

OSPO: Object-Centric Self-Improving Preference Optimization for Text-to-Image Generation

Yoonjin Oh, Yongjin Kim, Hyomin Kim, Donghwan Chi, Sungwoong Kim*
 Korea University

{dhdbsr1w, rla020, khmiee, zheedong, swkim01}@korea.ac.kr

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have enabled unified multimodal understanding and generation. However, they still struggle with fine-grained text–image alignment, often failing to faithfully depict objects with correct attributes such as color, shape, and spatial relationships. To mitigate this issue, previous studies have explored preference optimization methods such as DPO and GRPO, but these approaches incur substantial computational cost, both in constructing preference data and in performing optimization. This has motivated self-improving preference optimization approaches, in which the MLLM autonomously generates its own training data, self-estimates preference feedback, and self-optimizes using the resulting self-constructed preference pairs. However, existing self-improving methods still overlook fine-grained, object-level semantics, allowing object hallucination to persist. To tackle this problem, we propose Object-centric Self-improving Preference Optimization (OSPO), a self-improving framework designed to enhance object-level text–image alignment. OSPO explicitly constructs object-centric preference data without relying on any external data or models. We also introduce a new approach that leverages attention-based object masks together with an object-weighted SimPO loss to enhance object-specific fidelity. Extensive experiments on three compositional image generation benchmarks demonstrate that OSPO significantly improves fine-grained alignment and reduces object hallucination, outperforming prior self-improving methods and even specialized diffusion-based text-to-image models. The project page is available at <https://ku-agi.github.io/OSPO/>.

1. Introduction

Recent advancements in Unified Multimodal Large Language Models (Unified MLLMs) [11, 38–40, 45, 49–51] have achieved remarkable performance, demonstrating the ability to perform diverse vision tasks ranging from image

understanding to generation within a single model. Despite these successes, a persistent challenge remains in achieving fine-grained alignment between text and image. In text-to-image (T2I) generation, MLLMs often fail to faithfully render the details specified in the input prompt, including spatial relationships and precise visual attributes of objects. A critical and related failure mode is object hallucination, which encompasses not only the generation of non-existent objects but also the omission or distortion of objects described in the prompt.

Early work has addressed fine-grained alignment and hallucination using feedback-based post-training methods such as DPO [33, 43], PPO [36] and GRPO [37]. Despite their effectiveness, these approaches suffer from two key limitations: high cost and off-policy optimization. DPO requires large amounts of preference pairs curated by humans or stronger external models; collecting such data for image generation is inherently far more complex and costly than for text generation. PPO and GRPO do not require pre-collected preference pairs, yet involve multiple models during training incurring substantial computational overhead. Furthermore, all three methods introduce off-policy bias from external sources: DPO is trained on preference pairs whose distribution diverges from the model’s own outputs, while PPO and GRPO optimize against reward models trained on externally sourced data, creating a distributional mismatch that leads to unstable training.

To overcome these constraints, recent works have introduced self-improving approaches [14, 26, 31] that eliminate reliance on external data or models by generating training data or reward signals internally. This direction is especially promising with the advent of Unified MLLMs, whose dual abilities in image understanding and generation enable a self-improving loop in which the model evaluates and refines its own outputs. Such a loop is particularly viable because MLLMs’ visual comprehension has matured to the point where the model can serve as a reliable judge of its own generations. However, despite these strengths, current self-improving methods still lack mechanisms that explicitly enforce fine-grained text–image alignment, leaving them insufficient for fully resolving object-level hallucination and detailed alignment failures.

*Corresponding author

In this paper, we propose **Object-centric Self-improving Preference Optimization (OSPO)**, a self-improving framework that autonomously constructs fine-grained, object-focused T2I preference pairs and performs preference optimization guided by predicted object masks. Here, *object-centric* refers to treating each object independently as a distinct unit of alignment, rather than treating the image holistically. This framework enables the model to better capture fine-grained text–image correspondences at the individual object level. OSPO operates in five stages: prompt generation, prompt perturbation and densification to create prompt pairs sharing global context but differing in fine-grained details, candidate image generation with object mask, preference pair construction based on Self-VQA, and preference optimization. We validate the effectiveness of OSPO across multiple MLLMs on three established benchmarks for fine-grained T2I generation. Notably, OSPO substantially improves fine-grained alignment, outperforming existing self-improving methods and even surpassing diffusion-based models specifically designed solely for image generation.

To summarize, our main contributions are as follows:

- We present OSPO, a five-stage self-improving framework that mitigates object hallucination in T2I generation without relying on either external datasets or auxiliary models.
- We propose a pipeline that constructs high-quality, object-centric preference data with shared global semantics but fine-grained local differences, and incorporates object-aware supervision using attention-based object masks with an object-weighted SimPO loss.
- We demonstrate the empirical effectiveness of OSPO with substantial object-level alignment improvements across multiple T2I benchmarks, surpassing prior self-improving and specialized diffusion models.

