CROSS-FEATURE GUIDED REINFORCEMENT LEARNING FOR UNIFIED MULTIMODAL UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) integrate multimodal understanding and image generation, but joint reinforcement learning (RL) often suffers from optimization conflict because the two branches are optimized together with limited interaction. To address this issue, we propose **Cross-Feature Guided Reinforcement Learning (CGRL)**, which enables reciprocal guidance between the two branches during joint RL. Leveraging the semantic correspondence between paired images and captions, CGRL introduces cross-feature guidance by comparing each branch's source representation with the generated token hidden states of the opposite branch at the same backbone layer. The resulting guidance modulates the sequence-level advantage to obtain token-level advantages, allowing more precise policy updates. Experiments on Janus-Pro and Show-o, covering pure autoregressive and hybrid autoregressive–diffusion architectures, show consistent improvements over joint GRPO in multimodal understanding, image captioning, and text-to-image generation. Further gradient analysis reveals increased cross-branch gradient agreement. These results demonstrate that CGRL reduces the optimization conflict between understanding and generation and enables mutually beneficial joint RL in unified multimodal models.

1 INTRODUCTION

Recent unified multimodal models (UMMs) have achieved remarkable progress in both multimodal understanding and generation (Wu et al., 2025a; Zhou et al., 2025; Chen et al., 2025a; Pan et al., 2025; Wang et al., 2024; Deng et al., 2025; Chameleon Team, 2024). By integrating perception and creation within a single model, UMMs provide a promising foundation for more general multimodal intelligence. More importantly, unification is expected to enable understanding and generation to benefit from each other. However, jointly improving both capabilities remains challenging.

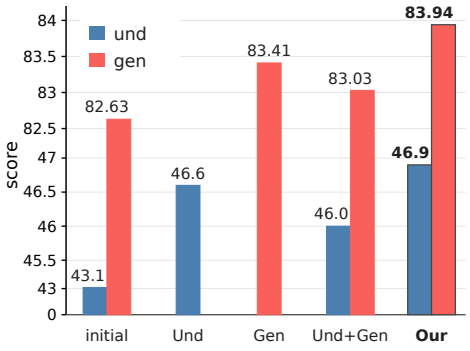


Figure 1: Joint GRPO performs worse than task-specific RL on both understanding and generation. CGRL improves both capabilities and outperforms task-specific RL. Results are reported on Janus-Pro 1.5B Chen et al. (2025b), with MMStar for understanding and DPG-Bench for generation.

Understanding and generation impose different optimization demands and can interfere when jointly trained through shared parameters (Wu et al., 2025a; Lu et al., 2026; Zhang & Tang, 2026). This conflict persists during reinforcement learning (RL). As shown in Figure 1, joint GRPO (Shao et al., 2024) underperforms task-specific RL on both capabilities. We attribute this limitation to two aspects: (1) joint training optimizes understanding and generation simultaneously yet largely independently, without an explicit mechanism for one task to inform the other's updates. (2) GRPO applies a sequence-level advantage uniformly to all tokens, lacking fine-grained token-level supervision. Together, these limitations restrict both cross-task interaction and fine-grained supervision,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

raising a key question: *how can understanding and generation effectively benefit from each other during RL?*

Our motivation is that an image and its paired caption describe the same semantic content in different modalities, naturally linking the two tasks. During joint RL, the understanding branch describes the image, while the generation branch synthesizes an image from the paired caption. This correspondence suggests that the internal representation of one branch can provide useful guidance for the optimization of the other.

Based on this insight, we propose **Cross-Feature Guided Reinforcement Learning (CGRL)**, which connects understanding and generation through cross-feature guidance. CGRL uses image representations from the understanding branch to guide the generation branch, while caption representations from the generation branch guide the understanding branch. CGRL computes same-layer similarities between opposite-branch input representations and generated-token hidden states, converts them into token-level weights, and uses these weights to modulate the sequence-level GRPO advantage. In this way, the two branches directly influence each other's policy updates while receiving finer-grained supervision.

We evaluate CGRL on Janus-Pro and Show-o, covering pure autoregressive and hybrid autoregressive–diffusion architectures. The evaluation spans diverse benchmarks across three categories: multimodal understanding, image captioning, and text-to-image generation. Across both architectures, CGRL consistently outperforms joint GRPO, while also achieving higher alignment rewards and stronger cross-branch gradient agreement. These results show that cross-feature guidance reduces the optimization conflict between understanding and generation and promotes mutually beneficial joint RL.

Our contributions are summarized as follows:

- We identify task conflict in joint RL for UMMs, where joint GRPO underperforms task-specific RL on both understanding and generation.
- We propose **CGRL**, which introduces token-level cross-feature guidance to modulate sequence-level advantages and enable reciprocal optimization between understanding and generation.
- We validate the effectiveness of CGRL across different UMM architectures through extensive experiments on diverse benchmarks.

2 Related Work

2.1 Unified Understanding and Generation Models

Existing UMMs broadly follow four architectural paradigms. *Pure autoregressive architectures* model both text and image tokens through next-token prediction, including Chameleon, Emu3, Janus, and Janus-Pro (Chameleon Team, 2024; Wang et al., 2024; Wu et al., 2025a; Chen et al., 2025b). *Cascaded architectures*, such as NExT-GPT and SEED-X (Wu et al., 2024; Ge et al., 2024), connect a multimodal language model to a separate diffusion-based image generator. *Hybrid architectures* combine autoregressive text prediction with diffusion- or flow-based image generation within a unified framework, as in Transfusion, Show-o, and BAGEL (Zhou et al., 2025; Xie et al., 2025; Deng et al., 2025). *Pure diffusion architectures*, including UniDisc and MMaDA (Swerdlow et al., 2025; Yang et al., 2025a), model both textual and visual outputs with diffusion. Since cascaded architectures still depend on external generation modules and pure diffusion architectures remain at an early stage of development, we focus on two representative unified paradigms: pure autoregressive architectures with Janus-Pro and hybrid autoregressive–diffusion architectures with Show-o. Their distinct generation mechanisms also allow us to examine whether CGRL generalizes across different UMM designs.

2.2 Post-Training Unified Multimodal Models

Recent post-training methods improve the interaction between understanding and generation through several forms of cross-task coupling. One line of work couples the two capabilities through preferences or reward signals. HermesFlow (Yang et al., 2025b) optimizes paired preferences with

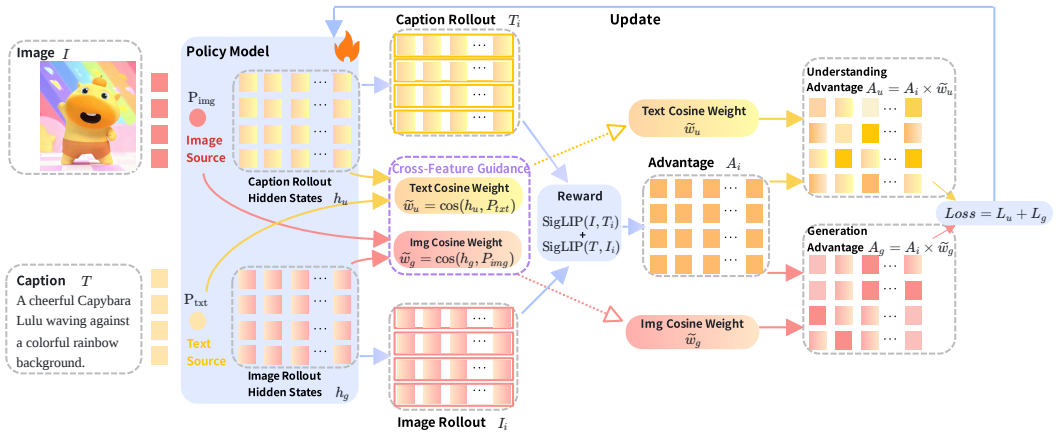


Figure 2: **Overview of CGRL.** Given an image–caption pair (I, T), the model samples caption and image rollouts. SigLIP2 rewards from both branches are summed and group-normalized to obtain a sequence-level advantage A_i. Cross-feature guidance computes similarities between rollout hidden states and source representations from the opposite branch. The resulting token-level cross-feature guidance weights $\widetilde{w}_u$ and $\widetilde{w}_g$ modulate A_i for the joint policy update.

Pair-DPO and iterative self-play, while CoRL (Jiang et al., 2025) and Unified-GRPO (Yan et al., 2026) introduce cycle-consistency, image–text matching, or reconstruction-based rewards for joint optimization. UniRL (Mao et al., 2025) further uses generated images as feedback for understanding through answer supervision. Another line explicitly incorporates one capability into the training process of the other: R3 (Ye et al., 2026) introduces reasoning, reflection, and refinement into image generation, while UniMRG (Su et al., 2026) improves understanding through auxiliary visual generation tasks. At the representation level, DIVA (Lu et al., 2026) decomposes intermediate features into shared and task-specific components to facilitate cross-task transfer. Despite these advances, most existing approaches couple the two tasks through outputs, rewards, or auxiliary representation objectives, while fine-grained interaction during policy updates remains underexplored. CGRL addresses this gap by using cross-feature guidance to let the two branches directly influence each other's policy updates through token-level modulation of sequence-level advantages.

