

Iterative Perceptual Alignment for VLMs via Deterministic Reconstruction Feedback

Xiaorui Chen^{1,2*}, Hanzhong Guo^{3,2*}, Nizhe Cai² , and Jieliang Luo^{2†} 

¹ Technical University of Munich, Munich, Germany

² Minus AI, Singapore

³ University of Hong Kong, Hong Kong
rluo@minus-ai.com

Abstract. While high-quality preference data is vital for fine-grained vision-language alignment, its creation remains bottlenecked by the high cost and subjective noise of human annotation. We propose a self-supervised framework that derives solid, highly discriminative supervisory signals through deterministic reconstruction feedback. By transforming abstract semantic matching into concrete visual reconstruction, our approach provides reliable consistency signals that overcome the input limits of traditional models. A reward model trained on our generated data achieves competitive accuracy across multiple zero-shot preference benchmarks using only 20% of the typical training data volume. Furthermore, fine-tuning VLMs via Direct Preference Optimization (DPO) based on this data significantly improves performance in handling complex, high-density descriptions. Our results demonstrate the potential of reconstruction-driven self-correction for low-cost, high-precision alignment of multi-modal models. Our code of the scoring model is released at <https://github.com/XC0053/iterative-perceptual-alignment>.

Keywords: Vision-Language Alignment · Iterative Self-Correction · Self-Supervised Dataset Generation

1 Introduction

In recent years, the rapid advancement of vision-language models (VLMs) [3, 30, 62] and text-to-image (T2I) models [12, 40, 48] has increasingly relied on large-scale, high-quality instruction and preference data [9, 31, 62]. As scaling laws [1, 23] drive demand beyond what human annotation can feasibly provide, automated and self-supervised data generation pipelines have become essential [7, 52]. However, this evolution is fundamentally bottlenecked by the lack of solid, highly discriminative supervisory signals during the data generation process. Existing automated metrics, such as the CLIP Score [19, 38], primarily focus on coarse-grained global matching and fail to adequately capture high-density detailed attributes [21, 22, 57]. Conversely, relying on proprietary large

* Work done during an internship at Minus AI. † Corresponding author.

multimodal models (e.g., GPT-4V [35]) or human evaluators introduces significant non-scalability and subjective noise [17, 28, 60]. This ambiguity makes it difficult to convert feedback into stable, deterministic supervisory signals for fine-grained model optimization.

To address this critical lack of solid signals, we propose that the optimal paradigm for evaluating text-to-image alignment—and thereby deriving reliable supervisory signals—is not abstract scoring, but concrete reconstruction. This hypothesis is grounded in a fundamental logical premise: if a generated text accurately parses and aligns with every subtle attribute of a source image, then mapping these semantics back to the visual space via a high-determinism synthesis engine (e.g., utilizing a distilled Diffusion Transformer with few-step deterministic sampling) [33, 37, 41, 48] should yield a reconstruction that is highly consistent with the original image. By transforming the elusive task of abstract semantic scoring into an observable, comparable alignment in the image domain, we establish an objective ground truth. Probabilistic generation inherently carries stochastic disturbances [20, 44], which this mechanism suppresses to a negligible level, converting ambiguous feedback into deterministic causal attributions.

Building on the intuition of "reconstruction as evaluation," we develop a comprehensive generative reconstruction feedback framework. Through an iterative, closed-loop process of generation, feedback, and correction [2, 6, 34, 43], this framework leverages reconstruction fidelity to filter out random noise and capture subtle alignment differences. The specific contributions of this paper are summarized as follows:

- Preference Dataset Construction: We construct a high-quality preference dataset by employing a generative reconstruction feedback loop. By utilizing direct image-to-image comparisons within the visual space as an intuitive metric, we generate preference data pairs with strong causal correlations and high discriminability.
- Scoring Model Training: We train a dedicated scoring model (Reward Model) capable of acutely capturing fine-grained differences in reconstruction fidelity. This model provides stable, objective supervisory signals, effectively mitigating the subjective noise and scalability issues inherent in traditional evaluation methods.
- VLM DPO Validation: We validate the effectiveness of our framework by applying Direct Preference Optimization (DPO) [39] to VLMs using the generated preference data. Experimental results demonstrate that this approach improves alignment accuracy, enabling models to better handle complex descriptions with dense details.

2 Related Work

2.1 Preference Alignment of Vision-Language Models

In recent years, Reinforcement Learning from Human Feedback (RLHF) has been widely applied to multi-modal alignment [46, 55], but its reliance on the

PPO algorithm [42] and explicit reward modeling makes training unstable and costly. Direct Preference Optimization (DPO) [39] was proposed as a lightweight alternative, improving efficiency and stability by parameterizing the implicit reward as the log-probability difference of the policy model. While DPO performs well in multi-modal tasks [13, 26, 27, 49, 59], its alignment quality is fundamentally limited by the quality of the input preference pairs (Input, Chosen, Rejected). We therefore adopt DPO as our final alignment method but shift the research focus forward, to the bottleneck of acquiring high-quality, fine-grained preference data.

2.2 Construction of Multi-modal Preference Datasets

Constructing high-quality preference data is the core of achieving fine-grained alignment for VLMs. Traditional preference dataset construction usually relies on heuristic rules (such as truncating positive samples or cross-image feature replacement) or utilizes existing VLMs for multiple re-samplings [26, 27, 56, 59]. However, these methods face significant limitations: negative samples constructed by heuristic rules often have an excessively large semantic gap compared to positive samples. They lack sufficient gradient discriminability, making it difficult to guide the model in learning fine-grained features. Meanwhile, descriptions sampled based on VLMs are easily restricted by the training data distribution, falling into the semantic homogenization of general patterns. Furthermore, when relying on VLM-as-a-judge to provide supervisory signals, the resulting signal often lacks robustness due to the model’s own hallucinations or inference randomness [17, 50, 60]. To break through these limitations, this paper abandons the traditional discrete negative sample construction method and proposes building a highly discriminative preference trajectory by generating continuous quality gradients.

