arXiv:2510.25682v2 [cs.CL] 30 Oct 2025

PairUni: Pairwise Training for Unified Multimodal Language Models

Jiani Zheng, **Zhiyang Teng**, **Xiangtai Li**, **Anran Wang**, **Yu Tian**, **Kunpeng Qiu**, **Ye Tian**, **Haochen Wang**, **Zhuochen Wang**

ByteDance

Abstract

Unified Vision-Language Models (UVLMs) must perform both understanding and generation within a single architecture, but these tasks rely on heterogeneous data and supervision, making it difficult to balance them during reinforcement learning (RL). We propose PairUni, a unified framework that reorganizes data into understanding–generation (UG) pairs and aligns optimization accordingly. We first use GPT-o3 to augment single-task data, generating captions for understanding samples and question-answer (QA) pairs for generation samples, forming aligned pairs from the same instance. Additionally, for each generation sample, we retrieve a semantically related understanding example to form a retrieved pair, linking different but related data points. These paired structures expose cross-task semantic correspondences and support consistent policy learning. To leverage this structure, we present Pair-GPRO, a pair-aware variant based on Group Relative Policy Optimization. It assigns a similarity score to each pair to modulate the advantage, strengthening learning from well-aligned examples and reducing task interference. We curate a high-quality dataset of 16K UG pairs named as PairUG for RL fine-tuning and evaluate PairUni on the powerful Janus-Pro UVLMs. Our approach achieves balanced improvements on various UVLMs, outperforming strong UVLMs RL baselines. Code: `https://github.com/Haochen-Wang409/PairUni`

Date: October 31, 2025

†**Correspondence:** `zhengjiani.0123@bytedance.com`, `zhiyang.teng@bytedance.com`,

1 Introduction

Unified vision–language models (UVLMs) have demonstrated strong performance in both multimodal understanding and image generation across a variety of recent systems and backbones [3, 8, 25, 40]. However, as evaluation steadily shifts toward complex, multi-step reasoning—covering mathematics, science, and multi-hop visual question answering—the goal of training a single system that balances these two capabilities under a unified learning paradigm remains non-trivial [5, 20, 24, 53]. This challenge is particularly acute when using reinforcement learning (RL), since understanding and generation are supervised with heterogeneous objectives and data formats, making the optimization process highly sensitive to data batching and cross-task credit assignment.

This fundamental disparity has largely led to siloed advancements. For instance, text-to-image RL has focused on improving object controllability and prompt adherence [13, 30], while visual reasoning improvements have separately targeted accuracy on benchmarks for math or science. Consequently, attempts at unified RL encounter significant practical obstacles: (i) task interference during joint optimization, where gains on one objective cause regressions on the other [14]; (ii) a broad and diverse understanding task space (e.g., math, charts, OCR) that resists a single reward design; and (iii) limited guidance on how to select and align data for unified RL at scale, constraining both stability and ceiling performance [12].

Prevailing GRPO-based strategies often sidestep the central issue by focusing on a single capability, such as using understanding signals to improve generation quality [13], or by adopting multi-stage schemes that alternate between tasks to find a fragile balance [14]. While effective to a degree, these methods do not directly tackle the core source of conflict: the lack of data-level semantic alignment between understanding and generation supervision and the absence of an optimization rule that respects this alignment. As a result, the shared policy is driven by competing gradients from unrelated signals, leading to unstable updates and uneven performance gains across tasks. This optimization-level conflict is not merely theoretical; it is empirically measurable and provides the direct motivation for our work.

A simple empirical regularity further motivates this study. When the data fed to understanding and generation are semantically aligned, the cosine similarity between their gradients, $\cos\big(\nabla_\theta \mathcal{L}_U,\ \nabla_\theta \mathcal{L}_G\big)$, increases, and higher gradient agreement correlates with stronger downstream results on MMMU, MMStar, and GenEval [5, 10, 53]. In contrast, randomly mixed or weakly matched samples reduce gradient agreement and hinder both objectives. This points to a data–optimization mismatch in unified RL: understanding and generation rely on heterogeneous supervision and data formats, and updating a shared policy without respecting semantic correspondence drives the model in conflicting directions, particularly in Janus-style architectures with a shared LLM [6]. The working hypothesis is that data alignment, rather than dataset size alone, is a key lever for balancing objectives, and that credit assignment should reflect the strength of that alignment.

We address this with PairUni, a unified RL framework that aligns the problem at both the data and optimization levels. On the data side, we reorganize heterogeneous supervision into understanding–generation (UG) pairs centered on the same or closely related images. Two complementary pair types are constructed. **Aligned pairs** are formed by completing single-task samples into unified quadruples, using GPT-o3 to add the missing caption or prompt for understanding data or to synthesize a question–answer pair for generation data, so that both objectives share the same instance. We use clustering to select representative high-quality medoids from the unified quadruples. **Retrieved pairs** link a generation sample to a semantically related understanding sample via similarity search over image embeddings, which expands coverage when exact matches are scarce. This paired view exposes cross-task correspondences, namely what to attend to for understanding and how to express it for generation, on related content rather than on unrelated batches.

On the optimization side, we develop Pair-GRPO, a pair-aware variant of GRPO that modulates the advantage by pair similarity. Aligned pairs receive full weight, and retrieved pairs are down-weighted by their pair-similarity scores. This mechanism strengthens updates from high-quality supervision and tempers weaker matches, which reduces cross-task interference while preserving GRPO stability through the clipped objective and KL regularization. In effect, the policy update is made to respect the semantic alignment already present in the data.

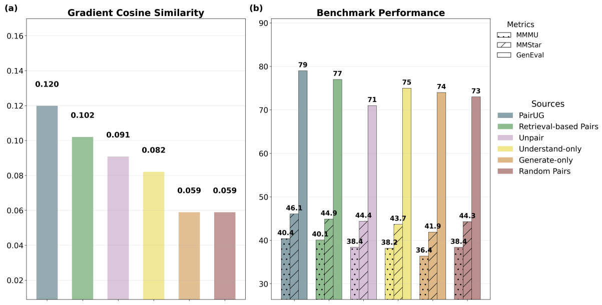
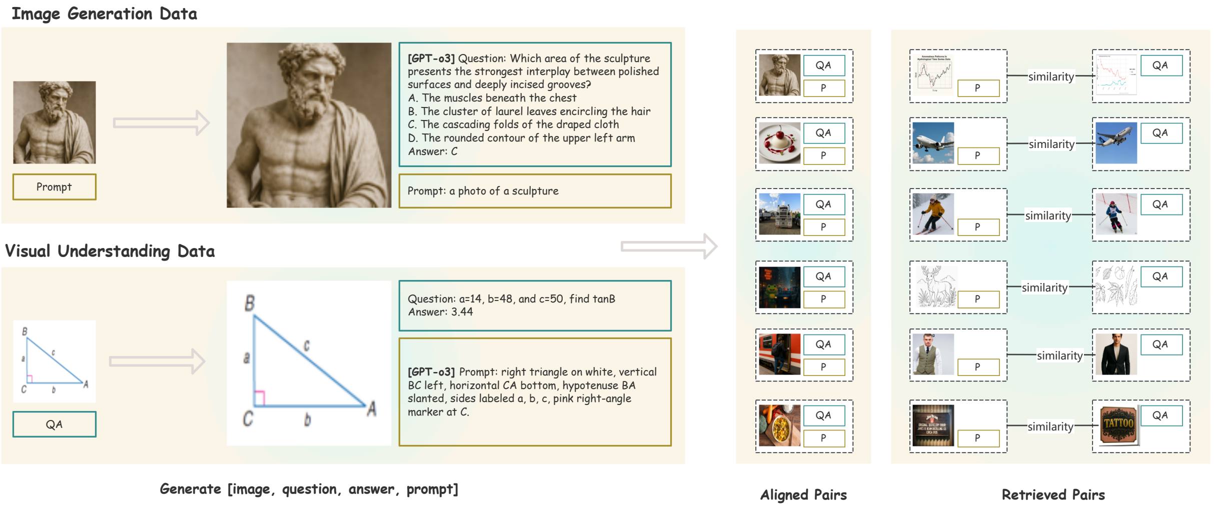


Figure 1 Performance Conflict Mechanism Analysis: Median gradient cosine similarity scores between understanding and generation components, alongside benchmark performance on two understanding benchmarks (MMMU [53], MMStar [5]) and one image generation benchmark (GenEval [10]). The analysis encompasses six distinct data combination scenarios: PairUG, Retrieval-based Pairs, Unpair data with low similarity scores, pure Generation-only data, pure Understanding-only data, and Random Pairs.

Our contributions are threefold. First, we introduce PairUni, a unified framework that reorganizes multimodal data into UG pairs (aligned and retrieved) and aligns optimization accordingly. We propose Pair-GRPO, a pair-aware GRPO variant that uses similarity scoring to modulate advantages, en-



[GPT-o3] Question: Which area of the sculpture presents the strongest interplay between polished surfaces and deeply incised grooves?
A. The muscles beneath the chest
B. The cluster of laurel leaves encircling the hair
C. The cascading folds of the draped cloth
D. The rounded contour of the upper left arm
Answer: C

Prompt: a photo of a sculpture

Question: a=14, b=48, and c=50, find tanB
Answer: 3.44

[GPT-o3] Prompt: right triangle on white, vertical BC left, horizontal CA bottom, hypotenuse BA slanted, sides labeled a, b, c, pink right-angle marker at C.

Figure 2 Data Pairing Pipeline. Left: examples of aligned quadruples from generation and understanding tasks. Right: pairing strategy using retrieval and clustering.

hancing learning from well-aligned examples while mitigating task interference between understanding and generation objectives. Second, we create a high-quality dataset of 16k PairUG for RL fine-tuning. Third, extensive experiments with Janus-Pro UVLMs demonstrate that PairUni achieves balanced improvements in both understanding and generation tasks, outperforming strong UVLMs RL baselines. Additional studies show transfer to a discrete diffusion backbone (Lumina-DiMOO [40]), indicating that the paired-data design and Pair-GRPO generalize across architectures and training regimes. These findings support a simple conclusion: pairing the data and weighting the advantage by pair-similarity is a general and effective ingredient for unified multimodal training, which improves understanding and generation together rather than trading one for the other.

2 Method

PairUni consists of two key components as shown in Figure 3: a data pairing pipeline (as shown in Figure 2) to generate training pairs for unified models (Section 2.1) and a Pair-GRPO algorithm (Section 2.2), which is specially designed for RL of understanding-generation pairs.