2. Related Works

2.1. Multimodal Large Language Models (MLLMs)

Early Multimodal Large Language Models (MLLMs), particularly large vision-language models [4, 7, 24, 54], primarily extended pretrained LLMs for visual understanding by integrating vision encoders. Recent advances, however, have moved toward unified MLLMs that can both understand and generate visual content. This unification enables a single model to handle diverse tasks, ranging from image understanding to generation, within a shared set of model parameters. Among them, Discrete MLLMs [3, 11, 38, 39, 45, 47, 49] represent visual information as quantized tokens produced by vector quantization (VQ) [10, 42] or learned tokenizers. These models employ vector quantization to discretize images into tokens, thereby allowing both vision and language to be modeled uniformly through next-token prediction. In this work, we focus on discrete unified MLLMs that generate images as sequences of discrete tokens.

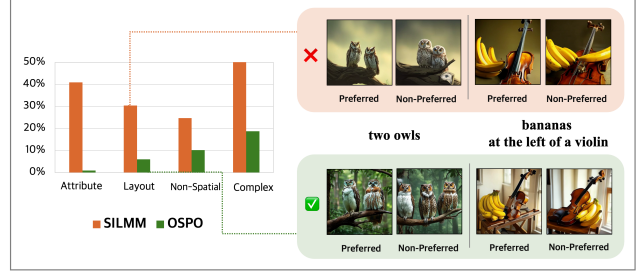


Figure 1. Proportion of Preference-Ambiguous Pairs across prompt categories in SILMM and OSPO (ours) datasets, along with representative examples. Lower is better.

2.2. Self-Improving Frameworks for T2I

Recent works have proposed self-improving preference optimization frameworks for T2I generation in MLLMs [14, 26, 31]. These frameworks commonly adopt a *Best-of-N sampling* strategy, which produces multiple candidate images from the same prompt and selects the highest- and lowest-scoring ones to form preference pairs. While SILMM [31], the first such framework, targets compositional T2I generation—a task that inherently requires fine-grained object-level alignment—this objective is only partially reflected in its design, limited to the preference assignment stage. Subsequent frameworks [14, 26] jointly train on T2I generation and image captioning but do not address fine-grained alignment as an explicit objective. Overall, existing frameworks lack object-centric design across their generation, preference assignment, and optimization stages, limiting their ability to address object hallucination. In contrast, our framework explicitly integrates object-level components throughout all stages, enabling more precise and semantically faithful T2I generation.

3. Framework

3.1. Preliminary

We analyze the 16,000-sample self-generated preference dataset produced using Janus-Pro-7B [3] following the workflow of SILMM [31]. We identify a systematic failure mode of **Preference-Ambiguous Pair**, where both images exhibit indistinguishable levels of prompt fidelity, providing no meaningful preference signal for training. A particularly problematic subcase is **Preference-False Pair**, where both images are uniformly correct or uniformly incorrect across all semantic units. The forced preferred/non-preferred label assignment by *Best-of-N sampling* actively introduces contradictory supervision. Figure 1 reports the proportion of such preference-ambiguous pairs across prompt categories, revealing that this problem is prevalent across all categories. This indicates that *Best-of-N sampling* is indiscriminate by nature and yields noisy preference assignments poorly aligned with the training objective. Formal definitions and the measurement protocol are provided in