3 Method

3.1 Preliminaries

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) extends Proximal Policy Optimization (PPO) (Schulman et al., 2017) by removing the critic network and estimating relative advantages from groups of sampled outputs. Given a conditioning input x, the sampling policy π_{old} produces K rollouts $\{o_i\}_{i=1}^{K}$ with rewards $\{r_i\}_{i=1}^{K}$. The group-relative advantage is obtained by group normalization:

$$A_i = \frac{r_i - \text{Mean}(\{r_1, r_2, ..., r_K\})}{\text{Std}(\{r_1, r_2, ..., r_K\})} \tag{1}$$

After estimating the group-relative advantages $\{A_i\}_{i=1}^{K}$, GRPO updates the policy by maximizing the following objective:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\{o_i\}_{i=1}^{K} \sim \pi_{\boldsymbol{\theta}_{\text{old}}}} \left[\frac{1}{K} \sum_{i=1}^{K} \left(\frac{\pi_{\boldsymbol{\theta}}(o_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(o_i)} A_i - \beta D_{\text{KL}}\left(\pi_{\boldsymbol{\theta}} \,\|\, \pi_{\text{ref}}\right) \right) \right] \tag{2}$$

Here, $\pi_{\boldsymbol{\theta}}$ denotes the policy. The KL term constrains the updated policy toward the reference policy π_{ref}, with β controlling the regularization coefficient.

3.2 CROSS-FEATURE GUIDED REINFORCEMENT LEARNING

We propose CGRL, which uses each branch's internal representations to guide the other branch's policy updates for joint RL of image captioning and text-to-image generation. As illustrated in Figure 2, CGRL follows four stages. First, the two branches sample caption and image rollouts, respectively. Second, SigLIP2 scores the alignment between each output and its conditioning input, and the branch rewards are combined into a sequence-level advantage. Third, *cross-feature guidance* computes cosine similarities between generated-token hidden states and opposite-branch input representations from the same backbone layer. Finally, these similarities are converted into token-level weights that modulate the sequence-level advantage during joint policy update.

Text–Image Alignment Reward. To encourage the generated outputs to remain semantically aligned with their conditioning inputs, we use SigLIP2 as the reward model S_ϕ. Given an image–caption pair (I, T), the understanding branch samples K caption rollouts $T_i \sim \pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot \mid I)$, while the generation branch samples K image rollouts $I_i \sim \pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot \mid T)$. For caption and image rollouts paired by index i, the joint reward is:

$$r_i = S_\phi(I, T_i) + S_\phi(T, I_i). \tag{3}$$

The rewards $\{r_i\}_{i=1}^K$ are group-normalized following Eq. 1 to obtain the sequence-level advantages $\{A_i\}_{i=1}^K$. Table 4 further compares joint and separate reward formulations.

Cross-Feature Guidance. Cross-feature guidance uses the source representation from one branch to guide token-level policy updates in the opposite branch. Let $\boldsymbol{h}^I_{u,j}$ and $\boldsymbol{h}^T_{g,j}$ denote the input-image and input-caption hidden states in the understanding and generation branches, respectively. We mean-pool the image and caption hidden states to obtain the corresponding source representations:

$$\mathbf{P}_{\text{img}} = \frac{1}{N_I}\sum_{j=1}^{N_I} \boldsymbol{h}^I_{u,j}, \qquad \mathbf{P}_{\text{txt}} = \frac{1}{N_T}\sum_{j=1}^{N_T} \boldsymbol{h}^T_{g,j}, \tag{4}$$

where N_I and N_T denote the numbers of image and caption tokens, respectively.

For rollout i, let $\boldsymbol{h}_{u,i,t}$ and $\boldsymbol{h}_{g,i,t}$ denote the hidden states of the t-th generated caption and visual token, respectively. All features are extracted from the same backbone layer to ensure a consistent feature space for similarity computation. The cosine similarity scores are:

$$w_{u,i,t} = \frac{\boldsymbol{h}^\top_{u,i,t}\mathbf{P}_{\text{txt}}}{\|\boldsymbol{h}_{u,i,t}\|_2\|\mathbf{P}_{\text{txt}}\|_2}, \qquad w_{g,i,t} = \frac{\boldsymbol{h}^\top_{g,i,t}\mathbf{P}_{\text{img}}}{\|\boldsymbol{h}_{g,i,t}\|_2\|\mathbf{P}_{\text{img}}\|_2}. \tag{5}$$

These similarities measure the correspondence between each generated token and the opposite branch's same-modality source representation.

Advantage Modulation. We transform the cosine similarity scores $w_{b,i,t}$ into cross feature guidance weights $\widetilde{w}_{b,i,t}$ to derive token-level advantages. Let $N_{b,i}$ denote the number of generated tokens in rollout i of branch $b \in \{u, g\}$. We first apply min max scaling within each rollout to normalize the similarity scores and stabilize the subsequent modulation:

$$\widehat{w}_{b,i,t} = \frac{w_{b,i,t} - \min_j w_{b,i,j}}{\max_j w_{b,i,j} - \min_j w_{b,i,j} + \epsilon_s}. \tag{6}$$

We then zero center the normalized scores and clip them into bounded guidance weights:

$$\widetilde{w}_{b,i,t} = \text{sg}\Big[\text{clip}\Big(1 + \gamma\left(\widehat{w}_{b,i,t} - \overline{\widehat{w}}_{b,i}\right), w_{\min}, w_{\max}\Big)\Big], \qquad \overline{\widehat{w}}_{b,i} = \frac{1}{N_{b,i}}\sum_{j=1}^{N_{b,i}} \widehat{w}_{b,i,j}. \tag{7}$$

Here, ϵ_s ensures numerical stability, γ controls the modulation strength, $[w_{\min}, w_{\max}]$ bounds the guidance weights, and sg denotes stop gradient. The modulated advantage is:

$$A_{b,i,t} = A_i \widetilde{w}_{b,i,t}. \tag{8}$$

Since $0 < w_{\min} < 1 < w_{\max}$, the guidance weights remain positive and preserve the sign of A_i while scaling each token's contribution. Appendix A.5 visualizes the resulting cross feature guidance weights. For Show-o, token similarities are averaged within each denoising transition before computing transition level guidance weights; details are provided in Appendix A.3.

Policy Update. Let $o_{b,i,t}$ denote the policy decision at step t of rollout i for branch b, and let $x_{b,i,t}$ denote its conditioning context. The policy ratio between the current and sampling policies is:

$$\rho_{b,i,t}(\boldsymbol{\theta}) = \frac{\pi^b_{\boldsymbol{\theta}}(o_{b,i,t} \mid x_{b,i,t})}{\pi^b_{\boldsymbol{\theta}_{\text{old}}}(o_{b,i,t} \mid x_{b,i,t})}. \tag{9}$$

Following CoRL (Jiang et al., 2025), we omit the KL penalty. We optimize each branch with the GRPO objective using the modulated advantages:

$$\mathcal{L}_b(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_{b,i}}\sum_{t=1}^{N_{b,i}}\min\{\rho_{b,i,t}A_{b,i,t}, \text{clip}(\rho_{b,i,t}, 1-\varepsilon, 1+\varepsilon)\, A_{b,i,t}\}\right]. \tag{10}$$

Here, ε is the clipping coefficient, and the expectation is over training pairs and sampled rollouts. Finally, the CGRL objective is defined as:

$$\mathcal{L}_{\text{CGRL}}(\boldsymbol{\theta}) = \mathcal{L}_u(\boldsymbol{\theta}) + \mathcal{L}_g(\boldsymbol{\theta}). \tag{11}$$

4 Experiments

4.1 Experimental Setup

Models and Training Tasks. We evaluate CGRL on two representative UMM architectures using image captioning and text-to-image generation as the joint RL tasks. Janus-Pro 1.5B and 7B (Chen et al., 2025b) model both text and visual tokens autoregressively, while Show-o 1.3B (Xie et al., 2025) combines autoregressive text generation with discrete diffusion for image generation. We use joint GRPO as the baseline, following the joint RL setup in CoRL (Jiang et al., 2025).

Implementation Details. All experiments are conducted for one epoch on 22K COCO-derived image–caption pairs[1] curated by CoRL (Jiang et al., 2025), with a global batch size of 16. We sample eight rollouts per input and use SigLIP2-SO400M/14@384 (Tschannen et al., 2025) as the reward model. Image generation uses classifier-free guidance with a scale of 5. Janus-Pro applies GRPO to both branches, while Show-o uses GRPO for text generation and Mask-GRPO (Luo et al., 2025) for image generation with 18 denoising steps. We select the guidance layer with the highest text–image cosine similarity score: layer 5 for Janus-Pro 1.5B, layer 10 for Janus-Pro 7B, and layer 11 for Show-o 1.3B. For all models, the visual modules are frozen and only the language model is optimized. Janus-Pro 1.5B and Show-o 1.3B use a learning rate of 2×10^{-6}, while Janus-Pro 7B is trained with LoRA (Hu et al., 2021) using a learning rate of 1×10^{-5}. Figure 5 and Appendix A.2 provide the layer-selection curves and detailed training configurations.