2.1 Paring Understanding–Generation Pairs

We aim to construct a unified paired dataset $\mathcal{S} = \{(I, C, Q, A)\}$, where each data item supports both generation and understanding capabilities within a single multimodal model. I is an image input, C is a text prompt that describes or motivates the image (used in generation), Q is a visual understanding question and A is the corresponding answer to Q. We build this paired dataset from two distinct sources: 1) **Understanding data** $\mathcal{U} = \{(I, Q, A)\}$, where the image is annotated with comprehension questions; 2) **Generation data** $\mathcal{G} = \{(I, C)\}$, where an image is paired with a generative prompt. These

Algorithm 1 Data Pairing Algorithm

```
Input: features $\mathcal{F}_u, \mathcal{F}_g$, clusters $K$, neighbors $n$
$\mathcal{F} \leftarrow \text{L2Norm}(\mathcal{F}_u \cup \mathcal{F}_g)$
$\mathcal{C} \leftarrow \text{MiniBatchKMeans}(\mathcal{F}, K)$
for $k = 1..K$ do
    $\mathcal{I}_k \leftarrow \{i : \text{assign}(x_i) = k\}$
    $i^* \leftarrow \arg\max_{i \in \mathcal{I}_k} \langle f_i, c_k \rangle$
    $\mathcal{D}_{\text{aligned}} \leftarrow \mathcal{D}_{\text{aligned}} \cup \{(x_{i^*}, y_{i^*})\}$
end for
$\mathcal{F}_u^{\text{rem}} \leftarrow \mathcal{F}_u \setminus \mathcal{D}_{\text{aligned}}$; $\mathcal{F}_g^{\text{rem}} \leftarrow \mathcal{F}_g \setminus \mathcal{D}_{\text{aligned}}$
for $x_i^g \in \mathcal{F}_g^{\text{rem}}$ do
    $\mathcal{J} \leftarrow \text{top-}n\ \text{kNN}(\phi_i^g, \mathcal{F}_u^{\text{rem}})$
    $\mathcal{D}_{\text{ret}} \leftarrow \mathcal{D}_{\text{ret}} \cup \{(x_j^u, x_i^g) : j \in \mathcal{J}\}$;
    $\mathcal{F}_u^{\text{rem}} \leftarrow \mathcal{F}_u^{\text{rem}} \setminus \mathcal{J}$, $\mathcal{F}_g^{\text{rem}} \leftarrow \mathcal{F}_g^{\text{rem}} \setminus \{x_i^g\}$
end for
Output: $\mathcal{D}_{\text{aligned}} \cup \mathcal{D}_{\text{ret}}$
```

two sources are inherently heterogeneous and rarely aligned. To unify them, we design a data pipeline that constructs either *aligned* or *retrieval-based* pairs.

2.1.1 Generating Aligned Pairs

Give U and G, we first augment each element in the the original datasets to the desired quratuples and then we design an algorithm to select representative pairs to construct $\mathcal{S}$.

Data augmentation For each understanding-only sample $(I, Q, A) \in \mathcal{U}$, we generate a compatible generation prompt C using GPT-o3, based on a template (as shown in Figure 11 in the appendix) that highlights salient image features. Similarly, for each generation-only sample $(I, C) \in \mathcal{G}$, we generate a question–answer pair (Q, A) using GPT-4o (as shown in Figure 10 in the appendix). Prompts are grouped by question type (e.g., multiple-choice, open-ended), and task-specific templates guide the generation process. Examples and prompts are provided in Appendix E. This yields a set of **aligned pairs**, where all four fields (I, C, Q, A) originate from a single source instance. These pairs eliminate semantic drift between tasks and ensure that both reward signals are grounded in the same visual context.

Data Selection Even after augmentation, many samples may be redundant or low-quality. To ensure coverage and diversity, we adopt a clustering-based strategy to identify representative aligned pairs ($\mathcal{D}_{\text{aligned}}$) from these datasets as shown in the first part of Algorithm 1. First, we extract image features using pretrained visual encoders, and then perform K-means clustering over the joint visual feature space of $\mathcal{U} \cup \mathcal{G}$. Second, for each cluster, we select the most central sample (i.e., the medoid) as a canonical representation of that cluster's content. This yields a curated set of self-referential pairs where understanding and generation annotations coexist for the same image. These samples are both semantically rich and geometrically representative of the data distribution, forming a strong backbone for joint training.

2.1.2 Constructing Retrieval-Based Pairs

While aligned pairs are semantically precise, their quantity is limited. To scale supervision, we introduce **retrieval-based pairs** $\mathcal{D}_{\text{ret}}$, where understanding and generation samples come from different images but share visual similarity. The second part of Algorithm 1 shows the algorithm. The main idea is to extract visually similar image pairs from two datasets to establish correspondences between "understand" and "generate" data. First, cosine similarity is computed across all remaining generation–understanding image pairs. For each generation image, we retrieve the top-n most similar understanding images above a similarity threshold δ. A greedy matching algorithm is used to ensure that each image is only used once. This retrieval mechanism exploits the fact that **semantically similar images often support related tasks**, even if not identical. By leveraging these approximate matches, the model can learn **cross-instance generalization**, which enhances its robustness and expands training coverage.

Together, $\mathcal{D}_{\text{aligned}}$ and $\mathcal{D}_{\text{ret}}$ form the UG pair set $\mathcal{S}$ used for policy optimization. These two pathways provide complementary benefits: aligned pairs deliver precise, high-quality supervision, while retrieval pairs enhance scale and semantic diversity, covering a wide spectrum of multimodal understanding and generation tasks.

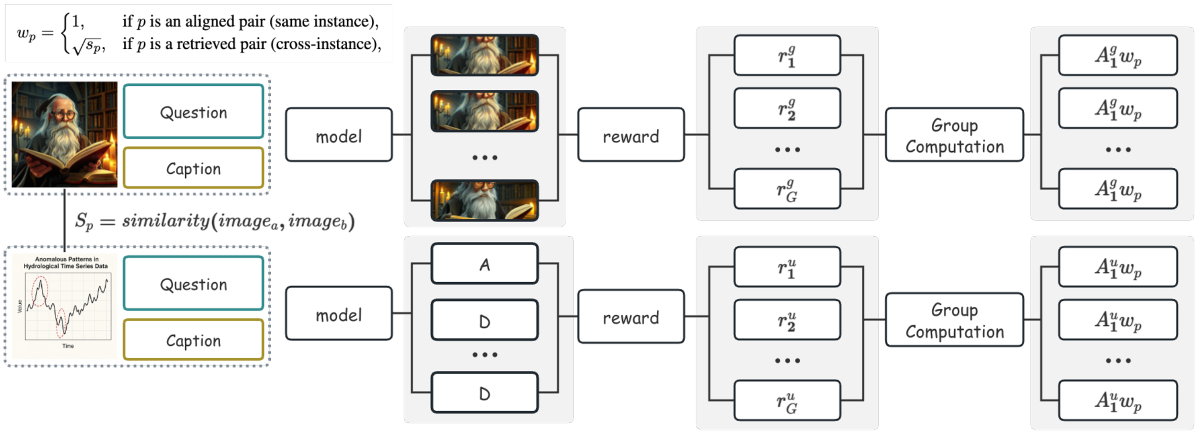
2.2 Pair-GPRO

This section describes: (1) vanilla GRPO with mixed tasks; (2) pairwise GRPO with UG-pairs; (3) Pair-GPRO, which incorporates pair similarity into advantage weighting. Each step is designed to better align understanding and generation, reduce conflict, and stabilize learning.

(1) Vanilla GRPO with mixed tasks. We form a batch $\mathcal{B} = \{\tau_i\}_{i=1}^{N}$ of N trajectories, where each trajectory τ_i is either an understanding or generation task. Each trajectory corresponds to output tokens $o_{1:T}$ given a prompt/input q, under the current policy π_θ and past policy $\pi_{\theta_{\text{old}}}$. For each token at timestep z, define the importance ratio

$$\rho_t(\theta) = \frac{\pi_\theta(o_z \mid q, o_{<z})}{\pi_{\theta_{\text{old}}}(o_z \mid q, o_{<z})}.$$

We compute a scalar reward r per trajectory (r_{Und} if it is an understanding sample, r_{Gen} if generation). Trajectories that share the same prompt q are grouped into sets of size G. Within each group, we normalize



$w_p = \begin{cases} 1, & \text{if } p \text{ is an aligned pair (same instance)}, \\ \sqrt{s_p}, & \text{if } p \text{ is a retrieved pair (cross-instance)}, \end{cases}$
Question
Caption
$S_p = similarity(image_a, image_b)$
Anomalous Patterns in Hydrological Time Series Data
Value
Time
Question
Caption
model
model
A
D
D
reward
reward
r_1^g
r_2^g
r_G^g
r_1^u
r_2^u
r_G^u
Group Computation
Group Computation
$A_1^g w_p$
$A_1^g w_p$
$A_1^g w_p$
$A_1^u w_p$
$A_1^u w_p$
$A_1^u w_p$

Figure 3 **Framework of PairUni**: A dual-component design integrating a data processing pipeline and the GRPO reinforcement learning algorithm.

the rewards by subtracting the group mean μ_r and dividing by the group standard deviation σ_r, obtaining a group-relative advantage

$$\widehat{A}_t = \frac{r - \mu_r}{\sigma_r},$$

which is then applied to all tokens in that trajectory. The vanilla GRPO objective is then

$$J_{\text{vanilla}}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\sum_{t=1}^{T} \min\big(\rho_t(\theta)\, \widehat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon)\ \widehat{A}_t\big) \right] \; - \; \beta\, D_{\text{KL}}\big(\pi_\theta \parallel \pi_{\text{old}}\big),$$

where ε is the clipping threshold limiting how much the policy can change per token, and β is the coefficient for a KL divergence term $D_{\text{KL}}(\pi_\theta \| \pi_{\text{old}})$, which prevents π_θ from drifting too far from π_{old}. In default settings, we set β zero to emphasize the clipped term [52].

Reward Functions Our reward functions are tailored to the specific goals of understanding and generation. For understanding tasks, typically formulated as multiple-choice question answering, we use prediction accuracy as our reward, a standard metric that directly measures correctness: $r_{\text{Und}} = \text{Acc}(y_{\text{pred}}, y_{\text{true}})$.

