1, the diffusion model can then re-generate a more faithfully aligned image from these re-aligned features.

In the next section, we will introduce Iteratively Updated Preference Objective, the training objective of our Implicit Aligner that leverages human preference datasets [34].

3.4. Iteratively Updated Preference Objective

As aforementioned, we leverage the same open-source human preference datasets [34] adopted in finetuning-based alignment methods [62] to train our Implicit Aligner. These datasets consist of triplets $\{\boldsymbol{c}_T, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l\}$, where $\boldsymbol{x}_0^w$ and $\boldsymbol{x}_0^l$ represent the human-preferred (winning) and non-preferred (losing) images based on the prompt $\boldsymbol{c}_T$ (*i.e.*, $\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l$). Let $\boldsymbol{c}_I^l$ and $\boldsymbol{c}_I^w$ denote the image prompt features of $\boldsymbol{x}_0^l$ and $\boldsymbol{x}_0^w$. The direction from $\boldsymbol{c}_I^l$ to $\boldsymbol{c}_I^w$ indicates a meaningful alignment-improving trajectory, conditioned on $\boldsymbol{c}_T$ [63, 73]. Such trajectory suggests a basic objective for our Implicit Aligner, in which the network f_θ fits $\boldsymbol{c}_I^w$ based on input conditions $\boldsymbol{c} = (\boldsymbol{h}, \boldsymbol{c}_I^l)$, where $\boldsymbol{h}$ represents the MLLM guidance features and $\boldsymbol{c}_I^l$ represents the image features, formulated as the following:

$$L_{\text{base}} = \mathbb{E}_{\boldsymbol{c}, \boldsymbol{c}_I^w} ||\boldsymbol{c}_I^w - f_\theta(\boldsymbol{c})||_2^2. \tag{4}$$

Besides the basics, we also draw inspiration from direct preference optimization (DPO) [55] and self-play finetuning (SPIN) [8] for enhanced feature alignment. In our case, we use DPO to finetune f_θ from a reference aligner f_{ref} (a copy of f_θ from an earlier training iteration), encouraging the network to output $f_\theta(\boldsymbol{c})$ close to the preferred features $\boldsymbol{c}_I^w$ while keeping distance from the non-preferred features $\boldsymbol{c}_I^l$. SPIN also forces $f_\theta(\boldsymbol{c})$ to be close to $\boldsymbol{c}_I^w$ while additionally forces $f_\theta(\boldsymbol{c})$ to be more preferred than $\boldsymbol{c}_I^{\text{ref}} = f_{\text{ref}}(\boldsymbol{c})$. Adapting both DPO [55, 62] and SPIN [8, 75], we derive our objective as the following:

$$L_{\text{pref}} = \mathbb{E}_{\boldsymbol{c}, \boldsymbol{c}_I^w, \boldsymbol{c}_I^l} \left[\ell \left(\underbrace{\log \frac{p_\theta(\boldsymbol{c}_I^w|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{c}_I^w|\mathbf{c})} - \log \frac{p_\theta(\boldsymbol{c}_I^l|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{c}_I^l|\mathbf{c})}}_{\text{DPO}} + \underbrace{\log \frac{p_\theta(\boldsymbol{c}_I^w|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{c}_I^w|\mathbf{c})} - \log \frac{p_\theta(\boldsymbol{c}_I^{\text{ref}}|\mathbf{c})}{p_{\text{ref}}(\boldsymbol{c}_I^{\text{ref}}|\mathbf{c})}}_{\text{SPIN}} \right) \right], \tag{5}$$

where ℓ is a monotonically decreasing convex function, usually implemented as $\ell(x) := \log(1 + e^{-x})$ [62, 75]. Following [33, 37, 58], we define p_θ and p_{ref} as Gaussian distributions with means predicted by f_θ and f_{ref} and a constant variance σ, *i.e.*, $p(\boldsymbol{c}_I|\mathbf{c}) = \mathcal{N}(\boldsymbol{c}_I; f(\mathbf{c}), \sigma)$. With some transformations, the objective in Eq. 5 can be simplified to:

$$\begin{aligned} L_{\text{pref}} = \mathbb{E}_{\boldsymbol{c}, \boldsymbol{c}_I^w, \boldsymbol{c}_I^l} [\ell(-[2(||\boldsymbol{c}_I^w - f_\theta(\boldsymbol{c})||_2^2 - ||\boldsymbol{c}_I^w - f_{\text{ref}}(\boldsymbol{c})||_2^2) \\ - (||\boldsymbol{c}_I^l - f_\theta(\boldsymbol{c})||_2^2 - ||\boldsymbol{c}_I^l - f_{\text{ref}}(\boldsymbol{c})||_2^2) \\ - ||f_{\text{ref}}(\boldsymbol{c}) - f_\theta(\boldsymbol{c})||_2^2])]. \end{aligned} \tag{6}$$

The detailed derivation of L_{pref} can be found in the supplementary. The final training objective is then a combination of L_{base} and L_{pref} with a ratio parameter λ:

$$L = L_{\text{base}} + \lambda L_{\text{pref}}. \tag{7}$$

Unlike prior works such as DPO [62] and SPIN [75], which require a fixed reference net f_{ref} with frozen weights, our Implicit Aligner can be trained from scratch and the reference net can be runtime updated. Specifically, we first randomly initialize f_{ref} and later iteratively copy f_θ to f_{ref} whenever f_θ outperforms f_{ref}. In practice, we execute the substitution when $f_\theta(\boldsymbol{c})$ is closer to $\boldsymbol{c}_I^w$ than $f_{\text{ref}}(\boldsymbol{c})$ for k consecutive iterations. Thereafter, we name our objective function L the Iteratively Updated Preference Objective.

4. Experiments

4.1. Experimental Setup

Baselines and models. Our experiments are conducted on two diffusion models: SDXL [52] and FLUX.1 [dev] (FLUX) [35]. We also compare IMG against Diffusion-DPO (SDXL-DPO) [62], the top-performing finetuning-based alignment method, and SLD [65], the leading editing-based alignment method. We further compare IMG with leading compositional generation methods, ELLA [24] and CoMat [31]. For MLLM, we finetune LLaVA 1.5-13b [40] on the Instruct-Pix2Pix dataset [6] for 1 epoch following [40] and extract the last hidden layer features for guidance. We use the IP-Adapter [67, 73], trained on SDXL and FLUX to enable image prompts and extract image features.

Implementation details. IMG's Implicit Aligner is trained on the Pick-a-Pic training set [34] (the same as SDXL-DPO [62]) for 100K iterations with 8 A100 GPUs and a batch size of 8. The training data contains 851K preferred and non-preferred image pairs under specific prompts. We use the AdamW [43] optimizer with a constant learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The ratio parameter λ in Eq. 7 is set to 1. The reference model updating step k in Sec. 3.4 is set to 10. We determine the optimal hyperparameters via the average Pick Score [34] across generated images on 500 Pick-a-Pic test set prompts. Pick Score is a caption-aware alignment scoring model trained on Pick-a-Pic. For evaluation, we report the Human Preference Scores (HPS) on the Human Preference Datasets (HPD) test set [66], which includes 3,400 prompts across 5 categories and on the Parpi-Prompts [74], a diverse prompt dataset of 1,632 prompts ranging from brief concepts to complex sentences. We also report results on the T2I-CompBench [26], which contains 1800 test prompts to validate compositional image generation capabilities. We apply 50 sampling steps for SDXL and SDXL-DPO, and 30 for FLUX. More implementation details are provided in the supplementary.