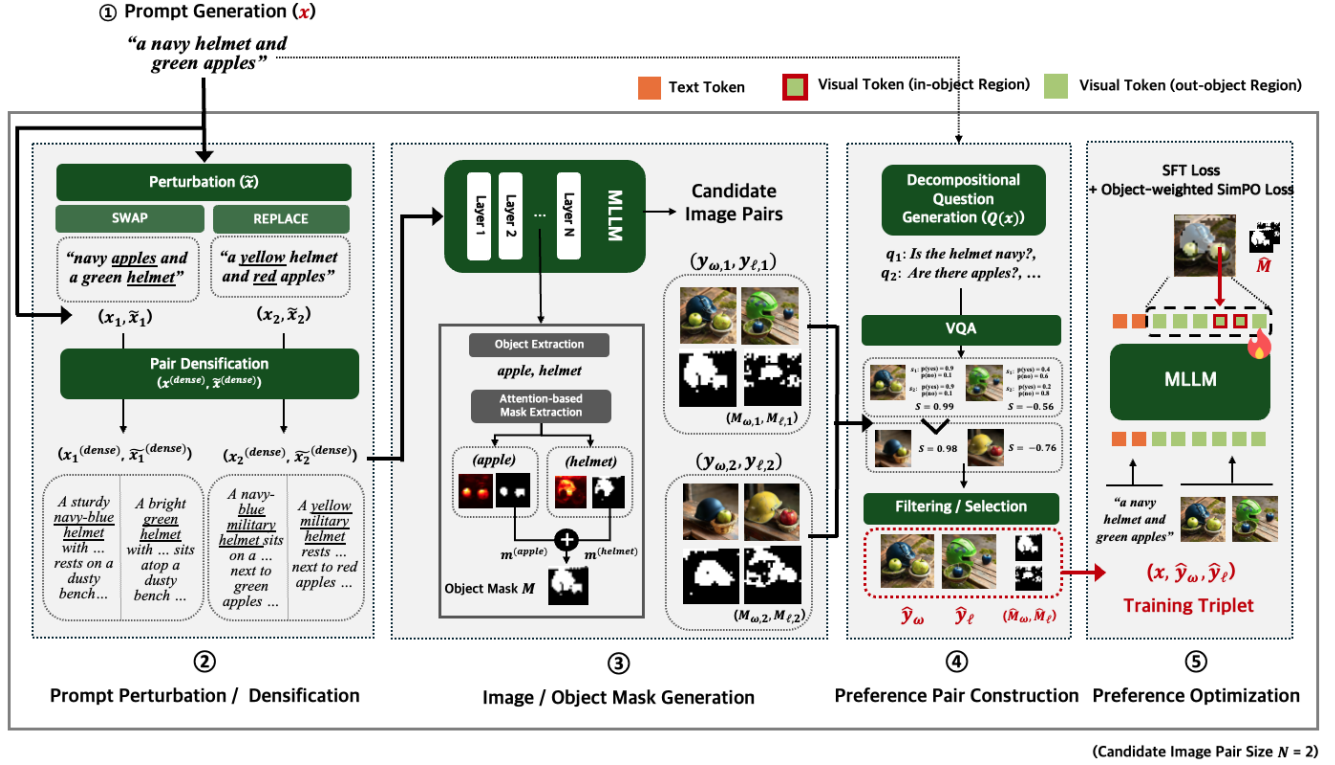


Figure 2. Overview of the OSPO framework: (Stage 1) The MLLM generates a set of base text prompts. (Stage 2) For each base prompt, the model generates multiple perturbed variants using three strategies (Replace, Swap, and Drop) and each original-perturbed pair is jointly densified. Each pair of original and perturbed prompt is pairwise densified by the MLLM. (Stage 3) The MLLM generates candidate preferred and non-preferred images from each densified prompt pair. (Stage 4) The model constructs atomic decompositional VQA questions and evaluates each candidate image’s prompt fidelity using the Self-VQA alignment score S , filtering out noisy supervision and selecting a single final image pair. (Stage 5) The MLLM is fine-tuned using preference optimization with the proposed Object-weighted SimPO loss together with SFT loss.

Appendix A.

3.2. Overview

We introduce **OSPO**, a five-stage self-improving preference optimization framework designed to mitigate object hallucination in T2I generation. The framework is fully self-contained, relying neither on external data nor pretrained reward models. By leveraging the dual capabilities of unified MLLMs, OSPO synthesizes text-image data from scratch without relying on *Best-of- N sampling*. It then constructs object-centric preference pairs that deliver precise, fine-grained supervision, addressing limitations of prior self-improving frameworks and mitigating object-level hallucination. The five stages are as follows: (1) Prompt Generation (2) Prompt Perturbation and Densification (3) Image and Object Mask Generation (4) VQA-based Preference Pair Construction (5) Preference Optimization. Figure 2 describes an overview of the OSPO framework.

3.3. Prompt Generation

Inspired by T2I-Compbench++ [19], we first construct the base set of training prompts, where each training prompt x

serves as the basis for subsequent data construction. The prompt set consists of four categories: Attribute, Layout, Non-spatial Relationship, and Complex Composition. The Attribute category is decomposed into Color, Shape, and Texture, while Layout is decomposed into 2D and 3D-Spatial Relationships. The remaining categories are single-level categories. Non-spatial Relationship describes dynamic actions or states, while Complex Composition combines multiple categories to form richer and longer prompts. All prompts are generated by MLLMs via in-context learning. The prompt template is provided in Appendix I.

3.4. Prompt Perturbation and Densification

To construct a preferred and non-preferred image pair (y_w, y_l) for each text prompt x , we generate the two images independently rather than using the common *Best-of- N sampling* strategy. We therefore generate preferred and non-preferred image variants using prompt pairs that share the same global semantics but differ slightly in object-level details, while leveraging the MLLM’s in-context learning capability. The templates are presented in Appendix I.

3.4.1. Prompt Perturbation

Inspired by SugarCrep [15] and WinoGround [41], multiple perturbed prompts $\tilde{x}$ are generated for each text prompt x using the following perturbation strategies:

- **Replace**: substitutes an object or attribute in the prompt with another that was not originally present, creating a novel compositional combination.
- **Swap**: exchanges the positions of objects or attributes within the prompt, inducing a different relational binding from the original.
- **Drop**: removes an object or attribute from the prompt, introducing semantic ambiguity.

After generating N perturbed prompts, each perturbed prompt is paired with the original text prompt to form N distinct pairs: $(x, \tilde{x}_1), (x, \tilde{x}_2), \dots, (x, \tilde{x}_N)$. The original prompt x is shared across all pairs, while each $\tilde{x}_n$, $n \in \{1, \dots, N\}$ represents a different perturbation. In our experiments, we set $N = 3$ by default.

3.4.2. Prompt Densification

Prior work [23, 46, 52, 53] shows that *Prompt Densification*—adding contextual detail to prompts—improves faithfulness in T2I generation. We extend this idea by applying densification to each prompt pair $(x, \tilde{x})$ before image synthesis, ensuring that paired images share global semantics but differ only in fine-grained object details. Each prompt pair is densified jointly, producing two prompts with consistent background context but distinct object-level semantics. We analyze the impact of densification in Section 5.1.3.

3.5. Image and Object Mask Generation

In this stage, we generate candidate image pairs and their corresponding object masks, which are used in the preference optimization stage to introduce object-centric guidance during training.