For generation tasks, we employ the HPSv2 reward model [46] to evaluate output quality. HPSv2 is a fine-tuned CLIP model [35] that effectively scores the alignment of generated content with human aesthetic preferences, showing strong performance in prior work [13, 23]. The reward is given by:$r_{\text{Gen}} = R_{\text{HPSv2}}(x, y_{\text{gen}})$, where x is the input prompt and y_{gen} is the corresponding generated image.

(2) Pairwise GRPO with UG data pairs. To reduce conflict between understanding and generation tasks, we reorganize training around a set of paired training examples $\mathcal{P} = \{p\}_{p=1}^{M}$. Each pair p consists of two datapoints: one generation example and one understanding example that are semantically aligned. This pairing is defined at the data level, not at the trajectory level: each data item in the pair can produce multiple trajectories through sampling.

For each paired datapoint p, we generate a set of trajectories $\{\tau_{p,k}^{(u)}\}_{k=1}^{K_u}$ for the understanding side, and $\{\tau_{p,k}^{(g)}\}_{k=1}^{K_g}$ for the generation side. These trajectories are grouped by task type and prompt to compute group-relative advantages. Specifically, we calculate $\widehat{A}_t^{(u)}$ and $\widehat{A}_t^{(g)}$ as the normalized reward within the respective task-type group, using GRPO's group-based normalization.

The pairwise GRPO objective is then defined as:

$$\begin{aligned}\max_{\theta}\ \mathbb{E}_{p\sim\mathcal{P}}\Bigg[&\sum_{\tau\in\{\tau_p^{(u)}\}}\sum_{t\in\tau}\min\big(\rho_t\,\widehat{A}_t^{(u)},\ \mathrm{clip}(\rho_t,1-\varepsilon,1+\varepsilon)\,\widehat{A}_t^{(u)}\big)+\\ &\sum_{\tau\in\{\tau_p^{(g)}\}}\sum_{t\in\tau}\min\big(\rho_t\,\widehat{A}_t^{(g)},\ \mathrm{clip}(\rho_t,1-\varepsilon,1+\varepsilon)\,\widehat{A}_t^{(g)}\big)\Bigg].\end{aligned} \tag{1}$$

This formulation ensures that policy gradients from understanding and generation are derived from semantically related training examples, even though multiple trajectories may be sampled per side. This pairing encourages more consistent policy updates across tasks.

(3) Pair-GPRO: similarity-weighted advantage. To further align training strength with semantic similarity, we introduce a pair-level similarity score $s_p \in [0,1]$ for each data pair p, computed via a pretrained image encoder. Based on this, we assign a pair weight w_p:

$$w_p = \begin{cases} 1, & \text{if } p \text{ is an aligned pair (same instance)},\\ \sqrt{s_p}, & \text{if } p \text{ is a retrieved pair (cross-instance)}, \end{cases} \tag{2}$$

and apply this weight to all advantages computed from trajectories originating from the pair:

$$\widetilde{A}_t^{(u)} = w_p\,\widehat{A}_t^{(u)}, \qquad \widetilde{A}_t^{(g)} = w_p\,\widehat{A}_t^{(g)}.$$

We use the square root to amplify the relative differences between similarity scores, as all selected pairs are drawn from a high-similarity candidate pool. The full PairGPRO objective becomes:

$$\begin{aligned}\max_{\theta}\ \mathbb{E}_{p\sim\mathcal{P}}\Bigg[&\sum_{\tau\in\{\tau_p^{(u)}\}}\sum_{t\in\tau}\min\big(\rho_t\,\widetilde{A}_t^{(u)},\ \mathrm{clip}(\rho_t,1-\varepsilon,1+\varepsilon)\,\widetilde{A}_t^{(u)}\big)+\\ &\sum_{\tau\in\{\tau_p^{(g)}\}}\sum_{t\in\tau}\min\big(\rho_t\,\widetilde{A}_t^{(g)},\ \mathrm{clip}(\rho_t,1-\varepsilon,1+\varepsilon)\,\widetilde{A}_t^{(g)}\big)\Bigg].\end{aligned} \tag{3}$$

This design modulates the trajectory-level credit assignment based on the quality of semantic pairing, strengthening updates from well-aligned pairs (aligned: $w_p = 1$) while attenuating noisy or weakly aligned ones (retrieved: $w_p = \sqrt{s_p}$). As a result, PairGPRO retains the stability of GRPO while better aligning cross-task gradients with semantic structure in the data.

3 Experiments

3.1 Experiment Settings

Training We adopt Janus-Pro [6] as a primary baseline because it is widely used as a comparator for unified multimodal understanding and generation and exhibits competitive performance. All experiments are conducted on 8× H100 GPUs. For the 7B model, we use a rollout size of 4 for both text and image generation and train for at most 1200 steps. The per-device batch size is 2 (global batch size 16). We set the classifier-free guidance (CFG) weight to 5, $\beta = 0$, the learning rate to 1×10^{-6}, and the sampling temperature to 1.0. For the 1B model, the rollout size for both modalities is increased to 8. All the visual features are extracted using a ResNet50 encoder[1] by removing the classification head [16] and L2-normalized.

We use the Orsta data [26][2] as the understanding data $\mathcal{U}$, which contains about 47K samples, and the BLIP3o data [3][3] as the generation dataset $\mathcal{G}$, which contains about 60K samples. We exclude the original

---

[1] https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html

[2] https://huggingface.co/datasets/One-RL-to-See-Them-All/Orsta-Data-47k

[3] https://huggingface.co/datasets/BLIP3o/BLIP3o-60k

Table 1 Main Results on multimodal understanding benchmarks. The PairUni method achieves the best performance across tasks and model sizes. The model displayed by $ULM - R1^{\dagger}$ reports results after training with UnifiedRL.

Model	LLM	MMMU	MMStar	MME(P)	POPE
Understanding Only					
InternVL3 [56]	Qwen2.5-1.5B [1]	48.6	**60.7**	-	**89.6**
Qwen2.5-VL [1]	Qwen2.5-3B	**51.2**	56.3	-	85.9
LMM-R1 [31]	Qwen2.5-3B	-	58.0	-	-
Unified Understanding and Generation					
Show-o[48]	Phi-1.3B [11]	26.7	-	-	80.0
HermesFlow [51]	Phi-1.3B	28.3	-	-	81.4
Janus-Pro-1B [6]	DeepSeek-LLM-1.5B [7]	36.3	-	1444.0	86.2
$ULM - R1^{\dagger}$ [14]	DeepSeek-LLM-1.5B	40.3	-	-	-
PairUni-1B	DeepSeek-LLM-1.5B	**40.4**	**46.4**	**1483.2**	**86.4**
Orthus [17]	Chameleon-7B [41]	28.2	-	1265.8	79.6
VILA-U [47]	LLaMA-2-7B [42]	-	-	-	85.8
UniToken [15]	Chameleon-7B	32.8	46.1	-	-
Janus-Pro-7B [6]	DeepSeek-LLM-7B	41.1	46.5	1567.1	87.4
DSR [12]	DeepSeek-LLM-7B	41.1	-	-	86.6
PairUni-7B	DeepSeek-LLM-7B	**47.0**	**49.5**	**1597.7**	**88.0**

detection and grounding QA pairs from Orsta since Janus-Pro fails on these tasks. The similarity threshold is 0.6. Our constructed PairUG dataset consists of 16,320 samples: including 4,971 aligned pairs and 11,349 retrieval-based pairs. The similarity score of each pair is provided.

Evaluation To comprehensively evaluate both multimodal understanding and generation capabilities, we adopt a range of established benchmarks. For understanding, MME Perception probes basic perceptual reasoning via binary (yes/no) questions; POPE [20] quantifies sensitivity to hallucinations through fine-grained visual grounding; MMStar [5] comprises multiple-choice questions that assess visual reasoning across diverse scenarios; and MMMU [53] offers a challenging suite spanning subjects such as mathematics and chemistry, requiring models to integrate information from multiple images and reason over complex semantics. For generation, we evaluate on GenEval [10], which assesses visual fidelity along dimensions such as object count, color consistency, and spatial arrangement, and Wise [29], which emphasizes knowledge-grounded image synthesis by evaluating the model's ability to generate semantically coherent and factually plausible visual content grounded in real-world knowledge.

3.2 Main Results

Multimodal Understanding. Table 1 summarizes the performance of various models on four representative multimodal understanding benchmarks. On the broad-coverage MMMU benchmark, which stresses general visual reasoning across science, math, and commonsense, PairUni delivers clear gains among unified models: at the 1B scale it attains 40.4 (vs. Janus-Pro-1B: 36.3; ULM-R1†: 40.3), and at the 7B scale it reaches 47.0, outperforming prior unified baselines such as Janus-Pro-7B (41.1) and UniToken (32.8). Beyond MMMU, PairUni also improves MMStar and MME: at 7B, it moves MMStar from 46.5 to 49.5 and MME from 1567.1 to 1597.7, while maintaining competitive POPE (88.0 vs. 87.4), showing gains in both perception-heavy and knowledge-heavy tasks. At the 1B scale, PairUni achieves **46.4** on MMStar and **1483.2** on MME with stable POPE (86.4), indicating that the data pairing and Pair-GPRO optimization are effective even under tight capacity budgets. The consistent improvements of PairUni suggest that PairUni strengthen unified visual reasoning without sacrificing perception robustness.

