FLUX LEARNS TO WRITE

Anonymous authors
Paper under double-blind review

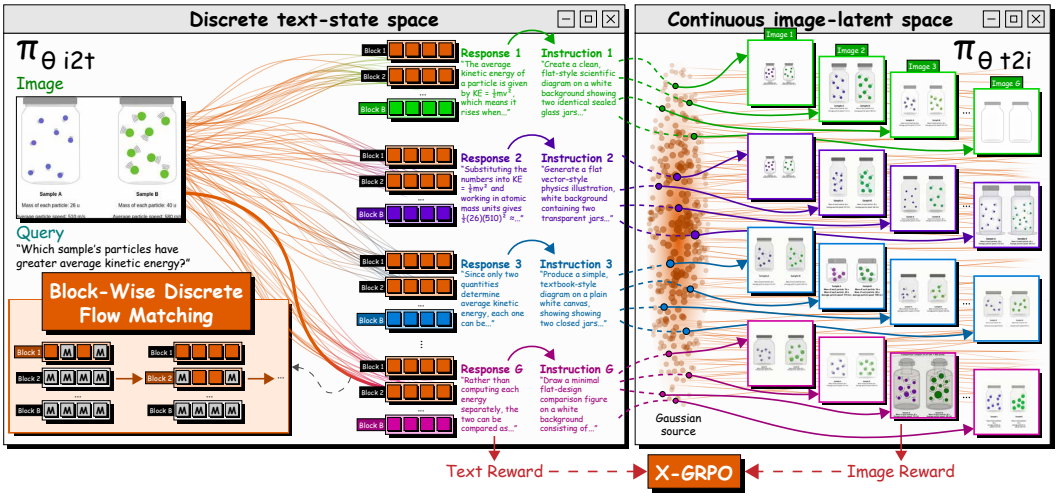


Figure 1: **We propose UniFlux** which transforms a pretrained image generator into a unified model with multimodal understanding and coherent reasoning. To further strengthen cross-modal grounding, multiple responses are converted into image-generation instructions, whose resulting images provide visual feedback to improve the underlying reasoning and semantic alignment.

ABSTRACT

Multimodal Diffusion Transformers (MM-DiTs) jointly attend to image and text tokens, introducing image-to-text feedback and implicit visually grounded semantics, making them promising for unified understanding and generation. However, existing approaches largely expose this capability only through pixel-level grounding and editing, while their reliance on full-sequence text denoising limits coherent long-range reasoning over jointly grounded visual and textual information. Moreover, weak cross-modal coupling limits visual feedback to textual reasoning, leading to semantic misalignment. We propose **UniFlux**, which adapts a pretrained MM-DiT for multimodal understanding using block-wise discrete flow matching, for long and coherent multimodal reasoning while preserving its generation prior. We further introduce X-GRPO, a cross-modal reinforcement-learning objective that couples separately normalized text- and image-side advantages through positive-gated multiplicative credit, assigning positive rollout credit only when both reasoning quality and image-regeneration fidelity exceed their respective group baselines. **UniFlux** achieves state-of-the-art performance across eleven understanding and generation benchmarks.

1 INTRODUCTION

Multimodal Diffusion Transformers (MM-DiTs) (Esser et al., 2024; Black Forest Labs, 2026b) have emerged as a prominent architecture for high-fidelity image generation, extending conventional Diffusion Transformers (DiTs) through joint modeling of text and image representations. In particular, several text-conditioned DiTs (Chen et al., 2024a; Li et al., 2024) employ a one-way conditioning paradigm, where fixed text representations guide evolving image latents but receive no visual feedback. This asymmetry prevents the textual representation from adapting to the evolving visual state, weakening fine-grained alignment between words and the visual objects and attributes formed during generation (Shin et al., 2025).

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.

MM-DiTs fundamentally modify this interaction by jointly processing image and text tokens. Their unified attention allows textual states to guide visual synthesis while simultaneously receiving information from the evolving image latents, introducing an explicit image-to-text feedback pathway at each denoising step. This bidirectional exchange suggests that pretrained MM-DiTs already contain an implicit form of visual understanding, even though this information is typically exposed only through generated pixels, attention or saliency maps, and editing behavior (Helbling et al., 2025; Ma et al., 2025a). Combined with their native capacity for high-fidelity image synthesis, this makes MM-DiTs a particularly promising foundation for unified multimodal models, where the same backbone can both interpret visual content and generate images from its semantic representations.