3.5.1. Image Generation

From each densified prompt pair $(x^{(dense)}, \tilde{x}^{(dense)})$, we generate one preferred and one non-preferred candidate image, denoted as (y_w, y_ℓ) . With N prompt pairs, N image pairs (a total of $2N$ images) are generated: $(y_{w,1}, y_{\ell,1}), (y_{w,2}, y_{\ell,2}), \dots, (y_{w,N}, y_{\ell,N})$.

3.5.2. Object Mask Generation

During image generation, we extract a binary token-level mask, referred to as the *Object Mask*, which indicates whether each visual token belongs to the object region of the image. This mask is obtained from attention weights, which makes the approach computationally efficient, as it does not require a separate segmentation model and leverages the model’s internal interactions [1, 21].

To obtain the object mask, we extract the attention distribution from a given object-representing text token over all visual tokens within the intermediate layers of the MLLM. Here, we exclude the top- k earliest and top- k latest layers to avoid unstable activations and over-smoothing. We then

average them across heads and layers, and reshape the result into a 2D spatial attention map aligned with the image grid. This map is binarized into an object-specific foreground mask m using OTSU’s adaptive thresholding method [28]. We set $k = 5$ by default. By repeating this process for all objects described in the original prompt x and taking the union of the resulting masks, we obtain a single object mask M for each generated image. Detailed formulations and algorithms are provided in Appendix B.2.

3.6. VQA-based Preference Pair Construction

Since preference optimization is highly sensitive to data quality, OSPO includes a pair filtering and selection stage to ensure the reliability of training pairs among the N candidate image pairs. The overall process is guided by the Self-VQA results.

3.6.1. Decompositional VQA

Decompositional Visual Question Answering (VQA) is a common method for fine-grained evaluation of text–image alignment [12, 17–19]. It evaluates an image at the level of atomic semantic elements [6]. This enables fine-grained evaluation across multiple semantic dimensions. In OSPO, every prompt x in the base prompt set is decomposed by the MLLM into a set of binary Yes/No questions, $Q(x) = \{q_1, q_2, \dots, q_K\}$, where K denotes the number of atomic semantic elements. Each of the N generated image pairs is evaluated using the corresponding question set, and the MLLM outputs Yes/No probabilities for every question. The alignment score $S(y)$ for an image y is then computed as the average probability margin between the Yes and No responses: $s_k(y) = p(\text{yes} \mid y, q_k) - p(\text{no} \mid y, q_k)$, $S(y) = \frac{1}{K} \sum_{k=1}^K s_k(y)$.

3.6.2. Pair Filtering and Selection

A candidate image pair (y_w, y_ℓ) is filtered based on each image’s per-question scores $s_1(\cdot), \dots, s_K(\cdot)$ and their aggregated alignment score $S(\cdot)$ for both y_w and y_ℓ . Since the two images form a coupled pair, if either image is deemed invalid, the entire pair is discarded. In particular, the pair is discarded if the aggregated alignment score for the preferred image y_w falls below the threshold: $S(y_w) < \tau$, where τ is a filtering threshold and we set $\tau = 0.6$ by default. Similarly, the pair is discarded if the non-preferred image y_ℓ satisfies $s_k(y_\ell) > 0$, $\forall k \in \{1, \dots, K\}$. By using both per-question scores s_k and the aggregated score S , we remove preference-ambiguous and preference-false pairs to avoid noisy supervision. If no candidate pair survives this process, the entire text prompt and its generated images are excluded from training.

After filtering, the final training image pair $(\hat{y}_w, \hat{y}_\ell)$ is selected as the pair whose preferred image attains the highest alignment score S . Finally, we construct **preference training triplet**: $(x, \hat{y}_w, \hat{y}_\ell)$.

3.7. Preference Optimization

Using the set of constructed preference triplets $\mathcal{D}$, we optimize the model using a combination of the Object-weighted SimPO loss and the Supervised Fine-Tuning (SFT) loss. Each loss term captures a complementary aspect of the preference learning objective.

Object-weighted SimPO Loss. Due to the nature of visual tokens, the standard SimPO objective [27], which aggregates reward over all tokens, can dilute the training signal by including many tokens that are irrelevant to the target object. Since our goal is fine-grained T2I alignment, we emphasize object-relevant visual tokens by applying spatial weights to the token-level reward in the standard SimPO. The Object-weighted SimPO loss is defined as:

$$\mathcal{L}_{\text{Obj-SimPO}} = -\mathbb{E}_{(x, y_w, y_\ell) \sim \mathcal{D}} \left[\log \sigma \left(\frac{\beta}{|y_w|} \sum_{t=1}^{|y_w|} (w_t)_t \log \pi_\theta((y_w)_t \mid x, (y_w)_{<t}) - \frac{\beta}{|y_\ell|} \sum_{t=1}^{|y_\ell|} (w_t)_t \log \pi_\theta((y_\ell)_t \mid x, (y_\ell)_{<t}) - \gamma \right) \right], \quad (1)$$

where π_θ denotes the MLLM. σ , β , γ , and w_t denote the sigmoid function, reward scaling factor, target reward margin, and token-level weighting factor computed by the object mask m_t , respectively. We define $w_t = 1 + \alpha m_t$, $m_t \in [0, 1]$, where t denotes token position and α controls the emphasis on object-relevant visual tokens. This weighting scheme strengthens gradients on object-relevant tokens, producing sharper and more targeted preference signals. We set $\alpha = 1$, $\beta = 5$, $\gamma = 2.5$ by default.