Text-to-Image Generation. We evaluate text-to-image generation on the WISE [29] and Geneval [10] benchmarks. On WiSE, as shown in Table 2, PairUni sets the best overall scores among unified models at both scales: 0.38 (1B) and 0.45 (7B). At 1B, it improves over Janus-1B (0.23) and ULM-R1 (0.33), and at 7B it surpasses Janus-Pro-7B (0.35) and Emu3 (0.39), effectively narrowing the gap with generate-only models. ULM-R1 [14] is not open-sourced, so we report its numbers only after Unified-RL and Refined-RL

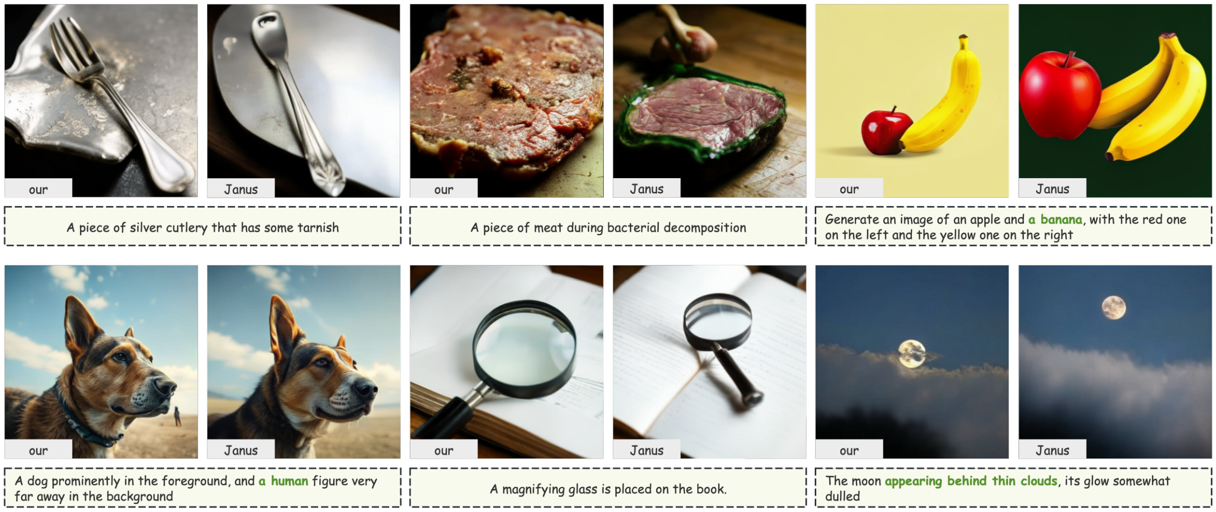

our
Janus
A piece of silver cutlery that has some tarnish
our
Janus
A piece of meat during bacterial decomposition
our
Janus
Generate an image of an apple and a banana, with the red one on the left and the yellow one on the right
our
Janus
A dog prominently in the foreground, and a human figure very far away in the background
our
Janus
A magnifying glass is placed on the book.
our
Janus
The moon appearing behind thin clouds, its glow somewhat dulled

Figure 4 **Case Study**: The generated image of Janus-Pro-7B and PairUni

training with approximately 40k samples. Subtask-wise, PairUni notably strengthens Space (0.56 at 1B; 0.62 at 7B) and Physics (0.44 at 1B; 0.55 at 7B), indicating better geographic and physical commonsense grounding. On GenEval (Table 3), PairUni exhibits strong compositional generalization at both scales. At 1B, PairUni-1B attains the best overall score (0.79), surpassing Janus-Pro-1B (0.73) and ULM-R1 (0.76). At 7B, PairUni-7B reaches 0.85, outperforming Janus-Pro-7B (0.79) and DSR (0.84), the latter showing comparatively weaker understanding and a notably underperforming 1B variant. Although Janus-Pro-R1 slightly exceeds PairUni-7B by one point (0.86), its understanding results and WISE scores are not reported. These gains align with our improvements on WISE and suggest that Pair-GPRO better enforces constraint following and object–attribute binding.

Taken together with the understanding results, these generation benchmarks indicate that PairUni improves **both** sides of the unified objective: it raises reasoning-heavy understanding performance while strengthening generation fidelity and knowledge. By contrast, understanding-focused baselines (e.g., InternVL3, Qwen2.5-VL) show strong comprehension but are not evaluated for generation; some unified baselines with competitive generation (e.g., DSR at 7B) exhibit weaker understanding; and diffusion-only generators (e.g., FLUX.1-dev) perform well on WISE yet offer limited understanding capability. By coupling paired data with Pair-GPRO, PairUni delivers balanced, cross-task gains without sacrificing either modality; qualitative cases in Figure 4 illustrate improved adherence to textual constraints and tighter semantic alignment.

3.3 Ablation Studies

Effect of Data Pairing. Across the **different pairing pipelines** in Table 4, we can observe three trends using Janus-Pro-1B. First, **single–source** training (either understanding–only or generation–only) yields specialization but not balance: models inherit strengths from their source (MMMU/MMStar or GenEval) while lagging on the complementary objective. Second, naive mixtures are insufficient: **unpaired** and **random** pairing under the same budget depress generative fidelity (e.g., 0.71–0.73 on GenEval), indicating that gradients from semantically unrelated samples behave like low–advantage noise. Figure 5 shows that compared to random pairs, our PairUG dataset leads to more stable training dynamics. Third, **pair–aware supervision** matters: aligned–based pairs outperform random pairing, and retrieval–based pairs are effective for both understanding and generation tasks.

Our proposed PairUG combines both pair types and scales effectively. At 7.5K examples ($\mathcal{U}\simeq$3.5K, $\mathcal{G}\simeq$4.0K, $\sim$761 aligned pairs), PairUG (7.5K) already matches or exceeds aligned–only training on understanding while preserving GenEval (0.76). Scaling the paired set to 16K (4,971 aligned + 11,349 retrieval pairs) delivers

Table 2 Results on WISE. PairUni achieves the highest overall score, with particularly outstanding performance in the space and physics subtasks.

Model	**Cultural↑**	**Time↑**	**Space↑**	**Biology↑**	**Physics↑**	**Chemistry↑**	**Overall**
Generating Only							
PixArt-α [4]	0.45	0.50	0.48	**0.49**	**0.56**	0.34	0.47
playground-v2.5 [19]	**0.49**	**0.58**	0.55	0.43	0.48	0.33	0.49
SD-v1-5 [37]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SD-XL-base-0.9 [32]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
FLUX.1-dev [18]	0.48	**0.58**	**0.62**	0.42	0.51	**0.35**	**0.50**
Unified Understanding and Generation							
VILA-U [47]	0.26	0.33	0.37	0.35	0.39	**0.23**	0.31
Janus-1B [6]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
ULM – R1† [14]	-	-	-	-	-	-	0.33
PairUni-1B	**0.31**	**0.39**	**0.56**	**0.38**	**0.44**	0.22	**0.38**
Emu3 [44]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Janus-Pro-7B [6]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
PairUni-7B	**0.36**	**0.46**	**0.62**	**0.42**	**0.55**	**0.29**	**0.45**

the best overall results across MMMU, MMStar, and GenEval. These findings support our central claim: **properly aligned** supervision (aligned and high–similarity retrieved pairs), rather than siloed or noisy mixtures, is essential for a unified model that improves **both** understanding and generation. Given the monotonic gains from 7.5K to 16K, we expect further improvements with larger, quality–controlled paired corpora.

Effect of Similarity-based Advantage Adjustment. Table 5 isolates the effect of similarity weighting in Pair-GPRO. At the 1B scale, adding pair similarity to the advantage computation yields consistent gains on understanding-heavy metrics while leaving generation quality unchanged: MME(P) rises from 1469.87 to 1483.18 (+13.31) and MMStar from 45.1 to 46.1 (+1.0), with GenEval fixed at 0.79. At 7B, we observe a similar pattern: MME(P) improves from 1554.91 to 1597.71 (+42.80) and MMStar from 47.7 to 49.5 (+1.8), while GenEval remains 0.85 and MMMU holds at 47.0. These trends are consistent with the intended role of similarity weighting. Without it, aligned and retrieved pairs contribute equally, so updates from weakly matched pairs can dilute task-specific signals and increase cross-task interference. Weighting the advantage by pair similarity amplifies well-matched supervision and attenuates noisier pairs, yielding measurable gains precisely on the metrics most sensitive to grounding and instruction-following (MME(P) and MMStar) while avoiding regressions in generation fidelity (GenEval) and preserving broad reasoning (MMMU). In short, similarity-aware credit assignment provides a simple, robust mechanism to trade up understanding quality without paying a cost on generation.

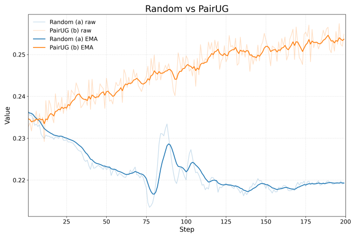


Figure 5 Training rewards of PairUG and random pairs.

Evaluation beyond autoregressive transformers. To assess architectural generality beyond autoregressive transformers, we evaluate PairUni on Lumina-DiMOO, a discrete diffusion model with the state-of-the-art results on unified tasks (Table 6). Applying PairUni yields consistent improvements: MMMU increases from 58.6 to 61.3 (+2.7), MMStar from 52.4 to 52.6 (+0.2), and GenEval from 0.88 to 0.89 (+0.01). These results indicate that pairing-based data construction and Pair-GPRO enhance both understanding and generation under a markedly different generative mechanism, supporting applicability across model families.

4 Conclusion

This paper introduces PairUni, a reinforcement learning framework for UVLMs that aligns understanding and generation via paired training signals, and PairUG, a curated dataset of understanding–generation pairs that

Table 3 Results on GenEval. PairUni achieves the top result among 1B-scale models, and in the 7B-scale category, it demonstrates competitive performances.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall
Generating Only							
PixArt-α [4]	**0.98**	0.50	0.44	0.80	0.08	0.07	0.48
SDXL [32]	**0.98**	0.74	0.39	**0.85**	0.15	0.23	0.55
DALL-E 3 [33]	0.96	**0.87**	**0.47**	0.83	**0.43**	**0.45**	**0.67**
Unified Understanding and Generation							
SEED-X [9]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
Show-o [48]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
ILLUME [43]	**0.99**	0.86	0.45	0.71	0.39	0.28	0.61
HermersFlow [51]	0.97	0.67	**0.65**	0.77	0.28	0.42	0.61
UniRL [27]	0.95	0.74	0.27	0.81	0.62	0.52	0.65
Janus-Pro-1B [6]	**0.99**	0.82	0.48	**0.90**	0.62	0.57	0.73
ULM-R1 [14]	-	-	-	-	-	-	0.76
Janus-Pro-R1 [30]	0.98	0.80	0.51	0.84	0.59	0.55	0.71
PairUni-1B	0.98	**0.91**	0.44	0.75	**0.95**	**0.69**	**0.79**
Chameleon [41]	–	–	–	–	–	–	0.39
D-DiT [21]	0.97	0.80	0.54	0.76	0.32	0.50	0.65
LWM [22]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
Transfusion [55]	–	–	–	–	–	–	0.63
TokenFlow-XL [34]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus-Pro-7B [6]	0.97	0.88	0.57	0.90	0.77	0.64	0.79
DSR [12]	–	–	–	–	–	–	0.84
Janus-Pro-R1 [30]	**0.99**	**0.94**	0.66	0.92	0.87	**0.78**	**0.86**
PairUni-7B	0.97	0.75	**0.78**	**0.97**	**0.91**	0.69	0.85

Table 4 Ablation Study of Pairing Algorithm

Model	MMMU	MMStar	GenEval
Pairs from $\mathcal{U}$ only	38.2	43.7	0.75
Pairs from $\mathcal{G}$ only	36.4	41.9	0.74
Unpair	38.4	44.4	0.71
Random Pair	38.4	44.3	0.73
Aligned-based Pairs	39.2	44.6	0.76
Retrieval-based Pairs	40.1	44.9	0.77
PairUG (7.5K)	39.6	43.7	0.76
PairUG (16k)	**40.4**	**46.1**	**0.79**

Table 5 Trajectory-level credit assignment.