2.3 Feedback-based Iterative Refinement Mechanisms

To obtain more accurate image-text alignment signals, recent research has begun to explore using external feedback or “understanding by generating” mechanisms to optimize model outputs [34, 43]. For example, some works use text-to-image (T2I) models to reversely reconstruct text into images, thereby verifying the accuracy of the descriptions [10, 21, 29, 45]. However, most existing works still remain at the stage of single evaluations or unstructured random resampling, failing to form an optimization trajectory with a clear evolutionary direction [4, 7, 16, 27, 56]. This paper models the alignment task as a reconstruction-driven optimization problem and designs a closed-loop iterative framework of “reconstruct-compare-refine.” By introducing a strong visual understanding model as a “Visual Critic” and combining it with a low-randomness image generator (Z-Image-Turbo [48]) and perceptual similarity evaluation (DreamSim [15]), this mechanism can perform targeted refinements on fine-grained features such as textures and spatial relationships. This closed loop not only provides a high-quality trajectory of steadily improving alignment (decreasing

perceptual distance) for DPO but also assists in training an efficient image-text consistency regression model, thereby eliminating the reliance on expensive generation steps during the inference stage.

3 Methodology

3.1 Problem Formulation

The fine-grained alignment of Vision-Language Models (VLMs) largely depends on the quality of the paired preference dataset $(I_{ref}, C_{chosen}, C_{rejected})$. Traditional sampling-based construction suffers from three coupled problems noted in Sec. 2.2: semantic homogenization under constrained training distributions, noisy supervision from VLM-as-a-judge evaluators, and insufficient discriminability from heuristic negatives with overly large semantic gaps.

To address these limitations, we model the alignment task as a reconstruction-driven optimization problem. We hypothesize that if a caption C can perfectly represent I_{ref} , then the image $I = g(C)$ reconstructed by a low-randomness text-to-image generator $g(\cdot)$ should be consistent with I_{ref} in the perceptual space. Let $\mathcal{D}(\cdot, \cdot)$ be the perceptual distance metric in the visual space. This optimization problem can be formalized as:

$$C^* = \arg \min_C \mathcal{D}(I_{ref}, g(C)) \quad (1)$$

where $g : \mathcal{C} \rightarrow \mathcal{I}$ is the mapping function from the caption space to the image space. Through a closed-loop iterative framework, we can generate a series of captions with continuous quality gradients, thereby approximating the optimal solution C^* .

3.2 Iterative Construction Pipeline

Initialization Stage (Fig. 1): We use a reference image set $\{I_{ref}\}$. For each reference image I_{ref} , we first utilize four Vision-Language Models (VLMs): Qwen3-VL-8B [3], InternVL3.5-8B [51], Gemma-3n-E4B-it [47], and LLaVA-1.5-7B [30] to obtain multiple candidate text captions $\{C_1, \dots, C_n\}$ of varying quality. Next, we reconstruct each candidate C_i into an image $I_i = g(C_i)$ with the Z-Image-Turbo [48] model and score it by DreamSim, $s_i = h(I_{ref}, I_i)$. The candidate with the best (lowest) score is selected as the initialization caption C_0 , serving as the starting point for iterative refinement. Because Z-Image-Turbo effectively deterministic generation randomness, the visual evaluation accurately reflects caption quality rather than sampling noise. We confirm this empirically: across repeated reconstructions, generator stochasticity is negligible relative to inter-caption score differences, and the resulting preference ordering is preserved when Z-Image-Turbo is replaced by other few-step T2I generators [25, 41]. Full quantitative results for both analyses are reported in Appendix A. Finally, we use the DreamSim model to evaluate the perception-level similarity between the generated and reference images: $s_0 = h(I_{ref}, I_0)$.

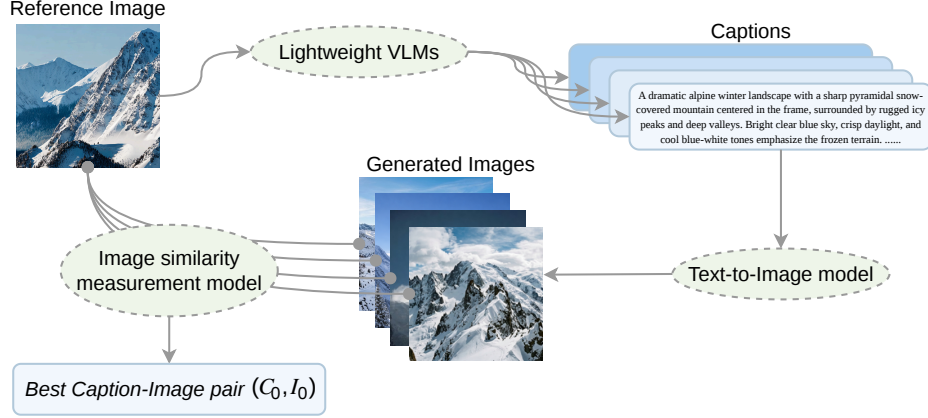


Fig. 1: Initialization stage of the iterative construction pipeline. For each reference image I_{ref} , multiple VLMs of varying scales generate a pool of candidate captions $\{C_1, \dots, C_n\}$. Each candidate is reconstructed into an image via Z-Image-Turbo [48], and DreamSim [15] perceptual similarity scores $s_i = h(I_{\text{ref}}, I_i)$ are computed to select the best-matching caption C_0 as the starting point for iterative refinement.