Recent unified multimodal approaches generally follow three directions. (Sun et al., 2024; Chameleon Team, 2024) models both text and images autoregressively, which leads to high generation cost for long visual sequences. (Wu et al., 2025; Shen et al., 2025; Deng et al., 2025) combine autoregressive language modeling with diffusion-based image generation, but still optimize the two modalities with different objectives. (Swerdlow et al., 2025; Yang et al., 2025) uses diffusion or flow modeling across both modalities, avoiding cross-modal objective conflict.

Despite these advances, existing unified models still exhibit several key limitations. **(1)** The visual-semantic information learned inside MM-DiTs remains largely implicit and is rarely verbalized directly (Li et al., 2025; Shi et al., 2026), limiting their use as unified understanding and generation models. **(2)** Existing diffusion/flow-based unified methods rely on full-sequence denoising, which weakens causal language modeling, constrains output length, and offers no explicit mechanism for incrementally building a stable reasoning trajectory (Arriola et al., 2025). This is particularly limiting for multimodal reasoning, where the model must progressively accumulate visual evidence and maintain coherent intermediate decisions. **(3)** Stronger language-visual generation alone does not ensure semantic alignment, causing reasoning and generation to diverge. Existing methods either optimize image generation with visual rewards (Xue et al., 2025; Liu et al., 2025; He et al., 2026), treat understanding and generation as separate tasks (Yang et al., 2025), or use staged reconstruction objectives (Bahng et al., 2025). Consequently, visual fidelity is rarely credited back to the textual response, allowing answers to be linguistically plausible yet visually incomplete, or visually faithful yet unresponsive to the prompt.

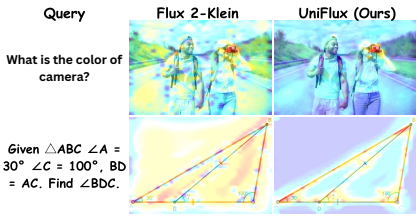


Figure 2: **Query-conditioned** attention maps for visual grounding.

We propose **UniFlux**, a unified framework that extends the text stream of a pretrained MM-DiT for coherent, long-form multimodal reasoning while preserving high-fidelity image generation. In Stage I, parameter-efficient adaptation with block-wise discrete flow matching (DFM) over textual answer representations transforms the text stream into an image-grounded reasoning pathway. Tokens are refined in parallel within each block, conditioned on previously completed blocks, enabling causal progression, stable context reuse, and variable-length reasoning. In Stage II, **X-GRPO** jointly aligns understanding and generation under a shared multimodal policy. By coupling language-side and image-side rewards through a shared cross-modal advantage, X-GRPO assigns downstream visual feedback to the textual decisions that produced it, encouraging cycle-consistent semantic alignment and jointly optimising both capabilities.

As shown in Figure 2, pretrained MM-DiTs (Flux 2-Klein) (Black Forest Labs, 2026b) attend diffusely, while **UniFlux** focuses on query-relevant evidence, enabling more precise and spatially grounded reasoning. We summarize our main contributions as follows:

1. We propose **UniFlux**, a two-stage framework that repurposes the text stream of a pretrained MM-DiT for unified multimodal understanding and generation. It turns the model's implicit visual-semantic representations into an explicit language reasoning pathway while preserving the native image-generation process, enabling both capabilities within a shared multimodal backbone.

2. We develop **block-wise discrete flow matching** for language generation within MM-DiTs, refining tokens in parallel within each block while conditioning on completed blocks to combine diffusion-style generation with causal progression, reusable context, and variable-length reasoning.

3. We introduce **X-GRPO**, a cross-modal reinforcement-learning objective that jointly aligns understanding and generation under a shared policy. By coupling language-side and image-side rewards

through a shared cross-modal advantage, X-GRPO feeds visual feedback back into the textual decisions that produced it, promoting stronger semantic consistency across both modalities.

4. **UniFlux** achieves state-of-the-art performance across eleven understanding and generation benchmarks. On multimodal understanding, it improves over the recent methods such as Mammoth-Moda2 (Shen et al., 2025) by up to 16%, and yields gains of over 83% relative to InternVL-2 (Chen et al., 2024c). **UniFlux** achieves an 12.35% improvement over Flux 2-Klein (Black Forest Labs, 2026b) for the image generation task on GenEval (Ghosh et al., 2023).