Evaluation Benchmarks. We evaluate CGRL across diverse benchmarks covering multimodal understanding, text-to-image generation, and image captioning. For multimodal understanding, we use MMStar (Chen et al., 2024b), MMBench (Liu et al., 2024b), MMMU (Yue et al., 2024), MMVP (Tong et al., 2024b), CV-Bench-2D and CV-Bench-3D (Tong et al., 2024a), and MMT-Bench (Ying et al., 2024). Text-to-image generation is evaluated on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and WISE (Niu et al., 2026). For image captioning, COCO and NoCaps assess in-domain and out-of-domain alignment using CLIP-L/14@336 (Radford et al., 2021) and SigLIP-L/16@384 (Zhai et al., 2023). Appendix A.1 provides details on the benchmarks and evaluation metrics.

4.2 Main Results

Multimodal Understanding. Although the understanding branch is trained only on image captioning without visual question answering (VQA) supervision, CGRL consistently improves VQA performance. As shown in Table 1, CGRL outperforms joint GRPO in 19 of 20 comparisons. On Janus-Pro 7B, MMMU increases from 40.6 to 45.3. These results suggest that cross-feature guidance strengthens visual understanding learned through captioning and transfers effectively to downstream VQA tasks.

[1] https://huggingface.co/datasets/mm-vl/x2x_rft_22k

Table 1: **Results on multimodal understanding benchmarks.** The best performance within each backbone block is highlighted in **bold**. † denotes results obtained from our evaluation.

Model	Params	MMStar	MMBench	MMMU	MMVP	CV-Bench		MMT-Bench
						2D	3D	
Understanding Only								
InstructBLIP (Dai et al., 2023)	7B	32.7	37.4	30.6	–	–	–	35.1
LLaVA-v1.5 (Liu et al., 2024a)	7B	33.1	64.3	35.7	–	**56.8**	**58.4**	48.9
Qwen2.5-VL (Bai et al., 2025)	3B	56.3	**79.8**	**51.2**	–	–	–	61.0
Ovis2 (Lu et al., 2024)	2B	**56.7**	–	45.6	–	–	–	**61.7**
SmolVLM (Marafioti et al., 2025)	1.7B	41.7	–	38.8	–	–	–	51.5
Unified Understanding and Generation								
Chameleon (Chameleon Team, 2024)	7B	31.1	35.7	22.4	–	–	–	23.9
Harmon (Wu et al., 2025b)	1.5B	35.3	50.4	38.9	–	–	–	–
Emu3-Chat (Wang et al., 2024)	8B	**46.6**	64.4	33.9	–	–	–	**43.4**
VILA-U (Wu et al., 2025c)	7B	37.2	–	30.7	–	–	–	–
BAGEL (Deng et al., 2025)	8B	–	**84.8**	**54.5**	–	–	–	–
HermesFlow (Yang et al., 2025b)	1.3B	–	–	28.3	–	–	–	–
DIVA (Lu et al., 2026)	1.3B	–	–	34.3	–	–	–	–
Show-o (Xie et al., 2025)	1.3B	29.3†	28.2†	25.4†	–	41.5†	47.5†	37.9†
+ GRPO	1.3B	30.2	28.7	25.9	–	41.7	48.2	38.3
+ CGRL (ours)	1.3B	**30.7**	**29.4**	**26.7**	–	**42.5**	**49.6**	**38.8**
Janus-Pro (Chen et al., 2025b)	1.5B	43.1	75.5	36.3	38.0†	56.4†	62.5†	47.4†
+ GRPO	1.5B	46.0	75.9	36.5	38.0	56.7	65.2	47.9
+ CGRL (ours)	1.5B	**46.9**	**76.4**	**36.9**	**40.0**	**57.6**	**66.1**	**48.3**
Janus-Pro (Chen et al., 2025b)	7B	46.5	79.2	41.0	48.6†	72.2†	69.8†	53.5†
+ GRPO	7B	48.7	**79.7**	40.6	50.0	71.9	69.5	54.1
+ CGRL (ours)	7B	**49.3**	79.6	**45.3**	**51.3**	**72.8**	**70.7**	**54.3**

Table 2: **Results on text-to-image generation benchmarks.** The best scores within each backbone block are highlighted in **bold**. † denotes results obtained from our evaluation. For WISE, we follow the official protocol and use GPT-4o (2024-05-13) as the evaluator.

Model	Params	GenEval							DPG-Bench	WISE
		Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall	Overall	Overall
Generation Only										
PixArt-α (Chen et al., 2024a)	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48	71.11	0.47
SDXL (Podell et al., 2024)	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55	74.65	–
DALL-E 3 (Betker et al., 2023)	–	0.96	0.87	0.47	0.83	**0.43**	0.45	0.67	83.50	–
SD3-Medium (Esser et al., 2024)	2B	**0.99**	**0.94**	0.72	**0.89**	0.33	**0.60**	**0.74**	**84.08**	0.42
FLUX.1-dev (Black Forest Labs, 2024)	12B	0.98	0.79	**0.73**	0.77	0.22	0.45	0.66	83.79	**0.50**
Emu3-Gen (Wang et al., 2024)	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54	80.60	0.39
Unified Understanding and Generation										
Janus (Wu et al., 2025a)	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61	79.68	0.23
Harmon (Wu et al., 2025b)	1.5B	**0.99**	0.86	0.66	0.85	**0.74**	0.48	0.76	–	0.41
MMaDA (Yang et al., 2025a)	8B	**0.99**	0.76	0.61	0.84	0.20	0.37	0.63	–	–
BLIP3-o (Chen et al., 2025a)	4B	–	–	–	–	–	–	0.81	79.36	0.50
BAGEL (Deng et al., 2025)	8B	**0.99**	**0.94**	**0.81**	**0.88**	0.64	**0.63**	**0.82**	**85.07**	**0.52**
DIVA (Lu et al., 2026)	–	–	–	–	–	–	–	0.64	76.03	0.41
HermesFlow (Yang et al., 2025b)	1.3B	0.98	0.84	0.66	0.82	0.32	0.52	0.69	70.32	–
ULM-R1-S1 (Jiang et al., 2025)	1.5B	–	–	–	–	–	–	0.76	83.30	–
Show-o (Xie et al., 2025)	1.3B	**0.97**†	0.81†	**0.64**†	**0.85** †	0.30†	0.47 †	0.67†	78.40†	0.35
+ GRPO	1.3B	**0.97**	**0.85**	0.63	0.82	0.31	0.52	0.69	79.14	0.39
+ CGRL (ours)	1.3B	**0.97**	**0.85**	0.63	**0.85**	**0.32**	**0.56**	**0.70**	**79.65**	**0.41**
Janus-Pro (Chen et al., 2025b)	1.5B	0.97†	0.82†	0.50†	0.89†	0.65 †	0.54†	0.73†	82.63	0.26
+ GRPO	1.5B	**0.99**	0.84	0.51	0.88	0.67	0.57	0.74	83.03	0.32
+ CGRL (ours)	1.5B	0.98	**0.88**	**0.54**	**0.90**	**0.70**	**0.60**	**0.77**	**83.94**	**0.35**
Janus-Pro (Chen et al., 2025b)	7B	**0.99**†	0.87†	0.57 †	**0.89**†	0.73†	0.64†	0.78 †	83.79 †	0.35
+ GRPO	7B	0.98	0.88	0.59	0.88	**0.77**	**0.66**	0.79	84.41	0.41
+ CGRL (ours)	7B	**0.99**	**0.90**	**0.61**	**0.89**	**0.77**	0.63	**0.80**	**84.68**	**0.43**

Text-to-Image Generation. CGRL improves text-to-image generation over joint GRPO across all three backbones, as shown in Table 2. On Janus-Pro 1.5B, CGRL reaches 76.9 on GenEval and 83.94 on DPG-Bench, compared with 75.9 and 83.3 for ULM-R1-S1 (Jiang et al., 2025). Notably, both methods use the same backbone and training data, while ULM-R1-S1 adopts a more elaborate reward design. CGRL also clearly improves over joint GRPO, which achieves 74.4 on GenEval and 83.03 on DPG-Bench. These results show that cross-feature guidance provides consistent gains without relying on more complex reward engineering.

Image–Caption Alignment. Table 3 evaluates caption alignment in both in-domain and out-of-domain settings. CGRL consistently outperforms joint GRPO on both CLIP and SigLIP across all backbones. On COCO_{val}, CGRL improves in-domain caption alignment, and these gains transfer to $\text{NoCaps}_{\text{val}}$. This demonstrates that CGRL improves both in-domain caption Alignment and out-of-domain generalization.

Table 3: **Results on image captioning benchmarks.** The best score within each backbone block is highlighted in **bold**.

Model	Params	COCO_{val}		$\text{NoCaps}_{\text{val}}$	
		CLIP	SigLIP	CLIP	SigLIP
Show-o (Xie et al., 2025)	1.3B	68.33	51.27	64.38	45.83
+ GRPO	1.3B	69.54	62.36	65.31	54.21
+ CGRL (ours)	1.3B	**69.89**	**64.99**	**65.53**	**56.10**
Janus-Pro (Chen et al., 2025b)	1.5B	73.69	81.91	71.33	78.34
+ GRPO	1.5B	75.63	90.35	72.16	86.31
+ CGRL (ours)	1.5B	**75.79**	**90.64**	**72.80**	**87.38**
Janus-Pro (Chen et al., 2025b)	7B	73.90	82.14	71.48	78.70
+ GRPO	7B	75.38	90.54	72.88	87.10
+ CGRL (ours)	7B	**75.95**	**90.89**	**73.45**	**88.04**

Qualitative Analysis. Figure 4 compares text-to-image generation from the base models, joint GRPO, and CGRL. In these examples, CGRL better follows compositional constraints involving attribute binding, object counting, and spatial relations. Specifically, it preserves the requested *purple parking meter* and *red laptop* on Show-o, correctly generates *three refrigerators* on Janus-Pro 1.5B, and better captures the *dining table to the right of the oven* on Janus-Pro 7B. CGRL also preserves more fine-grained visual details in longer prompts. Additional text-to-image and image-captioning examples are provided in Appendix A.8.