Model	MME(P)	MMMU	MMStar	GenEval
PairUni-1B w/o sim	1469.87	40.0	45.1	0.79
PairUni-1B	**1483.18**	**40.4**	**46.1**	0.79
PairUni-7B w/o sim	1554.91	47.0	47.7	0.85
PairUni-7B	**1597.71**	47.0	**49.5**	0.85

Table 6 Lumina-DiMOO performance

Model	MMMU	MMStar	GenEval
Lumina-DiMOO [40]	58.6	52.4	0.88
Lumina-DiMOO w PairUni	**61.3**	**52.6**	**0.89**

supports consistent policy learning. On standard UVLMs evaluations with Janus-Pro backbones, our approach achieves strong, balanced improvements in both understanding and generation, surpassing competitive RL baselines.

References

[1] 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.

[2] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. arXiv:2504.13181, 2025.

[3] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568, 2025.

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems, 37:27056–27087, 2024.

[6] 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, 2025. URL https://arxiv.org/abs/2501.17811.

[7] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[8] 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.

[9] 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.

[10] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 36:52132–52152, 2023.

[11] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. arXiv preprint arXiv:2306.11644, 2023.

[12] Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. Reinforcing multimodal understanding and generation with dual self-rewards. arXiv preprint arXiv:2506.07963, 2025.

[13] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703, 2025.

[14] Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation, 2025. URL https://arxiv.org/abs/2505.17534.

[15] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding, 2025. URL https://arxiv.org/abs/2504.04423.

[16] Brett Koonce. Resnet 50. In Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, pages 63–72. Springer, 2021.

[17] Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads, 2025. URL https://arxiv.org/abs/2412.00127.

[18] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

[19] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

[20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023.

[21] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2779–2790, 2025.

[22] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2025. URL https://arxiv.org/abs/2402.08268.

[23] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 8009–8019, 2025.

[24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.

[25] 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:2405.20797, 2024.

[26] Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One rl to see them all: Visual triple unified reinforcement learning. arXiv preprint arXiv:2505.18129, 2025.

[27] Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning, 2025. URL https://arxiv.org/abs/2505.23380.

[28] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

[29] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265, 2025.

[30] Kaihang Pan, Yang Wu, Wendong Bu, Kai Shen, Juncheng Li, Yingting Wang, Yunfei Li, Siliang Tang, Jun Xiao, Fei Wu, Hang Zhao, and Yueting Zhuang. Unlocking aha moments via reinforcement learning: Advancing collaborative visual comprehension and generation, 2025. URL https://arxiv.org/abs/2506.01480.

[31] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025. URL https://arxiv.org/abs/2503.07536.

[32] 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, 2023. URL https://arxiv.org/abs/2307.01952.

[33] MM Pooja, PM Sulfath, and KM Sheena. Dall-e 3: Advanced ai image generation model. Authorea Preprints.

[34] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2545–2555, 2025.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, June 2022.

[38] 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, 2024. URL https://arxiv.org/abs/2402.03300.

[39] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John

Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

[40] Alpha VLLM Team. Lumina-dimoo: A unified masked diffusion model for multi-modal generation and understanding, 2025. URL https://github.com/Alpha-VLLM/Lumina-DiMOO.

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

[42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[43] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. arXiv preprint arXiv:2412.06673, 2024.

[44] 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, 2024. URL https://arxiv.org/abs/2409.18869.

[45] Yinjie Wang, Ling Yang, Bowen Li, Ye Tian, Ke Shen, and Mengdi Wang. Revolutionizing reinforcement learning framework for diffusion large language models. arXiv preprint arXiv:2509.06949, 2025.

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

[47] 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, 2025. URL https://arxiv.org/abs/2409.04429.

[48] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528, 2024.

[49] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564, 2025.

[50] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809, 2025.

[51] Ling Yang, Xinchen Zhang, Ye Tian, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. Hermesflow: Seamlessly closing the gap in multimodal understanding and generation, 2025. URL https://arxiv.org/abs/2502.12148.

[52] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

[53] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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, pages 9556–9567, 2024.

[54] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities, 2025. URL `https://arxiv.org/abs/2505.02567`.

[55] 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 The Thirteenth International Conference on Learning Representations, 2025. URL `https://openreview.net/forum?id=SI2hI0frk6`.

[56] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL `https://arxiv.org/abs/2504.10479`.

Appendix

Overview. Appendix A covers related work on UVLMs and the RL methods used in UVLMs. Appendix B reports additional training results, including further model results and systematic ablation studies of the image extractor. Appendix C provides data-centric details, including summary statistics of the original data distribution, representative cases of retrieved pairs, and the empirical distribution of PairUG. Appendix D presents qualitative case studies on understanding tasks. Appendix E includes the prompts used by GPT-o3. The closing sections include notes on the use of large language models.

A Related Work

Unified Vision-Language Model. The pursuit of unified multimodal frameworks has led to significant innovations in both architecture design and training paradigms. Early approaches [41, 48] like Show-o series [48, 49] establish the autoregressive foundations for joint vision-language processing. Meanwhile, Transfusion [55] introduces diffusion-based methodologies to enhance generation quality. These foundational works, as systematically analyzed in [54], demonstrate the potential of unifying modalities through shared representation learning. Recent advances have pushed the boundaries of unified modeling [6, 8]. For example, Janus-Pro [6] innovatively uses bidirectional encoder-decoder structures for understanding and generation, achieving stronger performance on both sides. Bagel [8] adopts transformer experts and is trained with massive image generation and understanding data, leading to the state-of-the-art performance. This architectural evolution aligns with the broader trend of developing modular yet integrated systems that can dynamically adapt to different modalities. In this context, our work enhances UVLMs post-training, in particular, during the reinforcement learning phase. We present a novel view of pair data generation and utilization of such pair data with proposed Pair-GRPO.

Reinforcement Learning in UVLMs. The integration of reinforcement learning (RL) has emerged as a critical component for advancing unified MLLMs during the post-training. Early RL applications focused on modality-specific enhancements: step-by-step rule-based rewards for mathematical reasoning [38], and bbox IoU rewards for visual grounding [13]. For text-to-image generation, CLIP-based rewards [35] became standard for aligning visual outputs with textual descriptions. The paradigm shifted with unified RL approaches that exploit cross-modal synergies. T2I-R1 [13] pioneered iterative refinement through GRPO [38], using detailed descriptions as intermediate rewards. Recently, several works also explore the RL-based post-training for UVLMs. In particular, UniRL [27] proposed a self-improving pipeline where generated QA pairs simultaneously serve as training data and reward signals, though this approach showed performance degradation in understanding benchmarks. More sophisticated reward mechanisms have since been developed. DSR [12] introduced dual-source rewards combining original image-caption pairs with generated content, while HermesFlow [51] implemented Pair DPO [36] to enforce consistency between modalities. Notably, CoRL [14] adopts a two-stage approach, first training unified RL on shared data before specializing in understanding/generation phases, demonstrating improved performance across multiple benchmarks. Different from these methods, which all focus on unified RL method design, our work provides a new view on understanding data and generation data. We propose to build the UG pairs to benefit both tasks. With proposed Pair-GRPO along with UG pairs, our work improves various UVLMs.

B More Training Setting

B.1 Results on more models

As shown in Table 7, to assess the generalizability of our data and training method, we instantiate the Pair-GRPO framework on Lumina-DiMOO [40]. Unlike Janus-Pro [6], Lumina-DiMOO is a multimodal discrete diffusion model [50]. To enable parallel RL training for both understanding and generation, we set the rollout size to 2 and adopt a fixed-step diffusion sampling scheme: we use 2 diffusion steps for text generation and 35 steps for image generation. We build upon the diffusion Large Language Model (dLLM) [28] and its Proximal Policy Optimization (PPO) [45] implementation, extending the framework to a multimodal dLLM and adding an implementation of the GPRO algorithm for policy optimization. Because Lumina-DiMOO

Table 7 More models. SFT experiments on BAGEL and PairUni method on Lumina-DiMOO using PairUG data

Model	MME(P)	MMMU	MMStar	POPE	GenEval
Lumina-DiMOO	1534.2	58.6	52.4	87.4	0.88
Lumina-DiMOO [40] w PairUni	**1551.8**	**61.3**	**52.6**	**87.5**	**0.89**
BAGEL	1687.0	55.3/52.8*	-	87.4*	0.82/0.78*
BAGEL [8] + SFT	1642.0	**53.7**	-	**87.7**	**0.79**

* Results obtained from the publicly released checkpoint rather than the original reported numbers.

has not released its paper nor the official training and evaluation code, this integration is preliminary and may benefit from further refinement. In the current implementation, PairUni improves both the model's comprehension and generation capabilities.

BAGEL [8] is a strong UVLM that integrates both understanding and generation components with flow-based image generation and auto-regressive text-generation architectures. Since Bagel does not support RL training in default settings, we employ the PairUG dataset for supervised fine-tuning (SFT) on BAGEL. The SFT model yields consistent improvements on the POPE benchmark. While the scores on MMMU and GenEval do not surpass those of the original publication, our model still outperforms the public checkpoint baseline denoted as $*$ in Table 7, improving the MMMU score from 52.8 to 53.7 (+0.9) and the GenEval score from 0.78 to 0.79 (+0.1). The sole performance degradation occurred on the MME(P) benchmark, a result we attribute to a reduced effective token capacity during SFT with the PairUG dataset.

B.2 Results on different image extractor.

Table 8 Ablation study. Different image features extractor.