Closed-loop Semantic Refinement Stage (Fig. 2): After obtaining the initial text caption C_0 and its perceptual score s_0 , the system enters multiple rounds of “reconstruct-compare-refine” closed-loop iterations. The core objective is to progressively eliminate semantic deviations and construct a preference trajectory with high discriminability. Specifically, when the perceptual score s_{k-1} generated in the previous round has not yet reached the preset convergence threshold τ , a visual critic (e.g., GPT-5.1 [36]) takes I_{ref} and the current candidate image I_{k-1} as comparative inputs. Through detailed cross-image comparison, the critic identifies deviations in the previous round’s caption C_{k-1} and performs targeted semantic refinement to generate the next round’s evolved caption C_k . This mechanism ensures a directional optimization driven by visual feedback. An ablation study isolating the contributions of the visual critic and fresh reconstruction signals is provided in Appendix B.

After obtaining the refined caption C_k , the system feeds it back into the Z-Image-Turbo model for image reconstruction:

$$I_k = g(C_k) \quad (2)$$

Then, the DreamSim [15] model updates the perceptual score $s_k = h(I_{\text{ref}}, I_k)$. If $s_k \leq s_{k-1}$ (note that a lower score represents higher similarity), it indicates an effective alignment improvement. The iterative process continues until $s_k \leq \tau$ (reaching a satisfactory alignment level) or until the maximum number of iteration rounds K is reached. Such a closed-loop design ensures that every step of caption improvement is built upon observable visual quality gains, thereby providing a high-quality trajectory with progressive semantics and steadily improving alignment for subsequent preference data extraction.

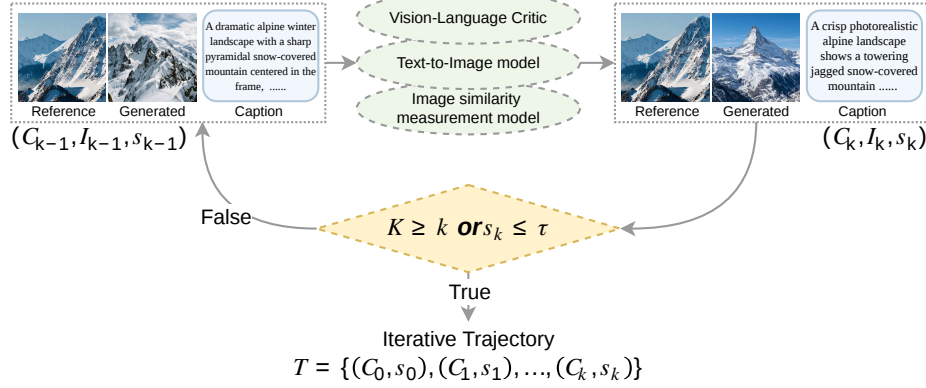


Fig. 2: Closed-loop semantic refinement stage. Given the current caption C_{k-1} and its reconstructed image I_{k-1} , a visual critic identifies semantic deviations by comparing I_{k-1} against I_{ref} and produces a refined caption C_k . The updated reconstruction $I_k = g(C_k)$ is evaluated by DreamSim [15] to update the perceptual score s_k . The loop continues until $s_k \leq \tau$ or the maximum iteration K is reached, yielding a trajectory $\mathcal{T}$ with steadily improving alignment scores.

Preference Pair Construction: After the iterative process ends, the system automatically extracts high-quality preference data by recording the complete iterative trajectory $\mathcal{T} = \{(C_0, s_0), (C_1, s_1), \dots, (C_K, s_K)\}$. This stage constructs preference datasets for different training objectives through two main strategies.

For the Pairwise Preference Dataset used in reward model pre-training, we first divide all samples into training, validation, and test sets based on the reference image I_{ref} . For multiple caption trajectories corresponding to the same I_{ref} , we form preference pairs from their scored iteration points (C_i, s_i) and filter them based on the perceptual score difference:

$$\mathcal{D}_{pair} = \{(C_i, C_j) \mid s_i < s_j - \delta, (C_i, s_i), (C_j, s_j) \in \mathcal{T}\} \quad (3)$$

where $\delta > 0$ is the minimum score interval threshold, used to filter out ambiguous pairs with too small score differences, thereby ensuring the clarity of the preference signal.

For the high-quality preference pairs required by DPO [39], to meet its high requirements for the quality of chosen-rejected pairs, we select the optimal caption (lowest score) and the suboptimal caption (second-lowest score) from the complete trajectory $\mathcal{T}$ of each reference image I_{ref} to form a preference sample pair:

$$(C^{chosen}, C^{rejected}) = (C_{k^*}, C_{m^*}), \quad k^* = \arg \min_k s_k, \quad m^* = \arg \min_{k \neq k^*} s_k \quad (4)$$

Furthermore, to control data quality, we introduce a score interval threshold τ_{dpo} , and only when it satisfies:

$$s_{m^*} - s_{k^*} \geq \tau_{dpo} \quad (5)$$

is the pair included in the final DPO training set. By adjusting τ_{dpo} , an effective trade-off can be made between data volume and preference signal strength. This adjacent-pair strategy ensures that the chosen-rejected pair encodes a fine-grained, non-trivial semantic difference, as overly distant pairs yield trivial distinctions that do not drive fine-grained alignment. The resulting DPO dataset is used to fine-tune vision-language models in Sec. 5.1.

3.3 Image-Text Consistency Scoring Model

We model image-text consistency as a scalar regression task: the scoring model $f_\theta(I_{\text{ref}}, C)$ maps a reference image and a candidate caption to a scalar reward without requiring image generation at inference time.

Backbone. We use a frozen InternVL3.5-1B [51] encoder to extract token-level representations independently for image and text. The image I_{ref} is encoded jointly with a fixed descriptive prompt, yielding image tokens $\mathbf{H}^I \in \mathbb{R}^{T_I \times D}$; the caption C is encoded standalone, yielding text tokens $\mathbf{H}^C \in \mathbb{R}^{T_C \times D}$, where $D = 1024$. No gradients flow into the backbone at any stage.