4.3 Reward and Gradient Analysis

We analyze the reward and gradient agreement between the two branches during joint optimization.

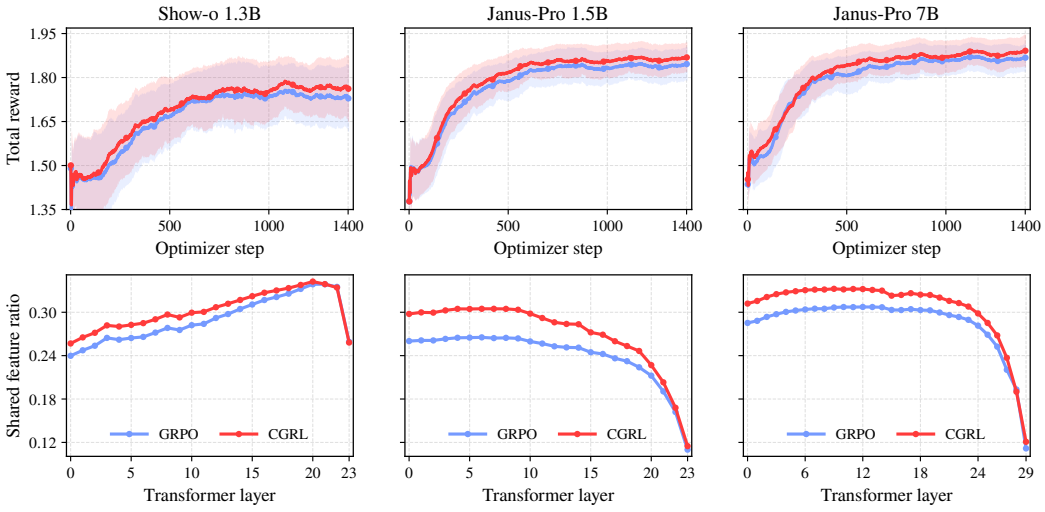


Figure 3: **Reward and gradient analysis.** Top: smoothed alignment reward during training. Bottom: shared feature ratio across transformer layers, reflecting gradient agreement between understanding and generation.



Figure 4: **Qualitative text-to-image comparison.** Each model group shows the base model, joint GRPO, and CGRL from left to right. The upper row evaluates compositional prompt following, including attribute binding, object counting, and spatial relations, while the lower row focuses on longer prompts with multiple fine-grained visual requirements. For these prompts, CGRL better preserves the specified attributes, counts, spatial relations, and detailed prompt content. Key objects and concepts in the prompts are highlighted in orange.

Reward. As shown in Figure 3, both joint GRPO and CGRL steadily increase the total alignment reward during training. CGRL generally achieves a higher reward, indicating more effective joint optimization of understanding and generation.

Gradient Agreement. We quantify gradient agreement between the two branches using the shared feature ratio, defined as the proportion of gradient dimensions with the same sign and similar magnitudes. The detailed definition is provided in Appendix A.4. As shown in Figure 3, CGRL achieves a higher shared feature ratio than joint GRPO across most layers, especially on Janus-Pro 1.5B. This indicates more consistent optimization directions between understanding and generation.

4.4 Ablations

We conduct ablation studies on Janus-Pro 1.5B to answer three questions: *(1) Can CGRL enable mutual improvement between understanding and generation and surpass task-specific RL? (2) Is cross-feature guidance more effective than self-guidance? (3) Which layer provides the most effective guidance, and why?* Table 4 summarizes the main results, while Appendix A.7 provides additional analyses of KL regularization, reward models, and cross-feature guidance strength.

Mutual Benefit from Joint Optimization. We compare three training strategies: *(1) Task-specific RL*: *U* and *G* optimize the understanding and generation branches independently; *(2) Single-branch Cross*: *U+Cross* and *G+Cross* optimize one branch with opposite-branch features; *(3) Joint RL*: $U{+}G_{\text{sep}}$ and $U{+}G_{\text{joint}}$ optimize both branches with separate or summed rewards.

$U{+}G_{\text{sep}}$ and $U{+}G_{\text{joint}}$ perform similarly, indicating that reward formulation alone does not resolve the conflict between understanding and generation. In contrast, *U+Cross* and *G+Cross* outperform their task-specific baselines *U* and *G*, showing that opposite-branch features provide useful guidance even when only one task is optimized. More importantly, *U+G+Cross* further outperforms both $U{+}G_{\text{joint}}$ and the single-branch Cross variants, demonstrating that jointly optimizing the complementary task provides additional gains beyond cross-feature guidance alone. These results show that CGRL enables understanding and generation to mutually reinforce each other during joint RL.

Cross-Feature vs. Self-Feature Guidance. We compare *Cross*, where each branch is guided by the opposite branch's input features, with *Self*, where each branch uses its own input features, and *Both*, which combines the two. *Cross* achieves the best overall performance, improving GenEval, DPG-Bench, and several understanding benchmarks, including MMStar, MMBench, and MMMU. This indicates that opposite-branch features provide complementary information during optimization. *Self* uses only within-branch features and performs worse, while *Both* yields intermediate results, suggesting that adding self-feature guidance provides no consistent benefit. These results support direct cross-feature guidance for joint optimization.

Table 4: **Ablation studies.** The four groups evaluate training strategy, single-layer selection, multi-layer fusion, and guidance source. **Bold** marks the best score. Green and red indicate performance above and below joint GRPO (U+G_{joint} without feature guidance), respectively.

Training Strategy	Layers	Guidance Source	Understanding							Generation	
			MMStar	MMBench	MMMU	MMVP	CV-Bench 2D	CV-Bench 3D	MMT	GenEval(%)	DPG
Janus-Pro 1.5B			43.1	75.5	36.3	38.0	56.4	62.5	47.4	72.8	82.63
U	–	–	46.6	76.3	36.0	38.6	57.9	65.4	48.0	–	–
U	5	Cross	46.9	76.2	36.6	39.4	58.2	66.0	48.2	–	–
G	–	–	–	–	–	–	–	–	–	76.0	83.41
G	5	Cross	–	–	–	–	–	–	–	76.4	83.67
U+G_{sep}	–	–	46.3	76.1	36.1	38.0	57.7	64.8	48.0	74.2	83.12
U+G_{joint}	–	–	46.0	75.9	36.5	38.0	56.7	65.2	47.9	74.4	83.03
U+G	3	Cross	46.6	**76.4**	36.8	38.7	57.7	66.5	48.2	75.0	83.26
U+G	5	Cross	**46.9**	**76.4**	36.9	40.0	57.6	66.1	**48.3**	**76.9**	**83.94**
U+G	12	Cross	46.5	76.1	36.7	38.0	57.1	66.5	**48.3**	74.6	83.24
U+G	19	Cross	46.3	76.0	36.6	39.3	58.0	66.9	48.2	75.3	83.79
U+G	23	Cross	**46.9**	76.0	36.7	38.7	57.5	66.2	48.1	75.0	83.45
U+G	24	Cross	45.9	76.1	36.6	40	**58.6**	**67.4**	48.2	75.1	82.95
U+G	{5, 12, 20}	Cross	46.8	75.8	36.1	38.0	57.9	67.3	48.1	75.5	82.52
U+G	{5, 7, 9}	Cross	**46.9**	75.9	**40.6**	40.6	56.7	66.4	48.1	75.3	83.25
U+G	{5, 24}	Cross	**46.9**	76.0	39.4	39.3	57.1	65.1	**48.3**	74.8	83.92
U+G	5	Self	46.6	76.2	36.6	40.0	57.7	66.6	**48.3**	74.6	83.19
U+G	5	Both	46.4	76.0	36.7	**40.7**	57.2	63.6	48.0	75.9	83.22

Guidance Layer Selection. We compare single-layer and multi-layer guidance and find that layer 5 provides the best overall balance between understanding and generation, achieving the highest scores on five benchmarks. Intuitively, shallower layers retain more general multimodal representations, whereas deeper layers become increasingly task-specific, making shallow features more transferable across branches. Multi-layer fusion improves selected understanding scores but lowers both generation scores relative to layer 5. Consistently, layer 5 also exhibits the highest text–image cosine similarity score in Figure 5. These results support using a well-aligned single layer for cross-feature guidance.

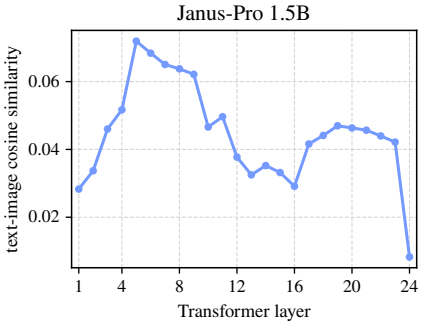


Figure 5: In Janus-Pro 1.5B, the hidden states cosine similarity score between an image and its paired caption peaks at layer 5.