SFT Loss. Visual token sequences encode spatial and structural patterns, where global coherence matters far more than the correctness of any single token. Unlike text generation, a token-level reward alone cannot effectively enforce object shape, geometry, or layout. Thus, leveraging only the PO objective may fail to provide sufficiently strong structural guidance. To complement this, we apply an SFT loss using preferred image as an anchor, ensuring consistent supervision over the full visual token sequence and reinforcing globally coherent image generation. The SFT loss is defined as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x, y_w) \sim \mathcal{D}} \left[\frac{1}{|y_w|} \sum_{t=1}^{|y_w|} \log \pi_\theta((y_w)_t \mid x, (y_w)_{<t}) \right]. \quad (2)$$

OSPO Loss. The final training objective is a weighted combination of Object-weighted SimPO loss and SFT loss:

$\mathcal{L}_{\text{OSPO}} = \mathcal{L}_{\text{Obj-SimPO}} + \lambda \mathcal{L}_{\text{SFT}}$. Here, λ is a loss weight for SFT loss, which we set $\lambda = 2$ by default.

4. Experiment

4.1. Experiment Setting

We conducted experiments using the Janus-Pro-1B and Janus-Pro-7B models [3] as our primary backbones. For training, we used a dataset of 20,000 samples, obtained after filtering, spanning four categories: Attribute, Layout, Non-spatial Relationship, and Complex Composition.

4.1.1. Benchmarks

We evaluated T2I generation capabilities on the three standard benchmarks: T2I-CompBench++ [19], DPGBench [17], and GenEval [12]. Each evaluation was conducted following the standard configurations and protocols. Detailed descriptions of the evaluation methodology for each benchmark are provided in Appendix D.

- **T2I-CompBench++:** Evaluates compositional image generation ability using 2,400 prompts across four categories—Attribute, Layout, Non-spatial Relationship, and Complex Composition.
- **DPGBench:** Evaluates image generation ability using 1,065 long, semantically rich prompts, which have an average token length of 83.91. It focuses on the semantic alignment of complex structures within extended narratives.
- **GenEval:** Evaluates compositional image generation ability using 550 prompts in six categories using an evaluation protocol similar to T2I-CompBench++.

4.1.2. Baselines

To evaluate the effectiveness of the OSPO framework, we compare OSPO with other self-improving preference optimization frameworks for T2I generation. Among self-improving frameworks, we include SILMM [31] and SUDER [14], both of which construct preference pairs by generating images and assigning rewards without relying on external supervision. Note that SUDER has not released its model checkpoint, thus its results are omitted for certain categories or benchmarks.

4.1.3. Implementation Details

All experiments were conducted on eight NVIDIA A100 (80GB) GPUs. Additional implementation details, including training hyperparameters, are provided in Appendix C.

4.2. Main Results

As presented in Table 1, the OSPO-trained model outperforms all self-improving baselines on T2I-CompBench++ for both the 1B and 7B model scales. Notably, substantial improvements are observed in the Attribute category, while the Non-spatial and Complex categories show relatively smaller improvements. This disparity arises from

Table 1. Comparison on T2I-CompBench++ [19]. ↑ indicates higher is better, with bold highlighting the best score and underline indicating the second best among the unified MLLMs.

Model	Size	Attribute ↑			Layout ↑			Non-	Complex ↑
		Color	Shape	Texture	Spatial-2D	Spatial-3D	Numeracy	Spatial ↑	
Diffusion Models									
SD-XL	–	0.6369	0.5408	0.5637	0.2032	–	–	0.3110	0.4091
DALL-E 3	–	0.7785	0.6205	0.7036	0.2865	–	–	0.3003	0.3773
FLUX.1	–	0.7407	0.5718	0.6922	0.2863	–	–	0.3127	0.3703
Unified MLLMs									
Show-o	7B	0.5600	0.4100	0.4600	0.2000	–	–	0.3000	0.2900
Unitok-MLLM	7B	0.7745	0.5195	0.6431	0.2668	0.4088	<u>0.5675</u>	0.3131	0.3617
Janus-Pro	1B	0.3542	0.2291	0.2843	0.0756	0.2367	0.2714	0.2809	0.2693
+ SILMM	1B	0.6097	0.3164	0.4474	0.1539	0.2632	0.2826	0.3122	0.3679
+ SUDER	1B	0.7765	0.5106	0.6767	0.2464	–	–	0.3130	0.3657
+ OSPO (ours)	1B	<u>0.8349</u>	0.5763	0.7274	<u>0.3077</u>	<u>0.4029</u>	0.5653	<u>0.3158</u>	<u>0.3935</u>
Janus-Pro	7B	0.5215	0.3272	0.4050	0.1654	0.2679	0.4431	0.3109	0.3868
+ SILMM	7B	0.7394	0.4325	0.5796	0.2105	0.3572	0.5073	0.3113	0.3725
+ SUDER	7B	0.7824	<u>0.5786</u>	<u>0.7292</u>	0.2524	–	–	0.3141	0.3858
+ OSPO (ours)	7B	0.8567	0.6386	0.7727	0.3562	0.3926	0.5908	0.3161	0.4147

Table 2. Comparison on DPGBench [17] and GenEval [12]. ↑ indicates higher is better, with bold indicating the best score and underline indicating the second best among the unified MLLMs.