Model	Human Preference Datasets (HPD)						Parti-Prompts	User Study
	Anime	Concept-Art	Drawbench	Painting	Photo	Average		
SDXL [52]	28.59	27.69	28.04	27.74	28.10	28.03	27.71	77.6%
SDXL + IMG (Ours)	**29.14 (85.5%)**	**28.31 (92.8%)**	**28.48 (81.5%)**	**28.33 (91.5%)**	**28.49 (80.3%)**	**28.56 (87.2%)**	**28.13 (79.2%)**	
SDXL-DPO [62]	28.98	28.12	28.40	28.18	28.29	28.39	27.99	75.5%
SDXL-DPO + IMG (Ours)	**29.42 (84.8%)**	**28.55 (86.1%)**	**28.68 (77.0%)**	**28.58 (83.8%)**	**28.66 (77.1%)**	**28.79 (82.6%)**	**28.34 (76.0%)**	
FLUX [35]	29.77	28.98	29.44	28.90	29.14	29.21	29.23	66.9%
FLUX + IMG (Ours)	**30.06 (67.0%)**	**29.17 (62.0%)**	**29.80 (64.0%)**	**29.20 (69.5%)**	**29.41 (64.8%)**	**29.48 (65.7%)**	**29.61 (71.7%)**	

Table 1. **Quantitative comparison with base models and finetuning-based alignment methods on HPD and Parti-Prompts**. We report the average HPS scores and the win rates of IMG over competing methods. Higher HPS scores indicate better preference alignment. Additionally, we conduct user studies to assess the real human preference rates of IMG.

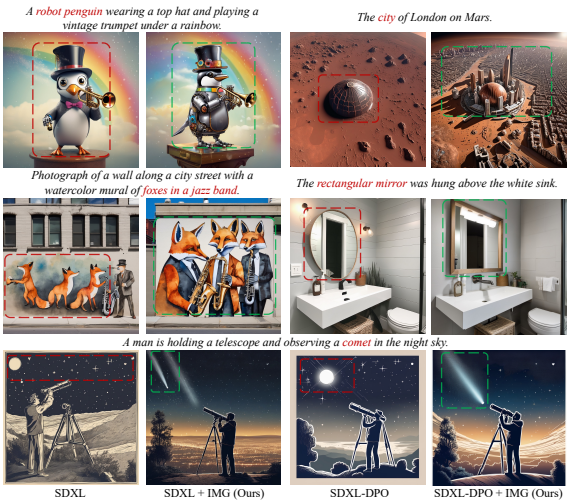


Figure 5. **Qualitative comparison with base models and finetuning-based alignment methods.** The first two rows show that IMG addresses various misalignment types across different prompts, while the last row shows that IMG resolves misalignment issues that challenge both models.

4.2. Comparison with Base Models and Finetuning-based Methods

Qualitative comparison. In addition to comparing with FLUX [35] in Fig. 1, we perform further quantitative evaluations by integrating IMG with the widely used SDXL [52] model and its finetuned variant, SDXL-DPO [62]. The Implicit Aligner of IMG is trained using SDXL's image prompt encoder [73] and seamlessly shared with SDXL-DPO. As shown in Fig. 5, the results highlight a) IMG's effectiveness in addressing diverse alignment issues from various aspects, such as concept comprehension (*e.g.*, a penguin with a robotic body and foxes playing musical instruments), aesthetic quality (*e.g.*, a well-constructed city on Mars), object addition (*e.g.*, a comet in the sky) and object correction (*e.g.*, a rectangular mirror), and b) IMG's flexibility to operate with SDXL-DPO without additional training. The last row also showcases cases where both SDXL and SDXL-DPO struggle, whereas IMG demonstrates clear improvements in alignment across both models.