2 Related Work

Unified Models for Generation and Understanding. Unified multimodal models broadly fall into three categories based on their generation objectives. Autoregressive models (Chameleon Team, 2024; Liu et al., 2024; Wang et al., 2024b; Chen et al., 2025) represent both modalities as discrete tokens and decode them causally, incurring a high cost for long visual sequences. Hybrid approaches combine autoregressive language modeling with diffusion or flow-based image generation (Xie et al., 2025a;b; Zhou et al., 2025; Deng et al., 2025; Shen et al., 2025; Ma et al., 2024; Zhao et al., 2024), improving visual efficiency but optimizing distinct objectives across modalities. A third line unifies understanding and generation through diffusion or flow modeling. D-DiT combines continuous image flow with discrete text diffusion (Li et al., 2025), while UniDisc, Muddit, MMaDa, and FUDOKI explore discrete diffusion or flow across modalities (Swerdlow et al., 2025; Shi et al., 2026; Yang et al., 2025; Wang et al., 2025b). However, these methods rely on full-sequence text denoising rather than causally structured long-form generation.

Reward Optimization for Generative Models. Preference alignment in language models has progressed from PPO-based RLHF (Ouyang et al., 2022; Schulman et al., 2017) to simpler objectives, such as DPO (Rafailov et al., 2023) and value-free methods like GRPO (Shao et al., 2024; Guo et al., 2025a). These ideas have been extended to visual generation through diffusion-policy optimization (Black et al., 2023; Fan et al., 2023), differentiable rewards (Xu et al., 2023), and preference learning (Wallace et al., 2024). More recently, GRPO-style methods have been adapted to autoregressive, diffusion, and flow-matching image generators (Xue et al., 2025; Liu et al., 2025; Wang et al., 2025c; Zheng et al., 2026; Duan et al., 2025; Yuan et al., 2025; Luo et al., 2025; Tong et al., 2025). T2I-R1 jointly optimizes textual planning and image generation through image-side rewards (JIANG et al., 2025). Most recently, unified methods couple understanding and generation through reconstruction or shared rewards (Yan et al., 2026; Liu et al., 2026; Zhang et al., 2026), but provide only coarse, largely terminal feedback with limited modality-specific credit assignment.

3 UniFlux

3.1 Background

Implicit Multimodal Understanding in MM-DiT. Conventional DiT models use text (q) to guide image denoising but receive no feedback from the evolving image latents (z_t). MM-DiT instead maintains image and text as interacting internal states, initialized as $\mathbf{H}_I^0 = P_I(z_t)$ and $\mathbf{H}_T^0 = P_T(\mathcal{E}(q))$. Here, $P_I(\cdot)$ and $P_T(\cdot)$ denote the image and text token projections, respectively, while $\mathcal{E}(q)$ denotes the embedding of the text tokens. Each MM-DiT block jointly processes both streams, allowing image tokens to attend to text tokens and text tokens to attend back to image tokens:

$$\mathbf{H}_I^{\ell+1} = \mathrm{Attn}_I\big(\mathbf{H}_I^\ell, [\mathbf{H}_I^\ell; \mathbf{H}_T^\ell]\big), \qquad \mathbf{H}_T^{\ell+1} = \mathrm{Attn}_T\big(\mathbf{H}_T^\ell, [\mathbf{H}_I^\ell; \mathbf{H}_T^\ell]\big), \qquad \hat{\mathbf{v}}_I = g_I(\mathbf{H}_I^L), \quad \ell \in [1, L]$$

where g_I is the MM-DiT's native continuous flow-matching head for image generation and decodes only the image stream after L layers, leaving $\mathbf{H}_T^{\ell+1}$ without an explicit output head. However, as the text stream is repeatedly conditioned on the evolving visual representation, it naturally accumulates visually grounded semantic information (Shin et al., 2025).