Cross-Attention Reward Head. The trainable head builds image-conditioned text representations through $L = 4$ stacked cross-attention blocks. In each block, caption tokens serve as queries while image tokens serve as keys and values, following the standard Transformer block design: LayerNorm, multi-head attention with 8 heads, residual connections, and a feed-forward network (FFN) with GELU activation [18], an expansion ratio of 4.0, and a dropout rate of 0.1. A learnable CLS token then aggregates the image-conditioned sequence via one additional cross-attention block, and the resulting vector $\mathbf{v}_{\text{cross}} \in \mathbb{R}^D$ is mapped to a scalar s_{cross} by a 5-layer multi-layer perceptron (MLP) with hidden dimension 2048, where each layer comprises a linear projection, a GELU activation, and dropout.

Score Decomposition. A naive cross-modal scorer may bypass image signals by learning text-quality shortcuts (fluency, length). To mitigate this tendency, we introduce a parallel *text-only branch* with independent parameters: a separate CLS token attends to raw (non-image-conditioned) text tokens and produces a text quality score s_{text} via an independent MLP. The final reward is defined as

$$r = f_\theta(I_{\text{ref}}, C) = s_{\text{cross}}(I_{\text{ref}}, C) - s_{\text{text}}(C), \quad (6)$$

where subtracting s_{text} encourages the model to ground preference signals in image-text correspondence rather than in text quality alone. We note that this design reduces, but does not eliminate, the model’s sensitivity to caption length; we quantify this residual length prior and its effect on downstream evaluation in Sec. 4.3.

Training Objective. The model is trained with a composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_t \mathcal{L}_{\text{text}} + \lambda_c \mathcal{L}_{\text{contrast}} + \lambda_x \mathcal{L}_{\text{cross}}. \quad (7)$$

$\mathcal{L}_{\text{main}}$ is a Bradley–Terry ranking loss [5] on the decomposed score, with s_{text} detached so that gradients flow exclusively through the cross-modal path: $\mathcal{L}_{\text{main}} = \mathbb{E}[\text{softplus}(-(r^+ - r^-))]$. $\mathcal{L}_{\text{text}}$ trains the text-only branch independently on the same training pairs. $\mathcal{L}_{\text{contrast}}$ randomly permutes image tokens within each mini-batch and penalises incorrect-image pairings, providing dense in-batch negative signal. $\mathcal{L}_{\text{cross}}$ operates on cross-image pairs to directly reinforce image–text grounding independently of caption quality. Loss weights and full architecture details are given in Appendix C.

3.4 Direct Preference Optimization

We align the generative model on our constructed preference data with Direct Preference Optimization (DPO), which avoids the explicit reward modeling of RLHF. Under the Bradley–Terry assumption, given a reference image I_{ref} and a chosen/rejected pair (C_w, C_l) , DPO parameterizes the implicit reward as the policy’s log-probability ratio and minimizes:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(C_w | I_{\text{ref}})}{\pi_{\text{ref}}(C_w | I_{\text{ref}})} - \beta \log \frac{\pi_\theta(C_l | I_{\text{ref}})}{\pi_{\text{ref}}(C_l | I_{\text{ref}})} \right) \right] \quad (8)$$

where π_{ref} is the SFT reference policy and β controls the KL regularization strength. The loss raises the relative likelihood of C_w over C_l , achieving preference alignment.

4 Training and Evaluation of the Scoring Model

4.1 Scoring Model Training Details

Training data. We use the NIGHTS dataset [15] as the reference image source, comprising 20,019 images generated by Stable Diffusion v1.4 and v1.5 [40]; 80% of the images are allocated to the training split, with the remainder reserved for validation and test under a strict image-level split to prevent leakage. For each reference image, the iterative pipeline runs for a maximum of $K = 4$ rounds, with early stopping triggered when $s_k < \tau = 0.25$. Each reference image thus accumulates a scored caption pool of 5 to 8 entries in total: the $n = 4$ initial candidates from independent VLMs [3, 30, 47, 51] (each evaluated by DreamSim [15]) and up to $K = 4$ additional captions produced by the critic-guided refinement loop. During preference pair construction, all entries in the same pool are exhaustively paired; pairs with a score gap $\Delta s < \delta = 0.03$ are discarded as ambiguous. The scoring model is then trained on two complementary pair types derived from these pools.

Table 1: Performance on in-domain preference tasks with varying difficulty levels (Accuracy %).

Model	Best-Worst	Best-Second	Best-Second-Thr
T2I-Consistency-Score (Ours)	94.44	60.91	76.8
Cyclereward-Combo [2]	90	55.35	62
Cyclereward-T2I	66.97	52.53	51
Cyclereward-I2T	90	55.35	61.33
HPSv2 [53]	66.87	55.86	62.43
PickScore [24]	51.01	54.04	59.67
LongCLIP-L [58]	87.11	52.63	54.7

Same-image pairs enumerate all ordered pairs (C_i, C_j) from the same reference image satisfying $s_i < s_j - \delta$, yielding **175,691 training triplets** across 15,747 unique images and 31,093 validation triplets across 2,002 unique images.

Cross-image pairs take each image’s highest-quality caption C_i^* and pair it against 3 randomly sampled alternative images $\{I_j\}_{j \neq i}$:

$$\text{correct: } (I_i, C_i^*), \quad \text{wrong: } (I_j, C_i^*), \quad j \neq i, \quad (9)$$

yielding **47,241** additional training triplets. These triplets directly penalise image-text mismatches regardless of caption quality, complementing the same-image pairs which capture caption quality differences.

Architecture and training hyperparameters are detailed in Appendix C.