5 Conclusion

In this work, we investigate the optimization conflict between multimodal understanding and image generation during joint RL for unified multimodal models. We propose **CGRL**, which leverages the semantic correspondence between the two branches to provide cross-feature guidance and modulate sequence-level advantages at the token level. Experiments on Janus-Pro and Show-o consistently demonstrate improvements over joint GRPO across multimodal understanding, image captioning, and text-to-image generation. Further analysis shows that CGRL strengthens cross-branch interaction and increases gradient agreement, providing evidence of reduced optimization conflict. Overall, CGRL demonstrates that incorporating cross-branch semantic guidance into policy optimization can turn joint RL from conflicting objectives into mutually beneficial learning.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking and language polishing. We have not used generative AI tools for research ideation, method design, or the generation of experimental results. All AI-assisted text revisions were carefully reviewed by the authors for technical accuracy, consistency with the original meaning, and correctness of the reported results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies post-training for existing multimodal model architectures and uses established research datasets and benchmarks. The resulting models may inherit biases, unsafe generation behavior, or licensing constraints from their base checkpoints and training data. CGRL does not directly remove those risks.

REPRODUCIBILITY STATEMENT

The available code is provided in the supplementary material. Section 3 specifies the rollout, reward, feature weighting, and policy update. Appendix A.2 lists the training hyperparameters. Appendix A.3 specifies the denoising transitions and policy updates. Appendix A.4 defines the gradient diagnostic. Appendix A.1 describes the evaluation benchmarks, splits, and metrics.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023.

Black Forest Labs. FLUX, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023.

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

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.

Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 5823–5846. Curran Associates, Inc., 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pp. 216–233. Springer, 2024b.

Renjie Lu, Xulong Zhang, Xiaoyang Qu, Jianzong Wang, and Shangfei Wang. DIVA: Harnessing the representation divergence in unified multimodal models for mutual reinforcement. In *Forty-third International Conference on Machine Learning*, 2026.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

Yifu Luo, Xinhao Hu, Keyu Fan, Haoyuan Sun, Zeyu Chen, Bo Xia, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement learning meets masked generative models: Mask-GRPO for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 108460–108485. Curran Associates, Inc., 2025.

Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. UniRL: Self-improving unified multimodal models via supervised and reinforcement learning. *arXiv preprint arXiv:2505.23380*, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. In *Conference on Language Modeling*, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. In *Forty-third International Conference on Machine Learning*, 2026.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. In *International Conference on Machine Learning*, 2026.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356. Curran Associates, Inc., 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12977, 2025a.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53366–53397. PMLR, 2024.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17739–17750, 2025b.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a unified foundation model integrating visual understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025c.

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. In *International Conference on Learning Representations*, pp. 28240–28264, 2025.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Haochen Wang, Zhendong Wang, Bin Lin, Hao Li, Xinyan Xiao, Jingdong Wang, Haifeng Wang, and Li Yuan. Unified multimodal models as auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41903–41912, 2026.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138867–138907. Curran Associates, Inc., 2025a.

Ling Yang, Xinchen Zhang, Ye Tian, Shiyi Zhang, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. HermesFlow: Seamlessly closing the gap in multimodal understanding and generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 62248–62272. Curran Associates, Inc., 2025b.

Sen Ye, Mengde Xu, Shuyang Gu, Di He, Liwei Wang, and Han Hu. Understanding vs. generation: Navigating optimization dilemma in multimodal models. In *International Conference on Learning Representations*, pp. 25118–25141, 2026.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMT-Bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57116–57198. PMLR, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Jiaxing Zhang and Hao Tang. Resolving task objective conflicts in unified model via task-aware mixture-of-experts. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 3131–3133. International Foundation for Autonomous Agents and Multiagent Systems, 2026.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, pp. 6446–6469, 2025.

A Appendix

This supplementary document is organized as follows:

- **A.1** provides evaluation benchmarks.
- **A.2** shows training hyperparameter configuration.
- **A.3** describes CGRL for discrete-diffusion transitions in Show-o.
- **A.4** provides gradient diagnostic method.
- **A.5** shows visualizing Cross-Feature Guidance.
- **A.6** shows detailed evaluation results.
- **A.7** shows additional ablation studies.
- **A.8** shows qualitative case study analysis.

A.1 Evaluation Benchmarks

Our evaluation covers diverse capabilities, including general multimodal understanding, compositional and knowledge-based image generation, and image–caption alignment.

A.1.1 Multimodal Understanding

MMStar. MMStar (Chen et al., 2024b) measures multimodal understanding with 1,500 multiple-choice questions selected to require image information. The benchmark balances six capabilities, with 250 questions each: coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics. Its construction combines automated filtering with human review to reduce questions that can be solved from text alone. We report overall answer accuracy across the benchmark.

MMBench. MMBench (Liu et al., 2024b) evaluates perception and reasoning through multiple-choice visual questions organized into 20 ability dimensions. These cover coarse perception, perception of individual and multiple instances, and attribute, relation, and logical reasoning. The benchmark includes English and Chinese versions and introduces circular evaluation, which permutes answer options to measure consistency across their orderings. We use the English development split of MMBench v1.0, MMBench_DEV_EN, and report its overall accuracy.

MMMU. MMMU (Yue et al., 2024) evaluates multimodal reasoning that requires college-level subject knowledge. It contains approximately 11,500 questions across 30 subjects in six disciplines, including science, medicine, business, and engineering. Problems combine text with diverse visual material, such as diagrams, charts, maps, and chemical structures, and include both multiple-choice and open-ended answers. We report overall accuracy on the 900-question validation split. Specifically, we use the validation Overall result from MMMU_DEV_VAL.

MMVP. MMVP (Tong et al., 2024b) tests fine-grained perception using 150 image pairs selected to have similar CLIP representations despite perceptually clear differences. Each image has an associated question, giving 300 questions in total. Questions probe details such as object orientation, presence, or position through short answer choices. The original pair-level metric awards credit only when both answers are correct, testing whether the model responds to visual changes across a pair of images.

CV-Bench. CV-Bench (Tong et al., 2024a) reformulates annotated computer vision tasks as multiple-choice visual questions. Its 2,638 manually inspected examples are derived from ADE20K, COCO, and Omni3D. CV-Bench-2D evaluates object counting and spatial relationships in the image plane, whereas CV-Bench-3D evaluates depth ordering and relative distance between objects. These tasks emphasize visual structure supported by existing annotations. We report accuracy for the 2D and 3D subsets separately, allowing the evaluation to distinguish gains in counting and spatial perception from gains in understanding three-dimensional scene geometry.

MMT-Bench. MMT-Bench (Ying et al., 2024) evaluates multimodal understanding across 32 meta-tasks and 162 subtasks. The full benchmark contains 31,325 multiple-choice questions covering recognition, localization, reasoning, and planning, with scenarios that include driving and embodied navigation. This breadth tests whether improvements extend across tasks with different visual inputs and knowledge requirements. We use the validation subset, MMT-Bench_VAL, and report its Overall accuracy, which includes visual recognition. Together with the more focused perception benchmarks, MMT-Bench assesses the range of understanding capabilities retained after joint post-training.

A.1.2 Text-to-Image Generation

GenEval. GenEval (Ghosh et al., 2023) evaluates compositional prompt following with 553 prompts covering six tasks: single-object presence, two-object composition, counting, colors, spatial position, and color-attribute binding. Its evaluation pipeline uses object detection to check the required entities, counts, and relative positions, and a color classifier to assess object attributes. Each image receives a binary success score according to whether it satisfies all requirements of its prompt. GenEval scores use a 0–1 scale, except in Table 4 and the main-results discussion, where they are percentages.

DPG-Bench. DPG-Bench (Hu et al., 2024) evaluates whether generated images satisfy long, semantically dense prompts. Each prompt is decomposed into questions about global properties, entities, attributes, relations, and other details. The official evaluator uses mPLUG-large to answer these questions and applies a dependency graph: a correct attribute or relation cannot receive credit when its prerequisite is missing. The overall score averages these dependency-aware question scores within each prompt and then across prompts. We report all scores on a 0–100 scale, while category scores summarize individual questions before dependency filtering.

WISE. WISE (Niu et al., 2026) evaluates the use of world knowledge in image generation through 1,000 prompts spanning 25 subdomains. Its six reporting categories cover cultural knowledge, time, space, biology, physics, and chemistry. We use the original WISE protocol with GPT-4o (2024-05-13), which assigns 0–2 scores for semantic consistency, realism, and aesthetic quality.

A.1.3 Image Captioning

COCO. COCO (Lin et al., 2014) contains photographs of everyday scenes with common objects in their natural context and human-written image descriptions. Captioning requires identifying the salient objects and expressing their actions, attributes, and relationships in text. We evaluate on COCO_{val}, which provides an in-domain assessment relative to the COCO-derived pairs.

NoCaps. NoCaps (Agrawal et al., 2019) evaluates novel-object captioning on images drawn from Open Images, extending the visual concepts represented in COCO caption training. The benchmark distinguishes in-domain, near-domain, and out-of-domain images according to their object categories. Relative to our COCO-derived RL data, this evaluation tests whether captioning improvements transfer to a broader object vocabulary and different scene distributions.