Block-Wise Language Generation. Block diffusion interpolates between autoregressive and diffusion language modeling by generating blocks sequentially while refining each block's tokens in parallel (Arriola et al., 2025). For a response $\mathbf{y} = [\mathbf{y}^1, \ldots, \mathbf{y}^B]$, its distribution factorizes as $p_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{b=1}^{B} p_\theta\big(\mathbf{y}^b \mid \mathbf{y}^{<b}, \mathbf{x}\big)$, across B blocks. Here, $\mathbf{x}$ denotes the multimodal context, and each block conditional is modeled through discrete denoising. Prior blocks remain fixed as causal

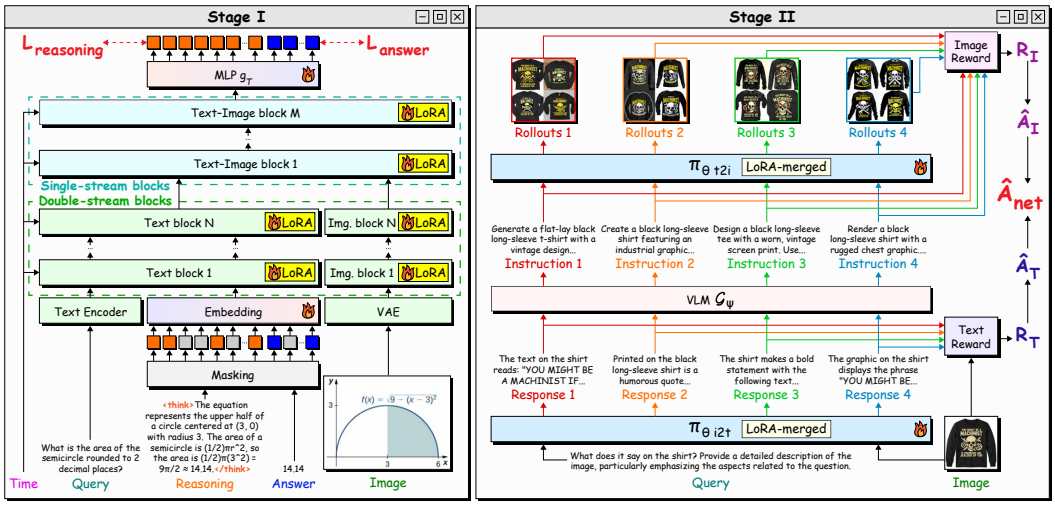


Figure 3: **Overview of UniFlux.** Stage I adapts the pretrained MM-DiT text stream with LoRA and block-wise discrete flow matching, enabling causal reasoning while preserving image generation. Stage II introduces X-GRPO, which samples text and image rollouts and couples their normalized rewards through a shared cross-modal advantage to jointly align understanding and generation.

context, while tokens in the current block are refined bidirectionally. This combines stable left-to-right progression with parallel within-block refinement, supporting coherent long-form reasoning and variable-length generation.

3.2 Architecture of UniFlux

UniFlux transforms a pretrained MM-DiT into a unified multimodal model for both understanding and generation while preserving its strong image-synthesis capabilities. As illustrated in Figure. 3, **UniFlux** follows a two-stage training framework. Stage I adapts the pretrained MM-DiT for image-grounded understanding by directly supervising the extended text stream using a Block-Wise Discrete Flow Matching (DFM) objective. Stage II jointly aligns understanding and generation through a cross-modal reinforcement learning objective that couples rewards across both modalities.

Stage I: Multimodal Understanding Alignment

A pretrained MM-DiT already learns image-grounded textual representations through image-to-text joint attention. However, these representations serve only as internal conditioning states up to several blocks and are not parameterized to predict language tokens. This provides a natural pathway for multimodal understanding without introducing a separate autoregressive decoder.

Extending the text stream for multimodal understanding. We extend and propagate the text stream from the N dual stream blocks into M single text-image stream blocks through the full MM-DiT backbone and introduce a vocabulary MLP projection head g_T at its output, mapping the final text-stream representations to vocabulary logits over the model's embedding ($\mathcal{E}$), enabling multimodal understanding. To preserve the pretrained multimodal and image-generation priors, we adapt the all MM-DiT layers using low-rank adapters (LoRA) (Hu et al., 2022), while jointly optimizing $\mathcal{E}$ and g_T. Formally, given an image-query-answer triplet $(\mathbf{I}, \mathbf{q}, \mathbf{y}) \sim \mathcal{D}$, the native image stream of MM-DiT produces visual tokens $\mathbf{x_I}$, while the frozen text encoder maps the query to $\mathbf{x_q}$. The target response $\mathbf{y} = [\mathbf{r}; \mathbf{a}]$ contains a coherent reasoning trace $\mathbf{r}$ followed by the final answer $\mathbf{a}$.