4.2 Performance on In-domain Preference Tasks

To verify the basic discriminative capability of our Scoring Model, we conduct a systematic evaluation on an in-domain held-out validation set based on 1k reference images. We construct three preference datasets with a gradient distribution of difficulty:

- **Best-Worst:** Selection of description pairs with the maximum reconstruction quality gap to assess fundamental identification of extreme semantic differences.
- **Best-Second:** Selection of pairs ranked first and second in quality to challenge fine-grained discriminability under minimal semantic gaps.
- **Best-Second-Thr:** Built upon the latter, requiring a score difference > 0.03 to eliminate random noise interference during inference.

As shown in Table 1, our T2I-Consistency-Score model exhibits clear advantages across the three difficulty levels compared to baselines like Cyclereward [2] and HPSv2 [53]. In the Best-Worst task, our model achieves a high accuracy of 94.44%. Notably, in the most challenging Best-Second task, while other baseline models generally remain at a low discrimination level (52%–55%), our model

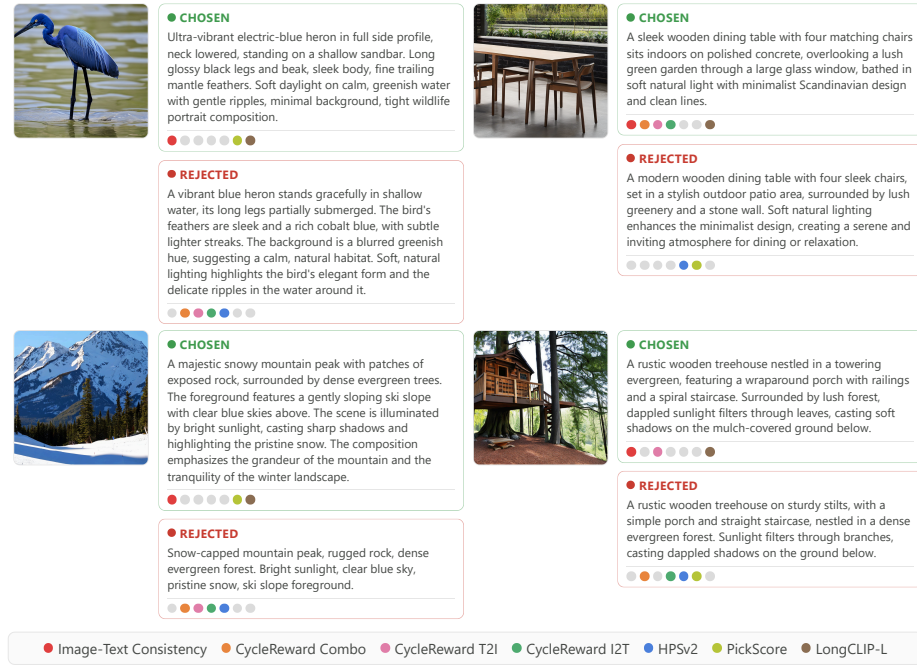


Fig. 3: Qualitative comparison of preference selections across multiple reward models. Each example presents a reference image alongside fine-grained ("CHOSEN") and flawed ("REJECTED") descriptions, with colored dots indicating model preferences. Our model (red dots) consistently and accurately identifies the high-quality descriptions. In contrast, baseline models are frequently misled by oversimplified texts, highlighting the superiority of our reconstruction-driven approach in capturing deep semantic consistency.

maintains a win rate of 60.91%. Furthermore, with the introduction of the threshold, the accuracy on Best-Second-Thr significantly increases to 76.80%. These results demonstrate the effectiveness of reconstruction-driven optimization in capturing fine-grained semantic alignment, providing high-quality supervisory signals for subsequent DPO training.

4.3 Zero-shot Generalization on Classic Benchmarks

To evaluate the robustness of the T2I-Consistency-Score model on unseen distributions, we conducted zero-shot testing on four authoritative public preference benchmark datasets. These include: RLHF-V [55], which focuses on human feedback alignment and hallucination correction; POVID [61], which targets hallucination and perceptual-distortion preferences on real photographs; VLFeedback [27], a large-scale dataset providing fine-grained quality evaluation signals; and RLAIIF-V [56], a visual alignment benchmark generated using AI feedback. For all four benchmarks we restrict evaluation to detailed-description samples

Table 2: Zero-shot generalization performance (Accuracy %) on public preference benchmarks.

Model	RLHF-V	POVID	VLFeedback	RLAIF-V
T2I-Consistency-Score (Ours)	74.12	67.43	75.1	53.47
Cyclereward-Combo [2]	65.25	77.41	59.38	54.09
Cyclereward-T2I	57.3	86.49	39.21	55.46
Cyclereward-I2T	63.52	78.9	54.39	54.22
HPSv2 [53]	63.52	82.69	53.83	53.62
PickScore [24]	59.73	76.88	48.49	51.94
LongCLIP-L [58]	58.55	84.15	58.03	54.33

to focus on fine-grained alignment, though selection differs by dataset. From POVID we keep samples whose instruction is "Describe this image in detail."; for RLHF-V and RLAIF-V we use their native detailed-description splits. These three provide native preference triplets. VLFeedback does not: we select detail-related instructions and form chosen/rejected triplets by Faithfulness score, discarding those with a score difference ≤ 1 . This yields 3,007 / 5,000 / 4,133 / 43,277 triplets for RLHF-V / POVID / VLFeedback / RLAIF-V. We compared our model with SOTA models such as the Cyclereward family [2], HPSv2 [53], PickScore [24], and LongCLIP-L [58].

As shown in Table 2, our model leads on RLHF-V (74.12) and VLFeedback (75.1) but trails Cyclereward variants on POVID and RLAIF-V. This pattern reflects a property of the preference signal rather than a general deficit. Length does not explain it: all four benchmarks use comparably long captions (median 104–136 tokens), and on length-matched pairs our scorer retains clear grounding on RLHF-V and VLFeedback while falling to near chance on POVID and RLAIF-V — the former a distributional mismatch (real-photo hallucinations, closer to Cyclereward’s training domain) and the latter a benchmark with limited resolving power, where all models cluster near chance. A full stratification by chosen/rejected length ratio is provided in Appendix D.