Quantitative comparison. In Tab. 1, we conduct quantitative evaluations on the HPD [66] and Parti-Prompts [74] datasets, which contain thousands of prompts across diverse categories and complexities. We evaluate images generated by the base models and those re-generated by IMG, using HPS scores [66], and report the win rates of IMG over the base models. The results demonstrate that IMG serves as a general framework that consistently enhances alignment across different base models. Notably, IMG shares the same training data as SDXL-DPO, yet when integrated with SDXL, it outperforms SDXL-DPO and achieves an average win rate of 84.6% over SDXL. Furthermore, when integrated with SDXL-DPO, IMG achieves even higher performance, with an average win rate of 80.5% over SDXL-DPO. This suggests that IMG not only enhances alignment more effectively but also complements finetuning-based methods boosting their performance without requiring additional data. When combined with the state-of-the-art FLUX model, IMG also achieves a strong alignment performance with an average win rate of 67.6%. In addition, we conduct user studies where 33 evaluators were asked to do an A-B test on 30 random image pairs generated by each base model and IMG with the same prompt. Each unique pair was assessed by 3 evaluators, and only fully consistent votes were used to compute the final win rates. Results show that approximately 70% of users prefer our re-aligned images over the originals. In Tab. 2, beyond general prompt sets, we further assess IMG on the compositional image generation benchmark, T2I-CompBench [47]. Without specialized training on compositional prompts like CoMat [31], IMG operates in a zero-shot manner while achieving leading performance with both SDXL and FLUX.

Model	Color↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	Complex↑
SDXL [52]	0.5552	0.4878	0.5213	0.1915	0.3125	0.3403
SDXL-ELLA [24]	0.7260	0.5634	**0.6686**	0.2214	0.3069	-
SDXL-CoMat [31]	**0.7827**	0.5329	0.6468	0.2428	0.3187	0.3680
SDXL-DPO [62]	0.6957	0.5430	0.6428	0.2415	0.3160	0.3561
SDXL-DPO + IMG (Ours)	0.7294	**0.5811**	0.6548	**0.2484**	**0.3199**	**0.3874**
FLUX [35]	0.7852	0.5633	0.6942	0.2792	0.3129	0.3891
FLUX + IMG (Ours)	**0.8060**	**0.5773**	**0.7008**	**0.2826**	**0.3139**	**0.3983**

Table 2. **Quantitative comparison with base models and finetuning-based alignment methods on T2I-CompBench.** The best results are in **bold** and the second-best results are underlined.

4.3. Comparison with Editing-based Methods.

Qualitative comparison. In Fig. 6, we compare IMG with the leading editing-based alignment method, SLD [65]. In the first case, although SLD recognizes the missing ten-

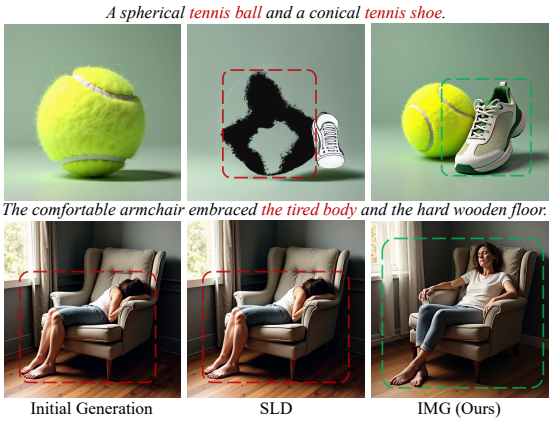


Figure 6. **Qualitative comparison with editing-based methods.** IMG surpasses SLD in visual quality and image comprehension.



Figure 7. **Multi-round generation results.** IMG continuously improves prompt-image alignment by executing multiple rounds.

	SDXL	SDXL-DPO	FLUX
Base Model	28.03	28.39	29.21
Base Model + SLD [65]	27.35	27.52	28.14
Base Model + IMG (Ours)	**28.56**	**28.79**	**29.48**

Table 3. **Quantitative comparison with editing-based methods.** SLD [65] is designed to enhance alignment through local editing but often at the cost of overall image quality.