Training using block-wise DFM for coherent reasoning. As long-form reasoning must progressively accumulate visual evidence and remain consistent with earlier decisions, jointly denoising the entire response provides an unsuitable fixed-length, non-causal generation process. Hence, we partition the response $\mathbf{y}$ into B blocks, $\mathbf{y} = [\mathbf{y}^1, \ldots, \mathbf{y}^B]$, and model it autoregressively across blocks while applying DFM within each block. This preserves stable causal context across all blocks and parallel refinement within each block. During training, the block b is corrupted at $\tau \sim \mathcal{U}(0, 1)$, while preceding blocks $\mathbf{y}^{<b}$ remain clean. Each token in block $\mathbf{y}^b$ is retained with probability $1 - \tau$ and otherwise replaced by `[MASK]`, producing the corrupted block $\tilde{\mathbf{y}}^b_\tau$. The text input is

$\mathbf{x}_T = [\mathbf{x}_\mathbf{q}; E(\mathbf{y}^{<b}); E(\tilde{\mathbf{y}}_\tau^b)]$. We inject $t = \tau$ through the native MM-DiT conditioning module, allowing the noised block to attend to the visual tokens, query, and prior reasoning blocks. Finally, the projection head $g_T(\mathbf{H}_{T,b}^L)$ decodes the current block's final hidden states.

Stage I training objective. Let θ represent the trainable parameter of LoRA, g_T and E. Using the clean-token parameterization of discrete flow matching, the model learns the reverse categorical flow by predicting the target tokens from their intermediate corrupted state:

$$\mathcal{L}_{\text{Stage-I}} = \mathbb{E}_{\tau,(\mathbf{I},\mathbf{q},\mathbf{y}),\tilde{\mathbf{y}}_\tau} \left[\sum_{b=1}^{B} \sum_{i:\, \tilde{y}_{\tau,i}^b = \texttt{[MASK]}} -\log p_\theta \left(y_i^b \mid \tilde{\mathbf{y}}_\tau^b, \mathbf{y}^{<b}, \mathbf{q}, \mathbf{I}, \tau \right) \right]. \tag{1}$$

Here, The inner sum $\sum i : \tilde{y}_{\tau,i}^b = \texttt{[MASK]}$ is taken only over masked positions in block b, and y_i^b denotes the corresponding clean target token at position i. Since $\mathbf{y} = [\mathbf{r}; \mathbf{a}]$, the objective supervises masked tokens across both the reasoning trace and final answer. During inference, each block starts masked and is refined from $\tau = 1$ to 0 with prior blocks as causal context. This enables coherent variable-length reasoning while preserving the image-generation pathway. The block-causal attention mask, discrete transition rates, and consistency analysis are detailed in Appendix D.

STAGE II: CROSS-MODAL REINFORCEMENT LEARNING

Stage I supervises reference reasoning–answer sequences through token reconstruction. However, this objective does not directly evaluate the correctness and visual grounding of the model's own sampled responses. In Stage II, we introduce **X-GRPO**, a cross-modal reinforcement learning method that jointly aligns multimodal understanding and image generation through a shared policy. Rather than optimizing text and image rewards independently, X-GRPO couples both modalities through a shared cross-modal advantage, thereby encouraging cycle-consistent semantic alignment across the entire multimodal generation process.

Generation as complementary grounding feedback. The main intuition of **X-GRPO** is a consistency principle where if a reasoning trace truly captures the visual content of the input image, then an instruction derived from that trace should let the model *regenerate* an image faithful to the original. Faithful regeneration is thus evidence that the reasoning is visually grounded, which is hard to satisfy with fluent but ungrounded language. Concretely, from each sampled response, we form an image-generation instruction with a VLM instruction generator $\mathcal{G}_\psi$, regenerate candidate images through the model's own generation pathway, and judge the fidelity of the generated image.