Model	Size	DPGBench ↑						GenEval ↑						
		Attribute	Relation	Entity	Other	Global	Overall	Single	Two	Count	Colors	Position	Color Attribution	Overall
Diffusion Models														
SD-XL	-	82.43	80.91	86.76	80.41	83.27	74.65	0.98	0.74	0.39	0.85	0.15	0.15	0.55
DALL-E 3	-	89.61	88.39	90.58	89.83	90.97	83.50	0.96	0.87	0.47	0.83	0.43	0.43	0.67
SD3-Medium	-	-	-	-	-	-	-	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Unified MLLMs														
Show-o2	7B	<u>89.96</u>	91.81	<u>91.78</u>	91.64	89.00	86.14	1.00	0.87	0.58	0.92	0.52	0.62	0.76
Emu3	8B	86.33	90.61	87.17	<u>89.75</u>	87.54	81.60	0.98	0.69	0.33	0.78	0.15	0.16	0.52
Unitok-MLLM	7B	88.37	91.39	88.13	87.54	83.98	81.87	<u>0.99</u>	0.75	0.39	0.80	0.21	0.41	0.59
Janus-Pro	1B	88.17	88.98	88.63	88.30	87.58	82.63	<u>0.99</u>	0.82	0.48	0.89	0.62	0.57	0.73
+ SILMM	1B	89.87	<u>91.91</u>	88.67	82.60	78.22	83.56	<u>0.99</u>	0.86	0.48	0.89	0.70	0.62	0.75
+ SUDER	1B	-	-	-	-	-	-	<u>0.99</u>	0.87	0.54	0.90	0.74	0.62	0.78
+ OSPO (ours)	1B	90.42	87.24	89.09	87.08	89.21	84.46	<u>0.99</u>	<u>0.89</u>	0.48	0.92	0.67	0.61	0.76
Janus-Pro	7B	86.11	92.72	88.30	88.85	82.62	83.81	0.98	0.87	0.57	0.89	0.78	0.67	0.80
+ SILMM	7B	89.47	91.26	90.33	88.37	91.46	84.56	0.98	0.87	0.58	0.90	0.77	<u>0.71</u>	0.80
+ SUDER	7B	-	-	-	-	-	-	<u>0.99</u>	<u>0.89</u>	0.70	0.92	<u>0.82</u>	<u>0.71</u>	0.84
+ OSPO (ours)	7B	89.03	89.45	92.14	84.22	<u>91.09</u>	<u>85.61</u>	1.00	0.90	<u>0.63</u>	<u>0.91</u>	0.83	0.72	<u>0.83</u>

differences in the evaluation metrics across categories, particularly the CLIP-T score [13, 32], which is used for the Non-spatial and Complex categories, has limited sensitivity within its scoring range. In this context, comparisons with other methods and models show that our model performs comparably well in those categories.

Our models of different sizes show consistent gains on other benchmarks, demonstrating the robustness of OSPO across scales. In Table 2, our 7B model achieves the second-highest overall score among unified MLLMs on DPGBench, while attaining the highest overall score among

Janus-Pro-based self-improving methods. Moreover, it achieves the second-best overall score on the GenEval benchmark with only a marginal performance gap. This gap mainly comes from the Count category, a known weakness of MLLMs. While SUDER outperforms ours in this category, this is largely attributable to its use of external COCO [2] image-text pair data, which provides counting supervision beyond what a purely self-improving framework can access. We also observe isolated cases where the 1B model outperforms the 7B model in specific categories; however, these inversions are category-specific, and the 7B model



Figure 3. Qualitative examples from Janus-Pro-7B + OSPO on T2I-CompBench++, GenEval, and DPGBench prompts.

demonstrates more balanced gains across categories, resulting in superior overall performance. Overall, results in Table 2 demonstrate strong generalization ability of our model to out-of-distribution input prompts. Additionally, we evaluate general T2I generation performance beyond compositional ability using CLIP and LPIPS score in Appendix E.3. Qualitative examples generated by the OSPO-trained model are shown in Figure 3, with additional samples provided in Appendix G.

5. Analysis

We conducted several experiments to analyze the contributions of individual components within the OSPO framework. All experiments were conducted using the Janus-Pro-7B model, focusing on three key aspects: (1) preference data (2) training loss and (3) computational cost. Appendix E provides additional analyses including iterative training, alternative 7B backbones, and data-generation biases.

5.1. Object-centric Preference Data Construction

OSPO is a self-improving framework that generates object-centric preference data that provides clear preference signals effective for training, as shown in Figure 1 representing a low proportion of preference-ambiguous pairs. It generates its own training data from scratch, enabling the data volume to scale with the available compute. Therefore, we

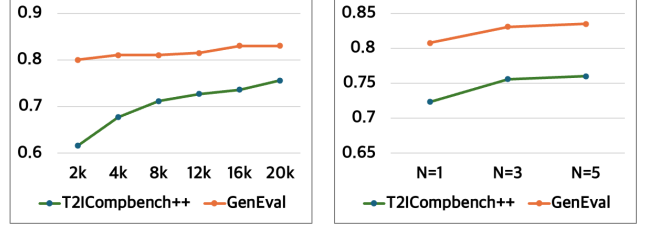


Figure 4. (Left) Effect of sample size on T2I-Compbench++ (attribute) and GenEval (overall) scores. (Right) Effect of candidate image pair size on T2I-Compbench++ (attribute) and GenEval (overall) scores. Higher is better for all benchmarks.

first investigate two factors that influence its performance: the sample size, which refers to the total number of text prompts x (see Section 3.3), and the candidate image pair size N (see Sections 3.4 and 3.5), which refers to the number of image pairs produced per prompt.