	SDXL	SDXL-DPO	FLUX
Base Model	28.03	28.39	29.21
Base Model + IMG (LLaVA)	**28.56**	**28.79**	**29.48**
Base Model + IMG (Qwen-VL)	**28.48**	**28.73**	**29.52**

Table 4. **Generalization capability to different MLLMs.** We further integrate IMG with Qwen-VL [4]. The results indicate consistent improvement across different MLLMs.

nis shoe, it incorrectly deletes the intended tennis ball. In the second case, SLD fails to recognize the unintended human gesture, resulting in no edits. These limitations arise primarily because the LLM in SLD interprets image content through text-based bounding box descriptions provided by a detector, which introduces several risks: a) inaccurate detection results that lead to incorrect editing instructions from the LLM (*e.g.*, removing the tennis ball) and b) an inability to address quality-related misalignments (*e.g.*, overlooking the unintended gesture). In contrast, IMG, leveraging an MLLM that processes both images and prompts as input, more accurately identifies potential misalignments.

Quantitative Comparison. In Tab. 3, we further compare IMG with SLD on the HPD benchmark. The results reveal that SLD leads to performance degradation, aligning with our qualitative observations and analysis. While SLD may enhance alignment in locally edited regions, the accumulated errors from the detector, LLM, and editing pipeline ultimately reduce overall image quality and alignment performance. In contrast, IMG improves alignment through a native and more reliable re-generation process.

4.4. Discussion

MLLM misalignment detection accuracy. We evaluate our finetuned MLLM as follows: a) We sample 300 instruction-based editing cases from SEED-Data-Edit [16] as the test set; b) Both the original and the finetuned MLLM detect misalignment by predicting editing instructions; and c) GPT-4 [1] is employed to verify whether the predictions are semantically consistent with the ground truth. As a result, the finetuned MLLM achieves a 76.3% "yes" rate, compared to 42.3% for the original MLLM.

Generalization to different MLLMs. In Tab. 4, we additionally integrate IMG with Qwen2.5-VL-7B. Evaluations on the HPD benchmark show consistent alignment improvements across different MLLMs, highlighting IMG's strong generalization capability.

Multi-round generation. In Fig. 7, we demonstrate that IMG functions as an iterable framework that continuously enhances alignment through multiple rounds. For details that are not fully aligned in the first round IMG, *e.g.*, the dog-like shape of the daikon radish and the action of diving, a second round of IMG further refines these details.

Dense prompt generation. Generating images from long, complex prompts remains a challenging task for diffusion models [24, 36, 74]. In Fig. 8, we integrate IMG with FLUX and compare it against leading community models, including Stable Diffusion 3.5 Large (SD3.5) [2], PixArt-σ [7] and Playground v2.5 [36]. The results demonstrate that IMG substantially enhances details in generated images (as highlighted by the colored texts and boxes), enabling FLUX to stand out among competitors.

4.5. Ablation Studies

In Tab. 5, we examine the impact of different implicit guidance features, loss components, and loss ratios on the Pick-a-Pic test set via Pick Score to determine the optimal training scheme and hyperparameters.



Figure 8. **Dense prompt generation.** We integrate IMG with FLUX [35] and compare it against leading community models, including Stable Diffusion 3.5 Large (SD3.5) [2], PixArt-σ [7] and Playground v2.5 [36], using complex dense prompts.

Guidance features	SDXL	SDXL-DPO
Base Model	22.34	22.57
No Guidance	22.22	22.39
Text Embedding	22.38	22.68
Original MLLM	22.59	22.81
Finetuned MLLM (IMG)	**22.74**	**22.98**

(a) Impact of different guidance features.

Ratio	SDXL	SDXL-DPO
Base Model	22.34	22.57
0	22.56	22.76
0.5	22.56	22.80
1 (IMG)	**22.74**	**22.98**
2	22.50	22.79

(b) Impact of different λ values.