Caption-Alignment Metrics. We measure alignment between each generated caption and its conditioning image with CLIP-L/14@336 (Radford et al., 2021) and SigLIP-L/16@384 (Zhai et al., 2023). Both encode images and text into a shared embedding space. Both models differ from the SigLIP2-SO400M/14@384 reward model, so their scores assess whether improved alignment extends beyond the encoder used during CGRL training.

A.2 Training Hyperparameter Configuration

Table 5 summarizes the complete training configuration for all three backbones. For Show-o 1.3B, we use the Show-o-512×512 checkpoint. We use AdamW with linear warmup followed by cosine learning-rate decay. For Janus-Pro 7B, LoRA is applied to the language model with rank $r = 128$, scaling factor $\alpha = 256$, and zero dropout, targeting both attention and MLP projection layers. Across all models, the GRPO clipping coefficient is set to 0.2, the cross-feature guidance strength to $\gamma = 1$. Experiments are conducted on NVIDIA A40 GPUs.

Table 5: **Detailed training configuration.**

Hyperparameter	Janus-Pro 1.5B	Janus-Pro 7B	Show-o 1.3B
Global batch size	16	16	16
Rollouts per input K	8	8	8
Optimizer	AdamW	AdamW	AdamW
Learning rate	2×10^{-6}	1×10^{-5}	2×10^{-6}
LR schedule	Cosine Decay	Cosine Decay	Cosine Decay
Warmup Ratio	0.05	0.05	0.01
DeepSpeed	ZeRO-2	ZeRO-2	ZeRO-2
Visual modules	Frozen	Frozen	Frozen
LLM optimization	Full	LoRA	Full
GRPO clip ε	0.2	0.2	0.2
LoRA rank r	–	128	–
LoRA α	–	256	–
LoRA dropout	–	0.0	–
LoRA target modules	–	Attn. + MLP	–
Cross-feature guidance layer ℓ	5	10	11
Guidance strength γ	1	1	1
Weight bounds $[w_{\min}, w_{\max})]$	[0.1, 2]	[0.1, 2]	[0.1, 2]
Image denoising steps	–	–	18

We select the guidance layer using the text–image cosine similarity score, computed as the similarity of paired image–caption hidden states minus that of randomly shuffled pairs. This subtraction reduces model-internal similarity bias and better isolates pair-specific semantic correspondence. Figures 5 and 6 show the resulting layer-wise scores.

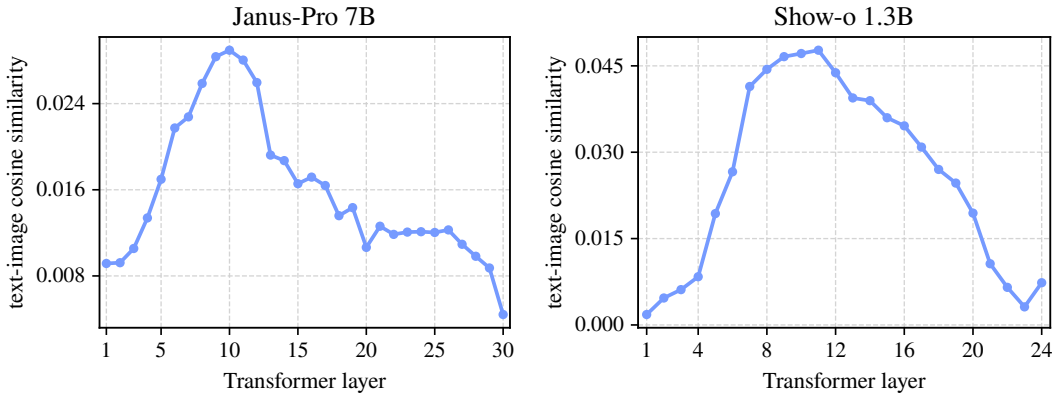


Figure 6: **Guidance layer selection.** The text–image alignment score peaks at layer 10 for Janus-Pro 7B and layer 11 for Show-o 1.3B. These layers are selected for cross-feature guidance in CGRL.

A.3 Discrete Diffusion Transition Decisions for Show-o

Show-o's generation branch follows Candidate 1 of Mask-GRPO (Luo et al., 2025), where each denoising transition is treated as a joint policy decision. Let t index visual tokens and τ index denoising transitions. At transition τ of rollout i, the accepted and remasked positions are denoted by $\mathcal{E}_{i,\tau}$ and $\mathcal{R}_{i,\tau}$, respectively.

Cross Feature Guidance. We compute the token-level cosine similarity scores $w_{g,i,t}$ following Eq. 5. Since policy decisions are made at the transition level, we average the similarity scores over the tokens accepted at each transition:

$$w_{g,i,\tau} = \frac{1}{|\mathcal{E}_{i,\tau}|} \sum_{t \in \mathcal{E}_{i,\tau}} w_{g,i,t}. \tag{12}$$

We then apply min max scaling across transitions within each rollout:

$$\widehat{w}_{g,i,\tau} = \frac{w_{g,i,\tau} - \min_{\tau'} w_{g,i,\tau'}}{\max_{\tau'} w_{g,i,\tau'} - \min_{\tau'} w_{g,i,\tau'} + \epsilon_s}. \tag{13}$$

Advantage Modulation. Following the autoregressive branches, we zero center the normalized scores and map them to bounded transition level guidance weights:

$$\widetilde{w}_{g,i,\tau} = \mathrm{sg}\Big[\mathrm{clip}\Big(1 + \gamma\left(\widehat{w}_{g,i,\tau} - \overline{\widehat{w}}_{g,i}\right), w_{\min}, w_{\max}\Big)\Big], \qquad \overline{\widehat{w}}_{g,i} = \frac{1}{N_{g,i}} \sum_{\tau'=1}^{N_{g,i}} \widehat{w}_{g,i,\tau'}. \tag{14}$$

Here, $N_{g,i}$ denotes the number of denoising transitions in rollout i. The transition level advantage is

$$A_{g,i,\tau} = A_i \widetilde{w}_{g,i,\tau}. \tag{15}$$

Policy Update. Let $o_{g,i,\tau}$ denote the joint policy decision at transition τ and $x_{g,i,\tau}$ its conditioning context. Following Candidate 1, the transition likelihood surrogate is

$$\pi^g_{\boldsymbol{\theta}}(o_{g,i,\tau} \mid x_{g,i,\tau}) = \prod_{t \in \mathcal{E}_{i,\tau}} p_{\boldsymbol{\theta},t}(o_t) \prod_{t \in \mathcal{R}_{i,\tau}} \left[\sum_{v \in \mathcal{V}^-_{\boldsymbol{\theta},t}} p_{\boldsymbol{\theta},t}(v) \right]. \tag{16}$$

Here, $\mathcal{V}^-_{\boldsymbol{\theta},t} = \{v : p_{\boldsymbol{\theta},t}(v) < c_{i,\tau}\}$ contains the visual codes below the recorded confidence threshold $c_{i,\tau}$. We substitute this likelihood into Eq. 9 to obtain the transition policy ratio and optimize the generation branch using Eq. 10, replacing the token index t with the transition index τ and using $A_{g,i,\tau}$ from Eq. 15.

A.4 Gradient Diagnostic Computation

We record gradients of the understanding and generation losses with respect to the monitored transformer-layer parameters in separate backward passes. For layer ℓ, let $g_{b,\ell,d}$ denote coordinate d of the recorded parameter-gradient vector for branch $b \in \{u, g\}$. The two branches are compared on the same parameter coordinates.

An active parameter-gradient coordinate contributes to the shared feature ratio when the branch gradients have comparable magnitudes and matching signs. We define the activity mask $a_{\ell,d}$ and magnitude ratio $c_{\ell,d}$ as

$$a_{\ell,d} = \mathbf{1}\left[|g_{u,\ell,d}| + |g_{g,\ell,d}| > \epsilon_g\right], \quad c_{\ell,d} = \frac{\min(|g_{u,\ell,d}|, |g_{g,\ell,d}|)}{\max(|g_{u,\ell,d}|, |g_{g,\ell,d}|) + \epsilon_g}. \tag{17}$$

The shared feature ratio at layer ℓ is

$$R_\ell = \frac{\sum_d a_{\ell,d}\, \mathbf{1}[c_{\ell,d} \geq 0.25]\, \mathbf{1}[g_{u,\ell,d} g_{g,\ell,d} \geq 0]}{\max(1, \sum_d a_{\ell,d})}. \tag{18}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function and $\epsilon_g = 10^{-12}$. The threshold 0.25 selects gradient magnitudes differing by approximately a factor of four or less. A layer with no active coordinates has $R_\ell = 0$.