5.1.1. Sample Size

As shown in Figure 4, model performance consistently improves as the total number of training samples increases. Notably, even with a minimal dataset, the model still achieves a substantial performance gain over the baseline. This highlights the data efficiency of OSPO framework.

5.1.2. Candidate Image Pair Size

We further analyze the effect of the candidate image pair size on overall performance. As shown in Figure 4, sampling multiple candidate pairs per prompt leads to consistent performance improvements compared to using a single pair. Although the performance gain gradually stabilizes with larger pair sizes, the results suggest that a moderate number of candidates is sufficient to capture diverse and informative training signals.

5.1.3. Image Sampling and Pair Filtering/Selection

Effect of Prompt Densification on Image Fidelity. To verify that densified prompts with richer semantic descriptions really lead to more accurate text-image alignment and better model performance, we conducted a comparative study between two settings: generating images from the original prompt x and from its densified version $x^{(dense)}$. As shown in Table 3, training the model with OSPO including the densification stage consistently yields higher performance on both benchmarks. Therefore, these results validate the effectiveness of the densification process. Also, the results of text-image alignment evaluation are provided in Appendix E.5.

Effect of Filtering and Selection Strategy. We further analyze the interaction between the filtering-selection stage and the prompt densification stage. These two stages are closely interdependent: the effectiveness of the filtering and

Table 3. Effect of filtering and selection strategy in two different settings (with and without densification) construction stage on T2I-Compbench++ (attribute) and GenEval (overall) scores. Higher is better for all benchmarks.

Model	Filtering	Selection	T2I++ $\uparrow$	GenEval $\uparrow$
Janus-Pro-7B			0.418	0.796
OSPO w/ Densification	$\times$	$\times$	0.716	0.813
	$\checkmark$	$\times$	0.725	0.810
	$\times$	$\checkmark$	0.724	0.805
	$\checkmark$	$\checkmark$	0.756	0.831
OSPO w/o Densification	$\times$	$\times$	0.618	0.816
	$\checkmark$	$\times$	0.549	0.814
	$\times$	$\checkmark$	0.627	0.814
	$\checkmark$	$\checkmark$	0.641	0.823

Table 4. Ablation study on loss components with T2I-Compbench++ (attribute, layout) and GenEval (overall, position) scores. Higher is better for all benchmarks.

Model	$\mathcal{L}_{\text{SimPO}}$	$\mathcal{L}_{\text{Obj-SimPO}}$	$\mathcal{L}_{\text{SFT}}$	T2I++ $\uparrow$		GenEval $\uparrow$	
				Attr.	Layout	Over.	Position
Janus-Pro-7B				0.418	0.292	0.796	0.570
	$\checkmark$			0.779	0.416	0.785	0.778
		$\checkmark$		0.776	0.428	0.794	0.795
	$\checkmark$		$\checkmark$	0.755	0.437	0.827	0.795
+ OSPO (ours)		$\checkmark$	$\checkmark$	0.756	0.447	0.831	0.828

selection processes largely depends on the diversity and semantic richness of the generated image samples. To capture this dependency, we investigate the combined impact of densification and filtering on benchmark performance, highlighting how the quality and distribution of generated samples directly influence the downstream preference pair selection process. Table 3 shows that the filtering and selection strategies consistently improve benchmark performance, regardless of whether prompt densification is applied. Notably, the effect becomes more pronounced when prompts are not densified and images are generated directly. This indicates that filtering and selection play a more critical role in mitigating noise and ensuring data quality when training on images with incomplete semantic fidelity. Additional experiments regarding the filtering threshold are provided in Appendix E.6.

5.2. Training Loss Design for Object-centric PO

As shown in Table 4, both loss components contribute to consistent performance gains across benchmarks. In particular, improvements are notable in spatial text-image alignment (the *Layout* category in T2I-CompBench++ and the *Position* category in GenEval). This suggests that the Object-weighted SimPO loss effectively focuses learning on object-relevant regions, while the SFT loss provides complementary structural supervision by acting as a stabi-

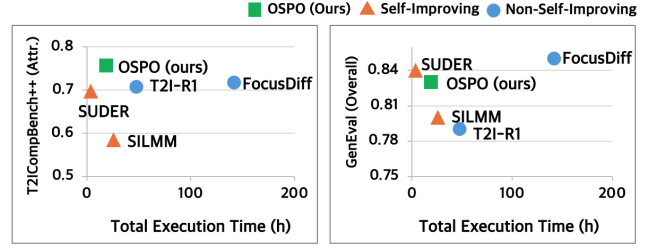


Figure 5. Performance–Efficiency trade-off across preference optimization frameworks for T2I generation. (Left) Comparison of T2I-CompBench++ (attribute) scores against execution cost. (Right) Comparison of GenEval (overall) scores against execution cost. Higher is better for all benchmarks.

lizing anchor during training. Together, these objectives enhance the model’s text-faithful and spatially coherent image synthesis.