Components	SDXL	SDXL-DPO
Base Model	22.34	22.57
+ L_{base}	22.56	22.76
+ $L_{\text{pref}-\text{DPO}}$	22.65	22.88
+ $L_{\text{pref}-\text{SPIN}}$ (IMG)	**22.74**	**22.98**

(c) Impact of each loss component.

Table 5. **Ablation studies.** We examine the impact of different (a) guidance features, (b) loss ratios, and (c) loss components on the Pick-a-Pic test set via the average Pick Score. Higher Pick Scores indicate better preference alignment.

Impact of different guidance features. In Tab. 5a, we examine various guidance features used by the Implicit Aligner to refine initial image features into better-aligned features. "No Guidance", which directly uses initial image features as aligned ones, performs worse than the "Base Model", underscoring the need for feature alignment. Using "Text Embedding" (text prompt embeddings from the diffusion text encoder) to guide feature alignment provides only minor improvements, as it lacks information on specific misalignments. "Original MLLM" features offer significant gains across two base models, making SDXL + IMG competitive with SDXL-DPO and highlighting MLLM's alignment capability. IMG achieves the best alignment scores with our "Finetuned MLLM". This aligns with the accuracy improvement in Sec. 4.4 and validates the effectiveness of our customized finetuning task in Sec. 3.2.

Impact of different λ values. In Tab. 5b, we examine the impact of varying the loss ratio λ in Eq. 7, which controls the strength of L_{pref}. Setting $\lambda = 0$ reduces the objective to L_{base} alone. We find that a small λ may not sufficiently activate the effect of L_{pref}, while an excessively large λ can lead to training instability and thus suboptimal performance, as the Implicit Aligner is trained from scratch rather than finetuned from a reference model. Empirically, $\lambda = 1$ provides a balanced and effective trade-off.

Impact of each loss component. In Tab. 5c, we examine each loss component in Eq. 7. Using the basic objective L_{base} (Eq. 4), SDXL + IMG already achieves alignment performance competitive with SDXL-DPO. Furthermore, IMG is compatible with SDXL-DPO, enabling additional performance gains. For L_{pref}, we separately evaluate the impact of its DPO and SPIN components in Eq. 5. Results show that these two components provide progressive gains.

5. Conclusion

In this paper, we propose Implicit Multimodal Guidance (IMG), a novel re-generation-based alignment framework. Unlike existing finetuning-based and editing-based approaches, IMG enhances alignment performance without requiring additional finetuning data or explicit editing operations. Specifically, given a generated image and its prompt, IMG involves an MLLM that identifies potential misalignments and an Implicit Aligner that reduces misalignments and facilitates re-generation by refining diffusion conditioning features. The Implicit Aligner is optimized through a trainable Iteratively Updated Preference Objective. Extensive qualitative and quantitative evaluations on SDXL, SDXL-DPO, and FLUX show that IMG outperforms existing alignment methods. Furthermore, IMG acts as a flexible plug-and-play adapter, seamlessly enhancing prior finetuning-based alignment methods.

Acknowledgments

This research was supported in part by National Science Foundation under Award #2427478 - CAREER Program, and by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education under Award #2229873 - National AI Institute for Exceptional Education. This project was also partially supported by cyberinfrastructure resources and services provided by College of Computing at the Georgia Institute of Technology, Atlanta, Georgia, USA.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 2, 3, 7

[2] Stability AI. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024. 7, 8

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 3

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[5] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2023. 2, 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3, 4, 5

[7] 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:2403.04692*, 2024. 7, 8

[8] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *ICML*, 2024. 2, 5

[9] Mang Tik Chiu, Yuqian Zhou, Lingzhi Zhang, Zhe Lin, Connelly Barnes, Sohrab Amirghodsi, Eli Shechtman, and Humphrey Shi. Brush2prompt: Contextual prompt generator for object inpainting. In *CVPR*, 2024. 3