Model	MMMU	MMStar	GenEval
PE [2]	40.1	45.5	0.77
DINOv3 [39]	**40.4**	46.0	**0.79**
ResNet [16]	**40.4**	**46.1**	**0.79**

We evaluate three alternatives for the image feature extractor: the Perception Encoder (PE) [2], DINOv3 [39], and ResNet [16]. The PE is designed for high-level semantic understanding, yet it underperforms on both the understanding and generation benchmarks (Table 8). In contrast, DINOv3 and ResNet—both emphasizing visual feature similarity—achieve comparable results. These findings indicate that, in our setting, enforcing consistency with respect to visual similarity is more critical than modeling semantic abstraction.

C More details about data

C.1 Data Distribution of Understanding and Generation Data

We curate two complementary splits covering multimodal understanding and image generation. For understanding, we adopt Orsta-47k [26], a high-quality and diverse set that spans chart analysis, counting, object detection, grounding, mathematical reasoning, OCR, puzzles, and scientific reasoning. For image generation, BLIP-3o-60k [3] comprises curated AI-generated images paired with detailed textual descriptions, including single- and dual-object scenes as well as text-containing visuals. We deduplicate and ensure there is no overlap with the data used during pretraining.

To characterize their composition, we apply unsupervised clustering over the union of the two splits and examine cluster proportions (Figure 6). The two distributions exhibit pronounced divergence: categories prevalent in understanding data—such as math- or OCR-intensive items—are rare in the generation split, whereas descriptive object-centric scenes are comparatively overrepresented in generation.

C.2 Retrieved Pairs Cases

Figure 7 presents representative retrieved examples.

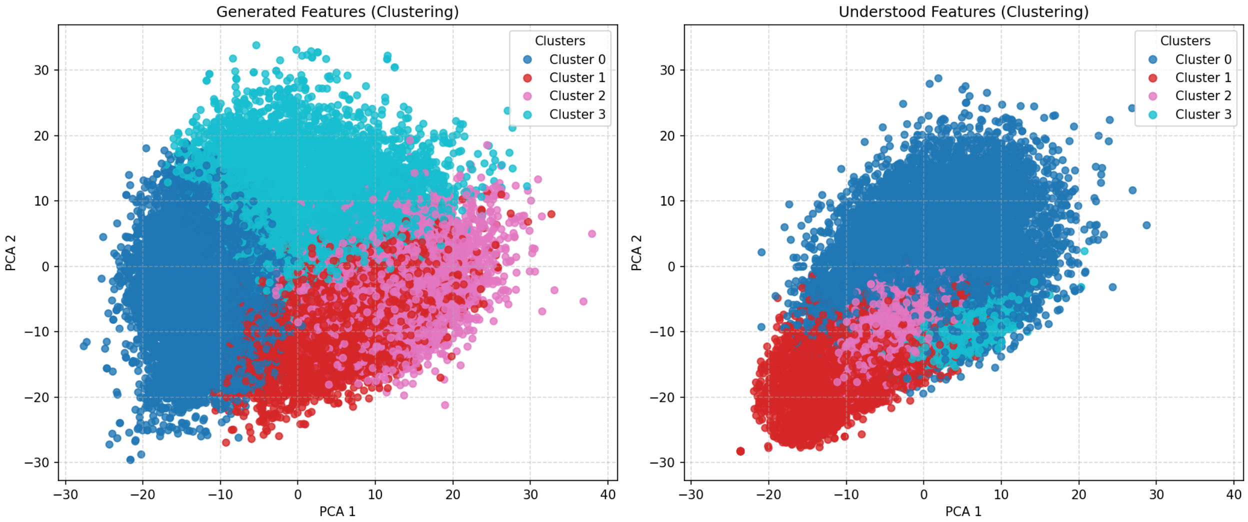


Figure 6 Distributional comparison between multimodal understanding and image generation data.

C.3 Distribution of PairUG

Figure 8 summarizes the composition of PairUG. We construct two complementary splits: Aligned Pairs and Retrieved Pairs. For the Aligned Pairs, the data originate from two sources: generation (3,043 examples) and understanding (1,928 examples). The class distribution is long-tailed: the head classes—such as Math (646), Human Gestures (602), and Puzzle (402)—account for a substantial portion of the data, while several categories (e.g., object count with 65 instances) appear infrequently. This split provides high-quality supervision across 17 labeled categories. The Retrieved Pairs display a different profile. The largest classes are JourneyDB (2,476) and GenEval (1,787), followed by Human Gestures (1,387), Object1 (1,232), Text2 (1,193), Text1 (992), Mscoco Human (857), and Occupation2 (708), with smaller categories such as Object2 (252) and Occupation1 (238). The similarity histogram is unimodal with most pairs in the 0.55–0.75 range, indicating that retrieval yields semantically related pairs while retaining diversity.

D Case Studies on Understanding Tasks

As shown in Figure 9, we present representative cases comparing Janus-Pro and PairUni on understanding tasks.

E Prompts for GPT-o3

We employ two instruction templates for GPT-o3: a generation prompt for producing structured quadruple data and an understanding prompt for obtaining human-readable interpretations and consistency checks. The complete prompt texts are shown in Figures 10 and 11. Both prompts specify the task, the expected input and output schema, and strict formatting constraints (e.g., forbidding extraneous commentary), which facilitate reliable downstream parsing and evaluation.

The Use of Large Language Models

We used a Large Language Model (LLM) only as a writing assistant to polish the language of the manuscript (e.g., grammar refinement, style adjustment, and clarity improvement). The research ideas, methodology design, experiments, and analysis were entirely conceived, implemented, and validated by the authors without reliance on the LLM. The LLM did not contribute to research ideation, experimental design, or result interpretation.

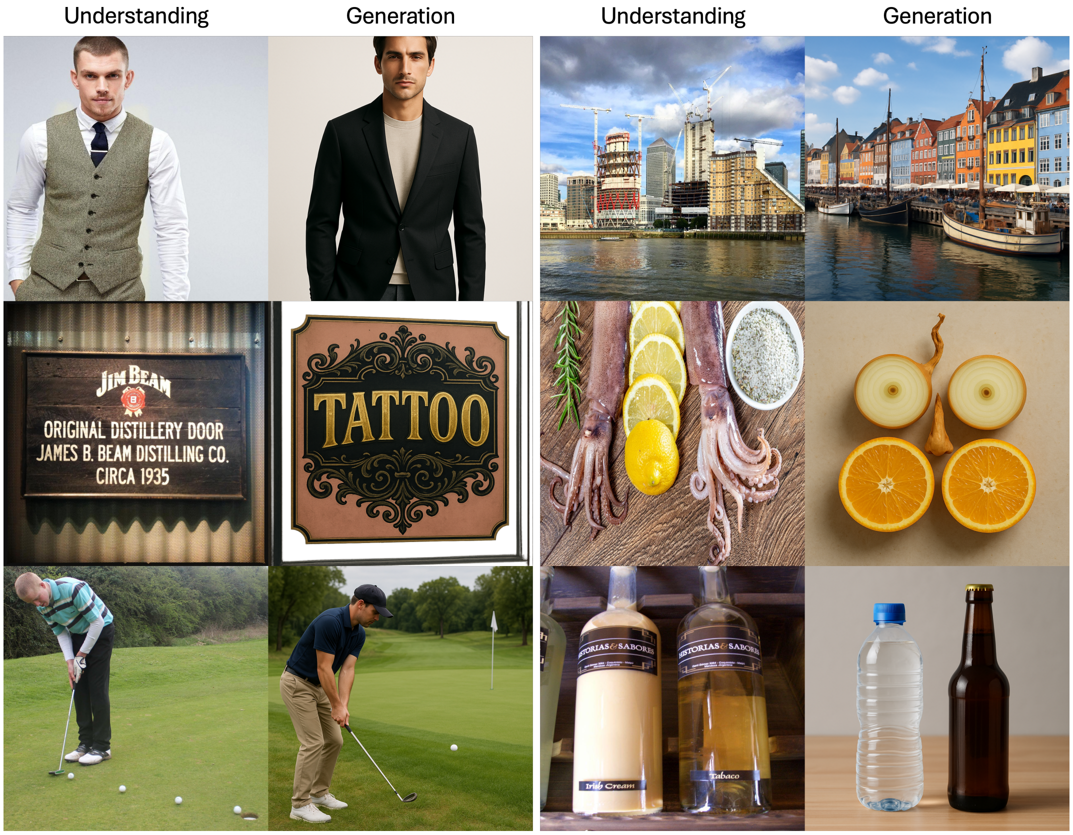


Figure 7 Representative paired cases for understanding and generation.