Notably, our scorer was trained on 175,691 same-image preference triplets (plus 47,241 auxiliary cross-image triplets), versus the $\sim 866\text{K}$ used by the closest competitor, Cyclereward-Combo — roughly 20% of its triplet volume. This data efficiency suggests that the reconstruction-driven objective captures image-text consistency more fundamentally.

5 Direct Preference Optimization

5.1 Implementation Details

To comprehensively verify the alignment effectiveness of our proposed reconstruction-driven preference data, we selected three representative open-source vision-language models as base models: Qwen3-VL-2B-Instruct [3], InternVL3.5-1B-HF, and InternVL3.5-2B-HF [51].

Dataset Construction: We utilized the positive and negative samples extracted based on the DreamSim [15] perceptual similarity scores for Direct Preference Optimization (DPO [39]), following the pair construction procedure described in Sec. 3.2:

- **Best-Second threshold > 0.05:** Filtered for high-quality samples with significant semantic differences, yielding a total data size of approximately 3K.
- **Best-Second threshold > 0.03:** Relaxed the semantic gap requirement while expanding the data scale, resulting in a total data size of approximately 5K.

Simultaneously, we sampled 3.5K instances from the public RLHF-V [55] preference dataset as a baseline training set to compare the actual efficacy of automatically constructed data against human-annotated data.

Training Hyperparameter Configuration:

For the three models mentioned above, we adopted a unified DPO [39] training configuration. The training objective used the standard Sigmoid preference loss function, and the KL penalty coefficient (β) controlling the model deviation was set to 0.1. The models underwent full-parameter fine-tuning for only 1 epoch to prevent overfitting; the learning rate was strictly set to 5×10^{-6} , combined with a cosine annealing strategy (Cosine Scheduler) and a 10% warmup ratio. For hardware optimization, the per-device batch size was set to 4, the gradient accumulation steps were set to 8, and gradient checkpointing was globally enabled to reduce memory consumption.

5.2 Quantitative Results on Benchmarks

To comprehensively evaluate the multimodal capability improvements brought by DPO [39] training, we conducted quantitative evaluations on multiple authoritative vision-language benchmarks. These include: POPE [28], focusing on object hallucination; Hallusion Benchmark [17], testing resistance to visual illusions and logical hallucinations; MME [14], a multi-dimensional set covering perception and commonsense reasoning; RealWorldQA [54], focusing on spatial and physical understanding in real-world scenarios; MM-Bench [32], providing fine-grained capability dimension divisions; and MMStar [8], aimed at filtering data contamination and evaluating core visual capabilities. Specific results are shown in Table 3.

1. Overall Performance Improvement: Models aligned through DPO achieve consistent improvements on the two 2B-scale backbones across most metrics. For example, after fine-tuning Qwen3-VL-2B-Instruct with our Thr > 0.03 dataset, its Hallusion score increased from 46.27 to 57.83, alongside growth on MM-Bench. This demonstrates that DPO effectively alleviates base-model deficiencies in fine-grained alignment; the backbone-specific behaviour of the 1B model is analysed in point 4 below.



Fig. 4: Qualitative comparison of generation results before and after DPO [39] alignment. DPO-aligned models (right) produce captions with more precise fine-grained attributes (textures, colors, spatial relationships), yielding reconstructed images that more closely match the reference I_{ref} .

2. Automated Data vs. Human-Annotated Data: Notably, our extremely small-scale automated datasets (3K and 5K) exhibited strong competitiveness against human-annotated data. Their performance matched or even surpassed models trained on 3.5K real human-annotated data (RLHF-V) across multiple core metrics. For instance, on InternVL3.5-1B-HF, our 3K dataset ($Thr > 0.05$) achieved better performance than RLHF-V on RealWorldQA (0.52 vs 0.51) and MMStar (0.34 vs 0.33), while remaining within 0.01 of RLHF-V on MM-Bench (0.49 vs 0.50). This indicates that reconstruction-driven preference scoring provides more precise fine-grained visual alignment gradients.

3. Trade-off Between Data Scale and Quality Threshold: Experiments revealed a trade-off between the discrimination threshold and data volume. The $Thr > 0.05$ (3K) dataset, featuring a higher discrimination threshold, performed better in comprehensive evaluations requiring fine-grained distinction, such as MM-Bench on Qwen3 (0.70 vs 0.67 for the 0.03 threshold). Conversely, the larger

Table 3: Quantitative comparison of DPO [39] alignment performance across multimodal benchmarks. “Ours (> 0.05)” and “Ours (> 0.03)” denote the DPO models trained on our automatically constructed datasets with different discrimination thresholds. Best results within each model group are highlighted in bold. POPE [28] and Hallusion [17] are reported as accuracy (%); RealWorldQA [54], MM-Bench [32], and MMStar [8] as accuracy (0–1); MME-P [14] is the perception score (max 2000); MME-C is the commonsense reasoning subscore (max 200). All benchmarks are evaluated with VLMEvalKit [11].