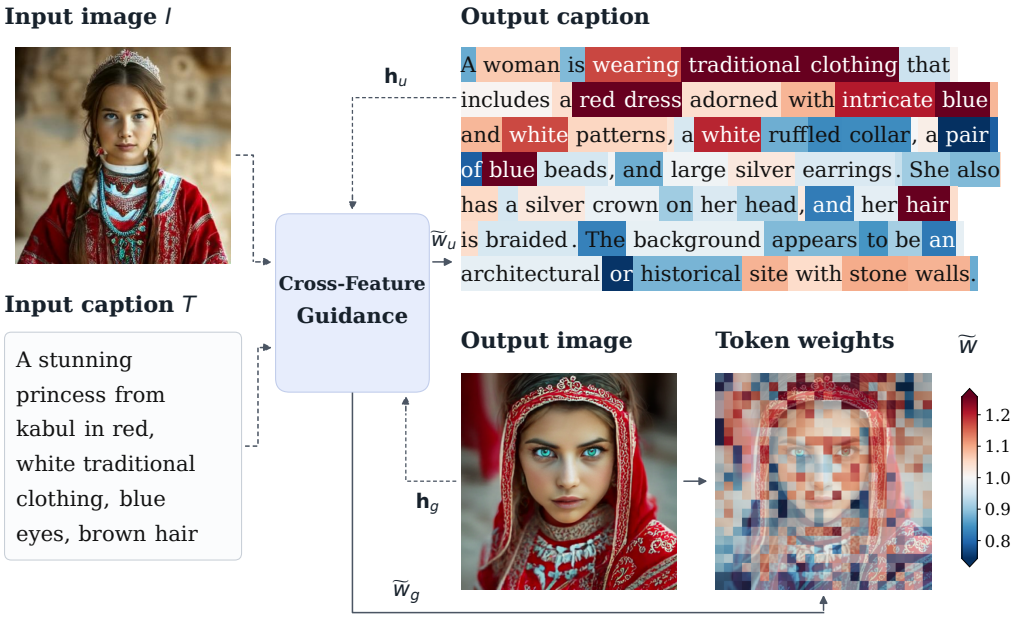


Figure 7: **Cross-feature guidance on Janus-Pro 1.5B.** The output caption is generated from image I and weighted using the input caption T. The output image is generated from T, with its token weights computed using I. Dashed arrows supply input and rollout features to the guidance module; $\widetilde{w}_u$ and $\widetilde{w}_g$ denote the token-level cross-feature guidance weights for caption and image rollouts.

A.5 Visualizing Cross-Feature Guidance

Figure 7 visualizes layer-5 cross-feature guidance on Janus-Pro 1.5B. The paired input caption and image provide the references for caption and image tokens, respectively. Both panels share a color scale: red indicates larger advantage multipliers and blue indicates smaller ones, with endpoint colors used for values outside the displayed range.

Caption guidance. The weights emphasize the subject and its visual attributes. The subject token "woman" receives a multiplier above one, while "clothing," "red," and "hair" receive larger values of 1.46, 1.44, and 1.29. In comparison, the adverb "also" and conjunction "or" receive 0.89 and 0.75, and the later occurrences of "and" are also downweighted. The highlighted phrases, such as "traditional clothing" and "red dress," describe attributes present in the paired caption. This pattern places greater emphasis on visually descriptive content and less on several words that mainly serve grammatical functions.

Image guidance. Higher weights appear around parts of the headscarf's curved upper border, its top ornaments, and the upper-right edge beside the face. These regions contain contours and fine details that define the subject's appearance. Several background patches and lower clothing regions receive smaller weights. The spatially selective pattern highlights particular structures within the portrait. Together, the two panels illustrate how cross-feature guidance emphasizes descriptive tokens and selected visual structures within each rollout.

A.6 Detailed Evaluation Results

Table 6: **Detailed results on WISE.** Scores follow the original WISE protocol with GPT-4o (2024-05-13) as the evaluator. The best score within each backbone block is highlighted in **bold**.

Model	Params	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Show-o	1.3B	0.28	0.40	0.48	0.30	0.46	0.30	0.35
+ GRPO	1.3B	0.32	0.42	0.59	**0.34**	0.53	0.31	0.39
+ CGRL (ours)	1.3B	**0.34**	**0.44**	**0.61**	**0.34**	**0.56**	**0.33**	**0.41**
Janus-Pro	1.5B	0.20	0.28	0.45	0.24	0.32	0.16	0.26
+ GRPO	1.5B	0.24	0.35	0.51	0.35	0.42	0.22	0.32
+ CGRL (ours)	1.5B	**0.26**	**0.40**	**0.53**	**0.38**	**0.51**	**0.23**	**0.35**
Janus-Pro	7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
+ GRPO	7B	0.34	0.42	0.57	0.42	0.54	0.32	0.41
+ CGRL (ours)	7B	**0.35**	**0.45**	**0.60**	**0.42**	**0.60**	**0.33**	**0.43**

Table 7: **Detailed results on DPG-Bench.** We report category and overall scores for the backbones and training methods in Table 2. The best score within each backbone block is highlighted in **bold**. GRPO denotes Mask-GRPO for Show-o's generation branch.

Model	Params	Global	Entity	Attribute	Relation	Other	Overall
Show-o	1.3B	80.78	85.14	85.83	90.64	68.00	78.40
+ GRPO	1.3B	**81.99**	85.78	**87.10**	**91.45**	68.80	79.14
+ CGRL (ours)	1.3B	81.69	**86.93**	86.04	91.03	**70.30**	**79.65**
Janus-Pro	1.5B	**87.58**	88.63	88.17	88.98	88.30	82.63
+ GRPO	1.5B	82.67	88.40	**88.76**	90.67	**88.78**	83.03
+ CGRL (ours)	1.5B	81.84	**90.34**	87.37	**92.40**	80.14	**83.94**
Janus-Pro	7B	**83.66**	89.64	87.53	93.13	80.30	83.79
+ GRPO	7B	82.67	**90.53**	**87.81**	93.54	83.00	84.41
+ CGRL (ours)	7B	83.28	**90.53**	87.73	**93.74**	**83.40**	**84.68**

Tables 6 and 7 provide the category-level results underlying the WISE and DPG-Bench overall scores in Table 2.

WISE. CGRL achieves the highest overall WISE score on all three backbones, improving over joint GRPO from 0.39 to 0.41 on Show-o, from 0.32 to 0.35 on Janus-Pro 1.5B, and from 0.41 to 0.43 on Janus-Pro 7B. Across the six categories, CGRL improves 16 of 18 scores and matches the remaining two. Physics shows the largest gain for every backbone, increasing from 0.53 to 0.56, from 0.42 to 0.51, and from 0.54 to 0.60, respectively. The gains extend across scientific and cultural knowledge, showing that cross-feature guidance improves performance on knowledge-intensive generation beyond joint GRPO.

DPG-Bench. CGRL attains the highest overall score within each backbone block, with different patterns across categories. Relative to joint GRPO, Janus-Pro 1.5B improves Entity from 88.40 to 90.34 and Relation from 90.67 to 92.40. Janus-Pro 7B matches joint GRPO on Entity and improves Global, Relation, and Other, with a small decrease in Attribute. For Show-o, CGRL raises Entity from 85.78 to 86.93 and Other from 68.80 to 70.30, while Global, Attribute, and Relation are lower. Thus, the higher overall scores accompany selective improvements in prompt alignment rather than uniform gains in every category.

A.7 Additional Ablation Studies

We evaluate KL regularization, reward model choice, and cross-feature guidance strength on Janus-Pro 1.5B. We report final overall GenEval and DPG-Bench scores and MMStar accuracy. Shaded rows denote the default CGRL configuration, and bold marks the highest score in each column.

KL Regularization. Table 8 compares KL coefficients $\beta \in \{0, 0.01, 0.04\}$. Omitting the KL penalty gives the highest scores on all three benchmarks. Relative to $\beta = 0.04$, the default $\beta = 0$ improves GenEval by 0.03, DPG-Bench by 0.70 points, and MMStar by 1.0 point. Reducing the coefficient to 0.01 narrows the gap, but still underperforms the default on every metric. These results support omitting KL regularization in our training setup.

Table 8: **Effect of KL regularization.**

KL regularization	GenEval	DPG-Bench	MMStar
With KL ($\beta = 0.04$)	73.8	83.24	45.9
With KL ($\beta = 0.01$)	75.6	83.47	46.2
No KL ($\beta = 0$, ours)	**76.9**	**83.94**	**46.9**

Reward Model. We replace the reward model S_ϕ in Eq. 3, retaining the same reward aggregation and group normalization. Table 9 shows that SigLIP2 outperforms CLIP ViT-L/14@336 on all three metrics. Qwen3-VL-Embedding-2B achieves the highest MMStar accuracy of 47.2, exceeding SigLIP2 by 0.3 points, but scores lower on GenEval (0.75 vs. 0.77) and DPG-Bench (83.37 vs. 83.94). The comparison reveals a trade-off between generation scores and MMStar accuracy.

Table 9: **Effect of the reward model.**

Reward model	GenEval	DPG-Bench	MMStar
CLIP ViT-L/14@336	75.4	83.53	46.4
Qwen3-VL-Embedding-2B	75.1	83.37	**47.2**
SigLIP2-SO400M/14@384 (ours)	**76.9**	**83.94**	46.9

Cross-Feature Guidance Strength. We vary $\gamma \in \{0.5, 1, 2\}$ in Eq. 8, keeping the guidance layer at layer 5. The default $\gamma = 1$ achieves the highest scores on all three benchmarks (Table 10). Increasing γ from 0.5 to 1 improves both understanding and generation. A further increase to 2 reduces GenEval from 0.77 to 0.76 and DPG-Bench from 83.94 to 83.72, while MMStar falls from 46.9 to 45.5. This comparison supports $\gamma = 1$ over the weaker and stronger modulation settings tested here.