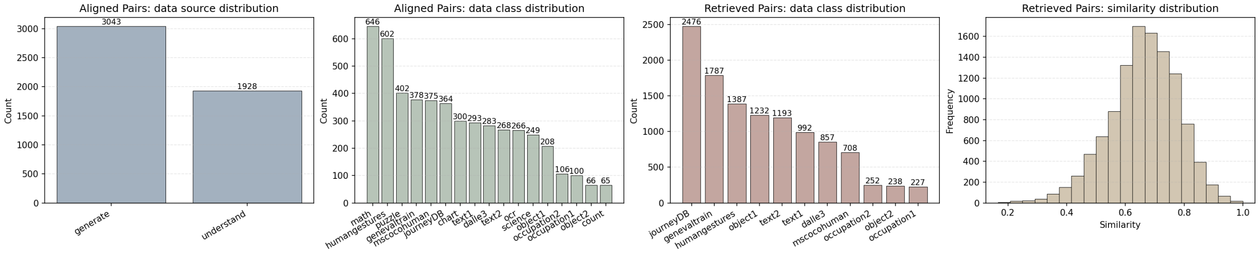
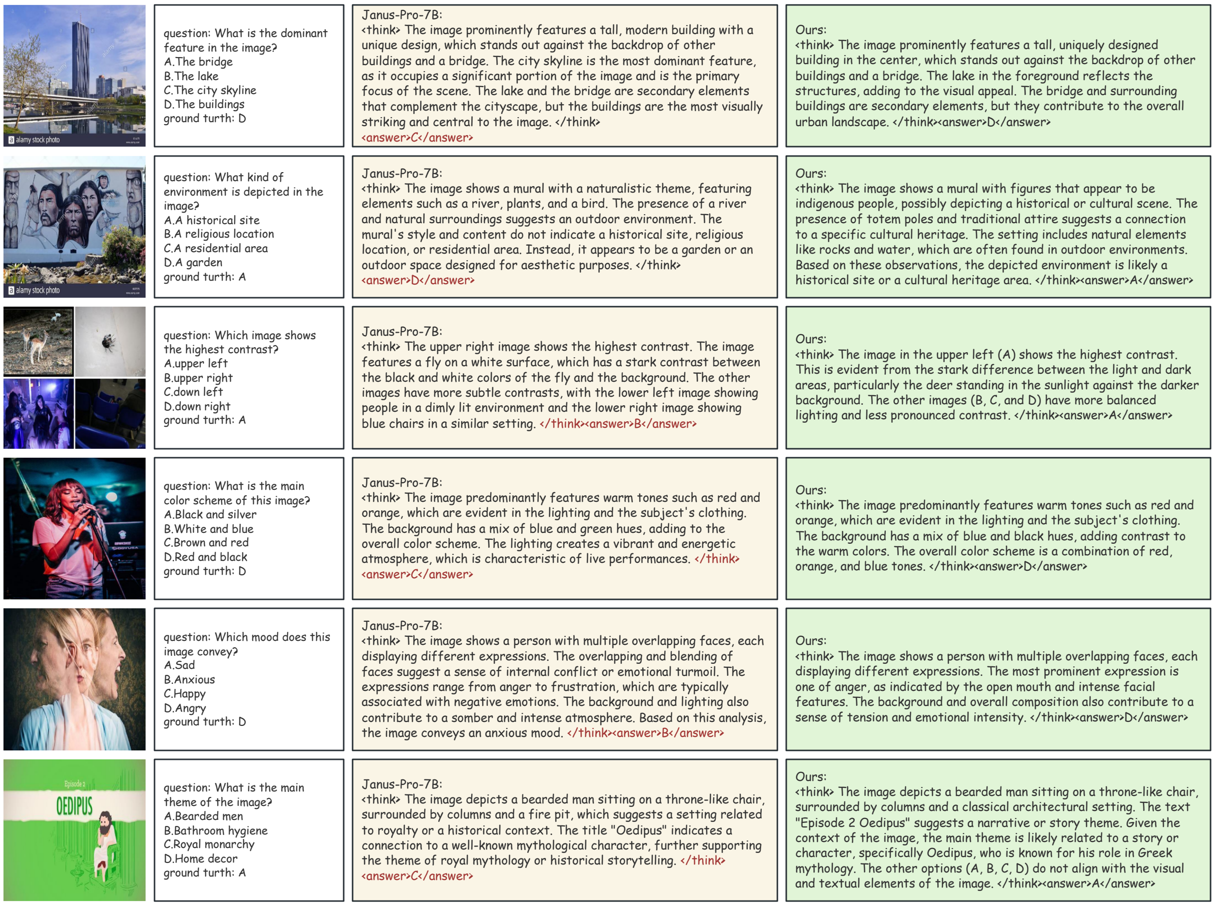
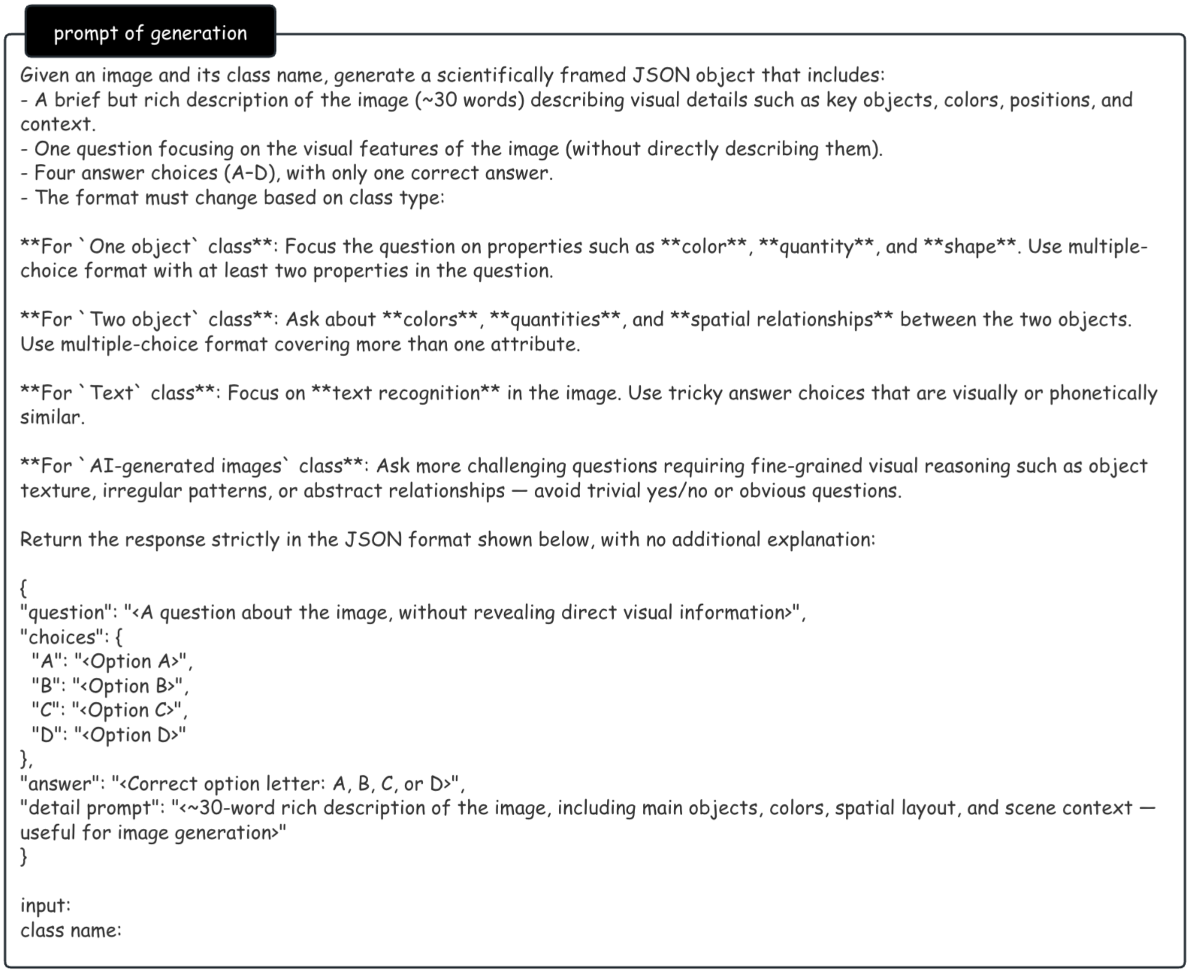


Figure 8 Distribution of PairUG. From left to right: (1) source breakdown for Aligned Pairs; (2) class distribution for Aligned Pairs; (3) class distribution for Retrieved Pairs; (4) similarity distribution for Retrieved Pairs.



question: What is the dominant feature in the image? A.The bridge B.The lake C.The city skyline D.The buildings ground turth: D	Janus-Pro-7B: <think> The image prominently features a tall, modern building with a unique design, which stands out against the backdrop of other buildings and a bridge. The city skyline is the most dominant feature, as it occupies a significant portion of the image and is the primary focus of the scene. The lake and the bridge are secondary elements that complement the cityscape, but the buildings are the most visually striking and central to the image. </think> <answer>C</answer>	Ours: <think> The image prominently features a tall, uniquely designed building in the center, which stands out against the backdrop of other buildings and a bridge. The lake in the foreground reflects the structures, adding to the visual appeal. The bridge and surrounding buildings are secondary elements, but they contribute to the overall urban landscape. </think><answer>D</answer>
question: What kind of environment is depicted in the image? A.A historical site B.A religious location C.A residential area D.A garden ground turth: A	Janus-Pro-7B: <think> The image shows a mural with a naturalistic theme, featuring elements such as a river, plants, and a bird. The presence of a river and natural surroundings suggests an outdoor environment. The mural's style and content do not indicate a historical site, religious location, or residential area. Instead, it appears to be a garden or an outdoor space designed for aesthetic purposes. </think> <answer>D</answer>	Ours: <think> The image shows a mural with figures that appear to be indigenous people, possibly depicting a historical or cultural scene. The presence of totem poles and traditional attire suggests a connection to a specific cultural heritage. The setting includes natural elements like rocks and water, which are often found in outdoor environments. Based on these observations, the depicted environment is likely a historical site or a cultural heritage area. </think><answer>A</answer>
question: Which image shows the highest contrast? A.upper left B.upper right C.down left D.down right ground turth: A	Janus-Pro-7B: <think> The upper right image shows the highest contrast. The image features a fly on a white surface, which has a stark contrast between the black and white colors of the fly and the background. The other images have more subtle contrasts, with the lower left image showing people in a dimly lit environment and the lower right image showing blue chairs in a similar setting. </think><answer>B</answer>	Ours: <think> The image in the upper left (A) shows the highest contrast. This is evident from the stark difference between the light and dark areas, particularly the deer standing in the sunlight against the darker background. The other images (B, C, and D) have more balanced lighting and less pronounced contrast. </think><answer>A</answer>
question: What is the main color scheme of this image? A.Black and silver B.White and blue C.Brown and red D.Red and black ground turth: D	Janus-Pro-7B: <think> The image predominantly features warm tones such as red and orange, which are evident in the lighting and the subject's clothing. The background has a mix of blue and green hues, adding to the overall color scheme. The lighting creates a vibrant and energetic atmosphere, which is characteristic of live performances. </think> <answer>C</answer>	Ours: <think> The image predominantly features warm tones such as red and orange, which are evident in the lighting and the subject's clothing. The background has a mix of blue and black hues, adding contrast to the warm colors. The overall color scheme is a combination of red, orange, and blue tones. </think><answer>D</answer>
question: Which mood does this image convey? A.Sad B.Anxious C.Happy D.Angry ground turth: D	Janus-Pro-7B: <think> The image shows a person with multiple overlapping faces, each displaying different expressions. The overlapping and blending of faces suggest a sense of internal conflict or emotional turmoil. The expressions range from anger to frustration, which are typically associated with negative emotions. The background and lighting also contribute to a somber and intense atmosphere. Based on this analysis, the image conveys an anxious mood. </think><answer>B</answer>	Ours: <think> The image shows a person with multiple overlapping faces, each displaying different expressions. The most prominent expression is one of anger, as indicated by the open mouth and intense facial features. The background and overall composition also contribute to a sense of tension and emotional intensity. </think><answer>D</answer>
question: What is the main theme of the image? A.Bearded men B.Bathroom hygiene C.Royal monarchy D.Home decor ground turth: A	Janus-Pro-7B: <think> The image depicts a bearded man sitting on a throne-like chair, surrounded by columns and a fire pit, which suggests a setting related to royalty or a historical context. The title "Oedipus" indicates a connection to a well-known mythological character, further supporting the theme of royal mythology or historical storytelling. </think> <answer>C</answer>	Ours: <think> The image depicts a bearded man sitting on a throne-like chair, surrounded by columns and a classical architectural setting. The text "Episode 2 Oedipus" suggests a narrative or story theme. Given the context of the image, the main theme is likely related to a story or character, specifically Oedipus, who is known for his role in Greek mythology. The other options (A, B, C, D) do not align with the visual and textual elements of the image. </think><answer>A</answer>