5.3. Computational Cost

Finally, we analyze the time cost of the OSPO framework compared with other preference optimization frameworks including both self-improving frameworks [14, 31] and non-self-improving frameworks [20, 29]. The time cost was measured on a system equipped with eight NVIDIA A100 (80 GB) GPUs for one iteration of each framework. We first compare OSPO with SILMM [31]. As shown in Figure 5, OSPO achieves higher benchmark performance than SILMM [31] while being more time-efficient, primarily because it generates smaller, targeted candidate sets by decoupling the source prompt. Compared with T2I-R1 [20] and FocusDiff [29], which rely on GRPO [37] with multiple reward models or combine SFT and GRPO with image-editing datasets for fine-grained alignment, OSPO achieves comparable performance at substantially lower computational cost. A detailed breakdown of the wall-clock time for each stage of OSPO is provided in Appendix C.3.

6. Conclusion

We propose an **Object-centric Self-improving Preference Optimization (OSPO)** framework for T2I generation with MLLMs. OSPO leverages the unified multimodal capabilities of MLLMs to self-enhance fine-grained text-image alignment. By self-generating object-centric preference data that emphasize fine-grained semantic differences and applying object-weighted preference optimization, the model effectively self-improves its ability to capture detailed object-level semantics. Experiments on compositional image generation benchmarks show that OSPO greatly improves fine-grained alignment in T2I generation.

7. Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00353007, 20%) and the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2026-RS-2024-00436857, ITRC(Information Technology Research Center), 20%), (No. RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 20%), (IITP-2026-RS-2025-02304828, artificial intelligence star fellowship support program to nurture the best talents, 20%), and (RS-2025-02653113, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale, 20%).

References

- [1] Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13242–13251, 2025. 4
- [2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6, 4
- [3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Januspro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 5, 13, 14
- [4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2
- [5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3
- [6] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 4, 3
- [7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 2
- [8] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 8
- [9] Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025. 8
- [10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2
- [11] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 1, 2
- [12] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 4, 5, 6, 13, 14, 15
- [13] 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. 6
- [14] Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Jirong Wen, and Rui Yan. Suder: Self-improving unified large multimodal models for understanding and generation with dual self-rewards. 2025. 1, 2, 5, 8
- [15] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 4
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2
- [17] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 4, 5, 6, 13, 14, 15
- [18] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.
- [19] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhengguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3, 4, 5, 6, 13, 14, 15
- [20] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 8
- [21] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350, 2025. 4
- [22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3
- [23] WeiJie Li, Jin Wang, and Xuejie Zhang. Promptist: Automated prompt optimization for text-to-image synthesis. In *CCF international conference on natural language processing and Chinese computing*, pages 295–306. Springer, 2024. 4
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2
- [25] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025. 2, 4, 15
- [26] Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via su-

- pervised and reinforcement learning. *arXiv preprint arXiv:2505.23380*, 2025. 1, 2
- [27] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235, 2024. 5
- [28] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 4, 1
- [29] Kaihang Pan, Wendong Bu, Yuruo Wu, Yang Wu, Kai Shen, Yunfei Li, Hang Zhao, Juncheng Li, Siliang Tang, and Yuet-ing Zhuang. Focusdiff: Advancing fine-grained text-image alignment for autoregressive visual generation through rl. *arXiv preprint arXiv:2506.05501*, 2025. 8, 4
- [30] Kaihang Pan, Yang Wu, Wendong Bu, Kai Shen, Juncheng Li, Yingting Wang, Yunfei Li, Siliang Tang, Jun Xiao, Fei Wu, Hang Zhao, and Yueting Zhuang. Janus-pro-rl: Advancing collaborative visual comprehension and generation via reinforcement learning. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. Poster. 8
- [31] Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. Silmm: Self-improving large multimodal models for compositional text-to-image generation. *arXiv preprint arXiv:2412.05818*, 2024. 1, 2, 5, 8
- [32] 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. 6, 3, 4
- [33] 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. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 1, 5, 6
- [34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3
- [35] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3
- [36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1
- [37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1, 8
- [38] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 1, 2
- [39] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 2
- [40] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1
- [41] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visiolinguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 4
- [42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2
- [43] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024. 1
- [44] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 5, 6
- [45] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1, 2
- [46] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024. 4
- [47] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025. 2
- [48] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β -dpo: Direct preference optimization with dynamic β . *Advances in Neural Information Processing Systems*, 37: 129944–129966, 2024. 5
- [49] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 1, 2
- [50] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [51] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 1
- [52] Hongji Yang, Yucheng Zhou, Wencheng Han, and Jianbing Shen. Self-rewarding large vision-language models for opti-

- mizing prompts in text-to-image generation. *arXiv preprint arXiv:2505.16763*, 2025. [4](#)
- [53] Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v (ision) for automatic image design and generation. *arXiv preprint arXiv:2310.08541*, 2023. [4](#)
 - [54] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. [2](#), [3](#)
 - [55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [4](#)
 - [56] Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl. *arXiv preprint arXiv:2505.24875*, 2025. [8](#)
 - [57] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7571–7580, 2022. [3](#)
 - [58] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018. [3](#)