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv:2309.17400*, 2023. 2, 3

[11] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv:2309.15807*, 2023. 2

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 2, 3

[13] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 2024. 2, 3

[14] 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. In *NeurIPS*, 2024. 3

[15] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *ICLR*, 2024. 3

[16] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 7

[17] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair diffusion: A comprehensive multimodal object-level image editor. In *CVPR*, 2024. 3

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[19] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *AAAI*, 2022. 3

[20] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023. 3

[21] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *CVPR*, 2024. 3

[22] Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. In *CVPR*, 2025. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[24] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv:2403.05135*, 2024. 5, 6, 7

[25] Jiannan Huang, Jun Hao Liew, Hanshu Yan, Yuyang Yin, Yao Zhao, Humphrey Shi, and Yunchao Wei. Classdiffusion: More aligned personalization tuning with explicit class guidance. In *ICLR*, 2025. 3

[26] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023. 5

[27] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv:2302.09778*, 2023. 3

[28] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, 2024. 3

[29] Arman Isajanyan, Artur Shatveryan, David Kocharian, Zhangyang Wang, and Humphrey Shi. Social reward: Evaluating and enhancing generative AI through million-user feedback from an online creative community. In *ICLR*, 2024. 3

[30] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *CVPR*, 2024. 3

[31] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. In *NeurIPS*, 2024. 5, 6

[32] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024. 3

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2015. 5

[34] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 3, 4, 5

[35] Black Forest Labs. Flux. https://blackforestlabs.ai/, 2024. 1, 2, 3, 5, 6, 8

[36] 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:2402.17245*, 2024. 7, 8

[37] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *CVPR*, 2021. 5

[38] Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. In *ACL*, 2021. 3

[39] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *TMLR*, 2024. 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 3, 5

[41] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2022. 3

[42] Zeyu Liu, Zanlin Ni, Yeguo Hua, Xin Deng, Xiao Ma, Cheng Zhong, and Gao Huang. Coda: Repurposing continuous vaes for discrete tokenization. *arXiv preprint arXiv:2503.17760*, 2025. 3

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[44] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 3

[45] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *CVPR*, 2023. 3

[46] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *NeurIPS*, 2023. 2, 3

[47] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. 3, 6

[48] OpenAI. Openai o1. https://openai.com/o1/, 2024. 2

[49] OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/, 2024. 2

[50] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *arXiv:2405.16537*, 2024. 3

[51] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *ICLR*, 2024. 3

[52] 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. In *ICLR*, 2023. 2, 5, 6

[53] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv:2305.11147*, 2023. 3

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[55] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2024. 2, 3, 5

[56] 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. *JMLR*, 2020. 3

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 3

[58] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *CVPR*, 2022. 5

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3

[61] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. In *ECCV*, 2024. 3

[62] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. 2, 3, 5, 6

[63] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv:2404.02733*, 2024. 5

[64] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. In *NeurIPS*, 2024. 3

[65] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *CVPR*, 2024. 2, 3, 4, 5, 6, 7

[66] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv:2306.09341*, 2023. 2, 5, 6

[67] XLabs-AI. X-flux. https://github.com/XLabs-AI/x-flux, 2024. 3, 5

[68] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv:2304.05977*, 2023. 3

[69] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile Diffusion: Text, images and variations all in one diffusion model. *arXiv:2211.08332*, 2022. 2

[70] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *CVPR*, 2024. 3

[71] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *CVPR*, 2024. 3

[72] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *ICML*, 2024. 3

[73] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023. 3, 5, 6

[74] 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. *TMLR*, 2022. 5, 6, 7

[75] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *arXiv:2402.10210*, 2024. 2, 3, 5

[76] Gong Zhang, Kihyuk Sohn, Meera Hahn, Humphrey Shi, and Irfan Essa. Finestyle: Fine-grained controllable style personalization for text-to-image models. In *NeurIPS*, 2024. 3

[77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[78] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 3

[79] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In *ICML*, 2024. 3