Table 10: **Effect of cross-feature guidance strength.**

Guidance strength	GenEval	DPG-Bench	MMStar
$\gamma = 0.5$	76.2	83.58	46.6
$\gamma = 1$ **(ours)**	**76.9**	**83.94**	**46.9**
$\gamma = 2$	75.7	83.72	45.5



Figure 8: **Additional text-to-image cases.** Columns show Show-o 1.3B (Show-o-512), Janus-Pro 1.5B, and Janus-Pro 7B after CGRL training. Each 2×2 panel contains four generated images for one prompt. Rows correspond to Prompts 1–4: a racing robot, an inhabited fantasy galaxy, an elderly winged gentleman, and a motorcycle in an ornate bank lobby.

A.8 Qualitative Case Analysis

A.8.1 Text-to-Image Generation

Figure 8 shows CGRL outputs for four prompts that combine a central subject with detailed attributes and surroundings. Each model provides four images per prompt, allowing us to inspect which requested elements persist across different samples. The cases cover object and attribute binding,

foreground–background composition, small accessories, and text rendered within an image. The complete prompts and corresponding observations follow in row order.

Prompt 1: Racing robot.

> An intricately designed robot with a polished metallic surface, donning a vibrant red and white race car suit, stands with a confident posture in front of a sleek F1 race car. The robot's black visor reflects the brilliant hues of the setting sun, which casts a warm glow over the futuristic cityscape depicted in the background. The illustration, reminiscent of a scene from a dynamic comic book, captures the essence of speed and technology.

The displayed samples combine a red-and-white racing figure, a Formula 1 car, and a city skyline under warm sunset lighting. The character occupies the foreground, with the car behind or beside it and the skyline farther away. This arrangement preserves the relationship among the main entities while keeping the dark visor and racing colors attached to the character. The visual realization varies across backbones. Show-o favors rounded mechanical bodies, several Janus-Pro 1.5B samples resemble helmeted human drivers, and Janus-Pro 7B depicts exposed metallic armor and joints. Thus, the samples share the requested scene composition, but differ in how explicitly they express the robot's mechanical identity. A racing outfit alone leaves that identity ambiguous.

Prompt 2: Inhabited fantasy galaxy.

> A striking image depicting an imagined distant galaxy teeming with life, inspired by the work of the renowned artist Caspar David Friedrich, is showcased in high quality on Artstation. The matte painting presents a vibrant scene filled with ethereal colors and pulsating with otherworldly energy. As viewers delve into the scene, they encounter intricate details of celestial bodies and the silhouettes of people populating this fantasy cosmos.

The scenes combine luminous galaxies and celestial bodies with human silhouettes on the ground. The inhabitants remain visible even in scenes dominated by the sky. Show-o places scattered figures beneath broad atmospheric arcs, with much of the image devoted to the sky. Janus-Pro 1.5B includes more individually defined planets and spiral structures, and several samples use larger foreground figures to establish depth. Janus-Pro 7B concentrates people and architectural silhouettes along a dark lower strip beneath an orange-and-blue sky. These compositions preserve the human and cosmic elements through different distributions of scale and foreground space. The repeated elements across the four samples are the inhabited setting and celestial structures, while the viewpoint and arrangement of the sky vary.

Prompt 3: Elderly winged gentleman.

> an intricately detailed character drawing by Bastien Lecouffe Deharme featuring an elderly gentleman with long, flowing grey hair. He possesses majestic, large grey wings that extend from his back, conveying both wisdom and strength. The refinement of the character is further accentuated by a polished monocle nestled over one keen eye, and he is dressed in sophisticated attire that hints at a bygone era of elegance.

The character samples retain an elderly face, gray or white hair, feathered wings, and formal clothing. Show-o uses outlined character drawings with relatively complete upper bodies, whereas the Janus-Pro samples favor closer portraits with more visible facial texture and shading. The wings remain associated with the character's back, although their full extent is sometimes cropped by the portrait framing. The monocle is less consistent than these larger attributes. Several Show-o samples lack a clearly identifiable single lens, while some Janus-Pro 7B samples replace the monocle with two-lens spectacles. Janus-Pro 1.5B includes visibly enlarged single lenses, but hair length and the framing of the wings vary across samples. The exact accessory remains less consistently reproduced than the elderly face, wings, and formal attire.

Prompt 4: Motorcycle in an ornate bank lobby.

> A sleek motorcycle with the word "BUZZ" emblazoned in bold letters on its side is parked inside an ornate bank lobby. The lobby features marble floors and intricate gold trimmings along the walls. The motorcycle's chrome accents gleam under the soft glow of the chandelier overhead, contrasting with the rich, dark wood of the bank's counters.

The samples place motorcycles in elaborate interiors with reflective stone floors, gold detailing, chandeliers, and dark wooden surfaces. The motorcycles remain the dominant foreground objects, while the architectural elements surround them and provide the requested indoor setting. Show-o uses comparatively sparse wall paneling and clear floor reflections; the Janus-Pro samples include denser gold ornamentation and more of the surrounding interior. Vehicle pose also varies, from side views that expose the tank to oblique views that emphasize the front wheel and handlebars. The requested word "BUZZ," however, is not consistently reproduced. Several motorcycles carry letter-like markings with incorrect or ambiguous characters, including samples where the side of the vehicle is visible. The examples therefore separate successful composition and material rendering from exact character reproduction. Fine scene details can be retained while a short, explicitly requested word remains inaccurate.

A.8.2 Image Captioning

Figure 9 compares captions from the base models, joint GRPO, and CGRL on six images. The examples examine whether a caption assigns attributes to the correct objects, retains relevant entities and relations, and transcribes visible text.

Attribute binding in the baseball scene. All three Janus-Pro 1.5B captions identify a baseball player swinging at a pitch or ball. Their difference lies in how the clothing attribute is expressed. The base caption omits the uniform color, and joint GRPO uses "light blue baseball player," which leaves the color attached grammatically to the person. CGRL describes a player "in a light blue uniform," assigning the observed color explicitly to the clothing while retaining the batting action. The added specificity comes from relating an attribute to its object, rather than from adding more scene entities.

Visible text and clothing in the racing scene. The second Janus-Pro 1.5B example requires both recognition of the interaction and attention to small details on the car. The base model identifies the inspection of the cockpit but transcribes the branding as "BALTIMORE OPEN." CGRL gives the annotated correct text, "BALTIC OPEN," and describes the man's red jacket. Joint GRPO instead mentions the Sauber and Audi logos. The comparison illustrates two distinct forms of detail coverage: CGRL combines a clothing attribute with the wording on the vehicle, while joint GRPO focuses on other branding. CGRL also retains the main relation between the man and the racing car, so the local details remain part of a coherent description of the scene.

Spatial context in the alpaca scene. The three Janus-Pro 7B captions agree on the alpaca, tree, and grassy field. Joint GRPO specifies the animal's light brown color, whereas CGRL adds that it stands "near a red fence." The fence is a separate background object, and the phrase relates it to the animal rather than merely listing its presence. This extends the caption from the central subject and its immediate surroundings to a broader description of the setting. The example highlights how a short spatial phrase can increase scene coverage without changing the main subject or action.

Object specificity in the sandwich scene. The base Janus-Pro 7B caption identifies meat, egg, melted cheese, and a white plate. Joint GRPO names beef more specifically, but its caption omits the cheese and plate. CGRL describes steak and retains the fried egg, melted cheese, and white plate. Its difference from the base caption is therefore mainly the specificity of the filling, while its difference from joint GRPO is the coverage of additional visible components. This example shows that naming the central ingredient and describing its surrounding objects are complementary aspects of caption quality. The useful detail lies in which components are retained, rather than caption length alone.

Entity coverage and a remaining error in the barge scene. The base Show-o caption includes a man in a blue shirt and a yellow construction vehicle on a boat. Joint GRPO describes the vehicle on a barge but omits the man. CGRL retains the barge description and adds the man in a blue shirt working on it, recovering a person–scene relation absent from the joint GRPO caption. Both trained variants nevertheless include the incorrect "number 6." Additional scene coverage leaves this local recognition error unresolved.



Figure 9: **Qualitative image captioning comparison.** Each model group shows input images and captions from the base model, joint GRPO, and CGRL. Green and red mark correct and incorrect details. The six examples cover clothing attributes, scene text, spatial relations, food descriptions, and interactions between people and objects.

Person-specific attributes in the boxing scene. The base Show-o caption distinguishes a boxer in a red shirt from one in a black shirt. Joint GRPO describes the two boxers jointly as wearing "red and black," leaving the correspondence between person, garment, and color implicit. CGRL gives the combinations separately: one wears a red shirt and black shorts, and the other a black shirt and red shorts. This is a more explicit description of the visible attributes because the same two colors occur on different garments across the two people. The caption preserves the shared fighting action while identifying which clothing belongs to each boxer.

These caption examples illustrate the attribute assignment, secondary-object coverage, and exact text recognition that cross-feature guidance targets at the token level. The retained "number 6" error, together with the monocle and "BUZZ" generation errors, identifies fine visual details that remain unresolved in these examples.