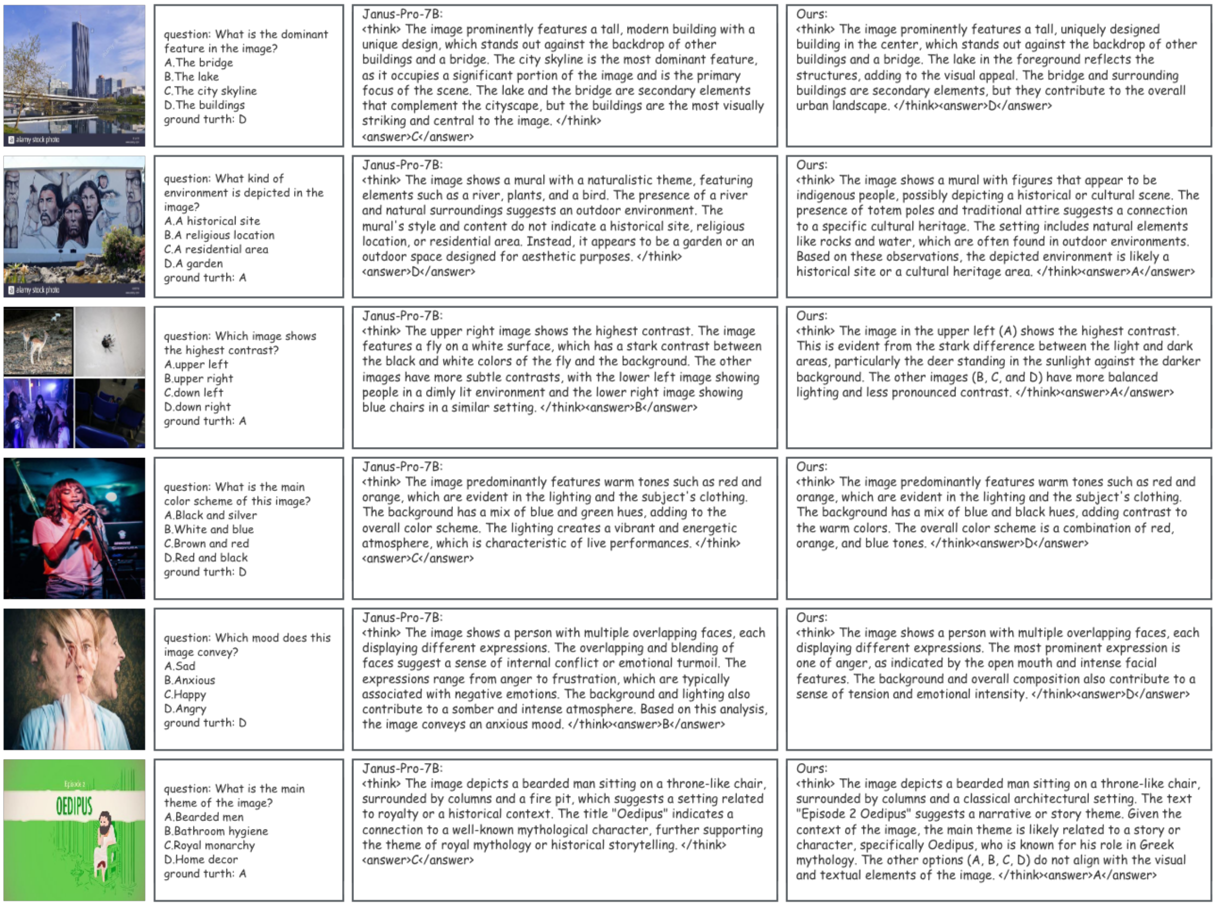
Figure 9 Representative cases comparing Janus-Pro and PairUni on understanding tasks.



prompt of generation

```
Given an image and its class name, generate a scientifically framed JSON object that includes:
- A brief but rich description of the image (~30 words) describing visual details such as key objects, colors, positions, and context.
- One question focusing on the visual features of the image (without directly describing them).
- Four answer choices (A–D), with only one correct answer.
- The format must change based on class type:

**For `One object` class**: Focus the question on properties such as **color**, **quantity**, and **shape**. Use multiple-choice format with at least two properties in the question.

**For `Two object` class**: Ask about **colors**, **quantities**, and **spatial relationships** between the two objects. Use multiple-choice format covering more than one attribute.

**For `Text` class**: Focus on **text recognition** in the image. Use tricky answer choices that are visually or phonetically similar.

**For `AI-generated images` class**: Ask more challenging questions requiring fine-grained visual reasoning such as object texture, irregular patterns, or abstract relationships — avoid trivial yes/no or obvious questions.

Return the response strictly in the JSON format shown below, with no additional explanation:

{
"question": "<A question about the image, without revealing direct visual information>",
"choices": {
  "A": "<Option A>",
  "B": "<Option B>",
  "C": "<Option C>",
  "D": "<Option D>"
},
"answer": "<Correct option letter: A, B, C, or D>",
"detail prompt": "<~30-word rich description of the image, including main objects, colors, spatial layout, and scene context — useful for image generation>"
}

input:
class name:
```

Figure 10 GPT-o3 prompt used for generating quadruple data.



question: What is the dominant feature in the image? A.The bridge B.The lake C.The city skyline D.The buildings ground turth: D	Janus-Pro-7B: <think> The image prominently features a tall, modern building with a unique design, which stands out against the backdrop of other buildings and a bridge. The city skyline is the most dominant feature, as it occupies a significant portion of the image and is the primary focus of the scene. The lake and the bridge are secondary elements that complement the cityscape, but the buildings are the most visually striking and central to the image. </think> <answer>C</answer>	Ours: <think> The image prominently features a tall, uniquely designed building in the center, which stands out against the backdrop of other buildings and a bridge. The lake in the foreground reflects the structures, adding to the visual appeal. The bridge and surrounding buildings are secondary elements, but they contribute to the overall urban landscape. </think><answer>D</answer>
question: What kind of environment is depicted in the image? A.A historical site B.A religious location C.A residential area D.A garden ground turth: A	Janus-Pro-7B: <think> The image shows a mural with a naturalistic theme, featuring elements such as a river, plants, and a bird. The presence of a river and natural surroundings suggests an outdoor environment. The mural's style and content do not indicate a historical site, religious location, or residential area. Instead, it appears to be a garden or an outdoor space designed for aesthetic purposes. </think> <answer>D</answer>	Ours: <think> The image shows a mural with figures that appear to be indigenous people, possibly depicting a historical or cultural scene. The presence of totem poles and traditional attire suggests a connection to a specific cultural heritage. The setting includes natural elements like rocks and water, which are often found in outdoor environments. Based on these observations, the depicted environment is likely a historical site or a cultural heritage area. </think><answer>A</answer>
question: Which image shows the highest contrast? A.upper left B.upper right C.down left D.down right ground turth: A	Janus-Pro-7B: <think> The upper right image shows the highest contrast. The image features a fly on a white surface, which has a stark contrast between the black and white colors of the fly and the background. The other images have more subtle contrasts, with the lower left image showing people in a dimly lit environment and the lower right image showing blue chairs in a similar setting. </think><answer>B</answer>	Ours: <think> The image in the upper left (A) shows the highest contrast. This is evident from the stark difference between the light and dark areas, particularly the deer standing in the sunlight against the darker background. The other images (B, C, and D) have more balanced lighting and less pronounced contrast. </think><answer>A</answer>
question: What is the main color scheme of this image? A.Black and silver B.White and blue C.Brown and red D.Red and black ground turth: D	Janus-Pro-7B: <think> The image predominantly features warm tones such as red and orange, which are evident in the lighting and the subject's clothing. The background has a mix of blue and green hues, adding to the overall color scheme. The lighting creates a vibrant and energetic atmosphere, which is characteristic of live performances. </think> <answer>C</answer>	Ours: <think> The image predominantly features warm tones such as red and orange, which are evident in the lighting and the subject's clothing. The background has a mix of blue and black hues, adding contrast to the warm colors. The overall color scheme is a combination of red, orange, and blue tones. </think><answer>D</answer>
question: Which mood does this image convey? A.Sad B.Anxious C.Happy D.Angry ground turth: D	Janus-Pro-7B: <think> The image shows a person with multiple overlapping faces, each displaying different expressions. The overlapping and blending of faces suggest a sense of internal conflict or emotional turmoil. The expressions range from anger to frustration, which are typically associated with negative emotions. The background and lighting also contribute to a somber and intense atmosphere. Based on this analysis, the image conveys an anxious mood. </think><answer>B</answer>	Ours: <think> The image shows a person with multiple overlapping faces, each displaying different expressions. The most prominent expression is one of anger, as indicated by the open mouth and intense facial features. The background and overall composition also contribute to a sense of tension and emotional intensity. </think><answer>D</answer>
question: What is the main theme of the image? A.Bearded men B.Bathroom hygiene C.Royal monarchy D.Home decor ground turth: A	Janus-Pro-7B: <think> The image depicts a bearded man sitting on a throne-like chair, surrounded by columns and a fire pit, which suggests a setting related to royalty or a historical context. The title "Oedipus" indicates a connection to a well-known mythological character, further supporting the theme of royal mythology or historical storytelling. </think> <answer>C</answer>	Ours: <think> The image depicts a bearded man sitting on a throne-like chair, surrounded by columns and a classical architectural setting. The text "Episode 2 Oedipus" suggests a narrative or story theme. Given the context of the image, the main theme is likely related to a story or character, specifically Oedipus, who is known for his role in Greek mythology. The other options (A, B, C, D) do not align with the visual and textual elements of the image. </think><answer>A</answer>

Figure 11 GPT-o3 prompt used for understanding quadruple data.