Model	POPE	Hallusion	MME-P	MME-C	RealWorldQA	MM-Bench	MMStar
Qwen3-VL-2B-Instruct							
Base (No DPO)	88.85	46.27	1510.46	125.71	0.66	0.66	0.42
+ Ours (> 0.05)	89.26	56.99	1486.91	121.43	0.63	0.70	0.45
+ Ours (> 0.03)	89.32	57.83	1493.92	121.43	0.63	0.67	0.45
+ RLHF-V [55] (3.5k)	89.55	56.15	1478.92	127.86	0.64	0.69	0.46
InternVL3.5-1B-HF							
Base (No DPO)	85.95	54.47	1369.87	112.86	0.45	0.32	0.25
+ Ours (> 0.05)	85.12	51.10	1357.73	117.14	0.52	0.49	0.34
+ Ours (> 0.03)	85.00	51.74	1359.83	117.14	0.52	0.49	0.33
+ RLHF-V (3.5k)	85.03	51.31	1351.56	115.00	0.51	0.50	0.33
InternVL3.5-2B-HF							
Base (No DPO)	86.67	59.62	1498.41	129.29	0.39	0.35	0.33
+ Ours (> 0.05)	88.66	60.15	1532.34	135.71	0.35	0.42	0.36
+ Ours (> 0.03)	88.54	60.99	1518.25	136.43	0.36	0.42	0.36
+ RLHF-V (3.5k)	88.56	60.46	1518.54	138.57	0.35	0.43	0.38

$Thr > 0.03$ (5K) dataset held an advantage in tasks assessing overall generalization and anti-hallucination, such as Hallusion Benchmark on Qwen3 (57.83 vs 56.99).

4. Analysis of InternVL3.5-1B Behaviour. InternVL3.5-1B exhibits a distinctly split response to DPO training. On hallucination-specific benchmarks, all three datasets incur mild regressions relative to the base model: POPE drops by 0.83–0.95 and Hallusion by 2.73–3.37, a pattern shared by RLHF-V [55] (−0.92 on POPE, −3.16 on Hallusion), indicating that the degradation is not unique to our data construction. By contrast, multi-modal comprehension benchmarks improve substantially: MM-Bench rises from 0.32 to 0.49 (+0.17), MM-Star from 0.25 to 0.34 (+0.09), and RealWorldQA from 0.45 to 0.52 (+0.07). We attribute this split to a capacity constraint: at 1B parameters, DPO training re-allocates representational capacity toward richer fine-grained visual descriptions, improving tasks that demand detailed attribute reasoning while disrupting the coarser yes/no calibration required by POPE and the subtle visual-linguistic coherence tested by Hallusion. The near-identical degradation profile of RLHF-V on these two benchmarks (−0.92 and −3.16) further supports the view that the regression is a model-size artefact rather than a data-quality issue.

5. Consistent Improvements on InternVL3.5-2B. In contrast to the mixed pattern observed on the 1B model, DPO training on InternVL3.5-2B yields consistent gains across six of seven benchmarks under all three training

datasets. Our $Thr > 0.05$ dataset achieves the largest gains on POPE (+1.99) and MME (+33.93), while $Thr > 0.03$ leads on Hallusion (+1.37), collectively demonstrating that the 2B model has sufficient capacity to absorb preference alignment without sacrificing existing capabilities. The sole exception is Real-WorldQA, which decreases marginally (-0.03 to -0.04) across all three methods including RLHF-V [55]. Consistent with the distributional analysis in Sec. 4.3, this marginal decrease reflects a dataset-level domain mismatch rather than a DPO-specific artefact: our synthetic, densely-described NIGHTS [15] training data transfers less well to real-photo spatial reasoning, as also shared by RLHF-V (-0.03 to -0.04). Against RLHF-V [55], our data remains competitive: $Thr > 0.05$ leads on POPE (88.66 vs. 88.56) and MME (1532.34 vs. 1518.54), while $Thr > 0.03$ outperforms RLHF-V on Hallusion (60.99 vs. 60.46), confirming that automatically constructed reconstruction-driven preference data can match human-annotated quality at a fraction of the annotation cost.

6 Discussion

In summary, the reconstruction-driven alignment method proposed in this paper significantly reduces the high cost of manual annotation while successfully guiding vision-language models to mine latent visual details—such as subtle perspective differences, environmental tones, and lighting changes—that are easily overlooked in traditional human or conventional preference data. However, this closed-loop feedback paradigm still faces several inherent challenges. On one hand, the inherent inference randomness of current Text-to-Image (T2I) base models introduces a certain degree of uncontrollable noise into the reconstruction process, thereby affecting the stability of fine-grained alignment gradients. On the other hand, the generalization ability of visual consistency evaluation models acting as “judges” (such as DreamSim [15] or our scoring model) needs further improvement when dealing with more extreme or out-of-distribution (OOD) complex cross-domain scenarios. Future research could explore the introduction of image generators with stronger determinism (e.g., control-conditioned diffusion models) and focus on building more robust, universal multimodal evaluation metrics to promote the application of reconstruction alignment techniques in broader scenarios.

References

1. Aghajanyan, A., Yu, L., Conneau, A., Hsu, W.N., Hambardzumyan, K., Zhang, S., Roller, S., Goyal, N., Levy, O., Zettlemoyer, L.: Scaling laws for generative mixed-modal language models (2023), <https://arxiv.org/abs/2301.03728> 1
2. Bahng, H., Chan, C., Durand, F., Isola, P.: Cycle consistency as reward: Learning image-text alignment without human preferences (2025), <https://arxiv.org/abs/2506.02095> 2, 9, 11
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li,

- Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), <https://arxiv.org/abs/2511.21631> 1, 4, 8, 11
4. Bai, S., Yang, S., Bai, J., Wang, P., Zhang, X., Lin, J., Wang, X., Zhou, C., Zhou, J.: Touchstone: Evaluating vision-language models by language models (2023), <https://arxiv.org/abs/2308.16890> 3
 5. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39**(3/4), 324–345 (1952), <http://www.jstor.org/stable/2334029> 8
 6. Chen, D., Chen, R., Zhang, S., Liu, Y., Wang, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark (2024), <https://arxiv.org/abs/2402.04788> 2
 7. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions (2023), <https://arxiv.org/abs/2311.12793> 1, 3
 8. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? (2024), <https://arxiv.org/abs/2403.20330> 12, 14
 9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023), <https://arxiv.org/abs/2305.06500> 1
 10. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: Dreamllm: Synergistic multimodal comprehension and creation (2024), <https://arxiv.org/abs/2309.11499> 3
 11. Duan, H., Fang, X., Yang, J., Zhao, X., Qiao, Y., Li, M., Agarwal, A., Chen, Z., Chen, L., Liu, Y., Ma, Y., Sun, H., Zhang, Y., Lu, S., Wong, T.H., Wang, W., Zhou, P., Li, X., Fu, C., Cui, J., Chen, J., Song, E., Mao, S., Ding, S., Liang, T., Zhang, Z., Dong, X., Zang, Y., Zhang, P., Wang, J., Lin, D., Chen, K.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models (2025), <https://arxiv.org/abs/2407.11691> 14
 12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), <https://arxiv.org/abs/2403.03206> 1
 13. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D.: Kto: Model alignment as prospect theoretic optimization (2024), <https://arxiv.org/abs/2402.01306> 3
 14. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), <https://arxiv.org/abs/2306.13394> 12, 14
 15. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: *Advances in Neural Information Processing Systems*. vol. 36, pp. 50742–50768 (2023), <https://neurips.cc/virtual/2023/poster/72380> 3, 5, 6, 8, 12, 15
 16. Gao, L., Schulman, J., Hilton, J.: Scaling laws for reward model overoptimization (2022), <https://arxiv.org/abs/2210.10760> 3

17. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models (2024), <https://arxiv.org/abs/2310.14566> 2, 3, 12, 14
18. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus) (2023), <https://arxiv.org/abs/1606.08415> 7
19. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning (2022), <https://arxiv.org/abs/2104.08718> 1
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020), <https://arxiv.org/abs/2006.11239> 2
21. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering (2023), <https://arxiv.org/abs/2303.11897> 1, 3
22. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* **36**, 78723–78747 (2023), <https://neurips.cc/virtual/2023/poster/73424> 1
23. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020), <https://arxiv.org/abs/2001.08361> 1
24. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation (2023), <https://arxiv.org/abs/2305.01569> 9, 11
25. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), <https://arxiv.org/abs/2506.15742> 4
26. Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L.: Silkie: Preference distillation for large visual language models (2023), <https://arxiv.org/abs/2312.10665> 3
27. Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L., Liu, Q.: Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment (2024), <https://arxiv.org/abs/2410.09421> 3, 10
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models (2023), <https://arxiv.org/abs/2305.10355> 2, 12, 14
29. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation (2024), <https://arxiv.org/abs/2404.01291> 3
30. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024), <https://arxiv.org/abs/2310.03744> 1, 4, 8
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), <https://arxiv.org/abs/2304.08485> 1
32. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), <https://arxiv.org/abs/2307.06281> 12, 14
33. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023), <https://arxiv.org/abs/2310.04378> 2

34. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoy, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023), <https://arxiv.org/abs/2303.17651> 2, 3
35. OpenAI: Gpt-4v(ision) system card (2023), <https://openai.com/index/gpt-4v-system-card/> 2
36. OpenAI: GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum (2025), <https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/> 5
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), <https://arxiv.org/abs/2212.09748> 2
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), <https://arxiv.org/abs/2103.00020> 1
39. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model (2024), <https://arxiv.org/abs/2305.18290> 2, 3, 6, 12, 13, 14
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), <https://arxiv.org/abs/2112.10752> 1, 8
41. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation (2023), <https://arxiv.org/abs/2311.17042> 2, 4
42. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), <https://arxiv.org/abs/1707.06347> 3
43. Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning (2023), <https://arxiv.org/abs/2303.11366> 2, 3
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations (2021), <https://arxiv.org/abs/2011.13456> 2
45. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality (2024), <https://arxiv.org/abs/2307.05222> 3
46. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented rlhf (2023), <https://arxiv.org/abs/2309.14525> 2
47. Team, G.: Gemma 3n (2025), <https://ai.google.dev/gemma/docs/gemma-3n> 4, 8
48. Team, I., Cai, H., Cao, S., Du, R., Gao, P., Hao, A., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jiang, Y., Jin, X., Li, L., Li, Z., Li, Z.Y., Liu, D., Liu, D., Wu, Q., Yu, F., Zhan, Z., Zhang, C., Zhang, S., Zhou, R., Zhou, S.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer (2026), <https://arxiv.org/abs/2511.22699> 1, 2, 3, 4, 5
49. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization (2023), <https://arxiv.org/abs/2311.12908> 3
50. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., Sang, J.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation (2024), <https://arxiv.org/abs/2311.07397> 3

51. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), <https://arxiv.org/abs/2508.18265> 4, 7, 8, 11
52. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions (2023), <https://arxiv.org/abs/2212.10560> 1
53. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis (2023), <https://arxiv.org/abs/2306.09341> 9, 11
54. xAI: Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v> (2024), accessed: 2024-04-12 12, 14
55. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., Chua, T.S.: Rlhf-v: Towards trustworthy mlms via behavior alignment from fine-grained correctional human feedback (2024), <https://arxiv.org/abs/2312.00849> 2, 10, 12, 14, 15
56. Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., Chen, D., Lu, X., Cui, G., Dang, Y., He, T., Feng, X., Song, J., Zheng, B., Liu, Z., Chua, T.S., Sun, M.: Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness (2025), <https://arxiv.org/abs/2405.17220> 3, 10
57. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? (2023), <https://arxiv.org/abs/2210.01936> 1
58. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip (2024), <https://arxiv.org/abs/2403.15378> 9, 11
59. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization (2024), <https://arxiv.org/abs/2311.16839> 3
60. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023), <https://arxiv.org/abs/2306.05685> 2, 3
61. Zhou, Y., Cui, C., Rafailov, R., Finn, C., Yao, H.: Aligning modalities in vision large language models via preference fine-tuning (2024), <https://arxiv.org/abs/2402.11411> 10
62. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023), <https://arxiv.org/abs/2304.10592> 1