Formally, we initialize the policy π_θ from the Stage-I and keep a frozen reference π_{ref} and a sampling policy $\pi_{\theta_{\text{old}}}$. For a query-image pair $(\mathbf{q}, \mathbf{I}) \sim \mathcal{D}$, the understanding pathway samples a group of G_T responses $\{A^{(i)}\}_{i=1}^{G_T}$, each a reasoning trace and a reference answer $a^\star$. A VLM judge scores them, $R_T^{(i)} = \mathcal{J}_\psi^T(\mathbf{q}, \mathbf{I}, A^{(i)}, a^\star)$. Each response is turned into an instruction $X^{(i)} = \mathcal{G}_\psi(\mathbf{q}, \mathbf{I}, A^{(i)})$, from which the generation pathway samples $G_\mathbf{I}$ images $\{\widehat{\mathbf{I}}^{(i,j)}\}_{j=1}^{G_\mathbf{I}}$ that a second judge rates for faithfulness to the original image, $R_\mathbf{I}^{(i,j)} = \mathcal{J}_\psi^\mathbf{I}(\mathbf{I}, \widehat{\mathbf{I}}^{(i,j)}, X^{(i)})$. The complete language- and image-side reward formulations and VLM judge prompts are provided in Appendix E. We apply GRPO normalization by subtracting the group mean and dividing by the population standard deviation plus $\delta = 10^{-8}$. For each input $(\mathbf{q}, \mathbf{I})$, text statistics are computed over its G_T responses, and image statistics over all $G_T G_\mathbf{I}$ generated images.

We use positive-gated multiplicative coupling to assign positive credit only when both modality-specific rewards exceed their respective group means:

$$\widehat{A}_{\text{net}}^{(i,j)} = \text{sg}\Big([\widehat{A}_T^{(i)}]_+ \cdot [\widehat{A}_\mathbf{I}^{(i,j)}]_+ \Big), \qquad [x]_+ = \max(x, 0). \tag{2}$$

This multiplicative weight is positive only when both advantages are positive and vanishes whenever either is nonpositive. Unlike the signed product, this weight is zero whenever either advantage is nonpositive, preventing positive reinforcement of jointly below-average pairs. Appendix F.2 establishes the sign consistency of this coupling, including its response-level averaged form.

Stage II (X-GRPO) training objective. We optimize both pathways using the PPO clipped surrogate $\ell(\rho, \widehat{A}) = \min(\rho \widehat{A}, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)\widehat{A})$ following Xue et al. (2025). For the understanding pathway, we use the text-policy surrogate ratio $\rho_T^{(i)} = \pi_{\theta i2t}(\xi_T^{(i)} \mid \mathbf{I}, Q) / \pi_{\theta_{\text{old}} i2t}(\xi_T^{(i)} \mid \mathbf{I}, Q)$, where $\xi_T^{(i)}$ denotes the blockwise recorded text rollout. For the generation pathway, we use the likelihood



Figure 4: **Qualitative comparison on fine-grained image understanding**, where UniFlux captures more specific visual details than BAGEL and MammothModa2.

ratio $\rho_{\mathbf{I}}^{(i,j)} = \pi_{\theta t2i}(\xi_{\mathbf{I}}^{(i,j)} \mid c_{\mathbf{I}}^{(i)}) / \pi_{\theta_{\text{old}} t2i}(\xi_{\mathbf{I}}^{(i,j)} \mid c_{\mathbf{I}}^{(i)})$, where $\xi_{\mathbf{I}}^{(i,j)}$ is the j-th image rollout conditioned on the instruction $c_{\mathbf{I}}^{(i)}$ derived from response i. The image likelihood is computed from the recorded stochastic transition densities. Both policies are evaluated on identical recorded states, targets, and conditioning inputs, with text scores and image log-likelihoods computed without length normalization. Text rollouts are generated using τ-leaping for the block-wise discrete flow (Gat et al., 2024). With rollouts sampled under $\pi_{\theta_{\text{old}}}$, the X-GRPO objective:

$$\mathcal{J}_{\text{X-GRPO}}(\theta) = \mathbb{E}_{(\mathbf{q},\mathbf{I})\sim\mathcal{D}}\left[\frac{1}{G_T G_{\mathbf{I}}}\sum_{i=1}^{G_T}\sum_{j=1}^{G_{\mathbf{I}}}\left(\ell\big(\rho_T^{(i)}, \widehat{A}_{\text{net}}^{(i,j)}\big) + \gamma\,\ell\big(\rho_{\mathbf{I}}^{(i,j)}, \widehat{A}_{\text{net}}^{(i,j)}\big)\right)\right] \tag{3}$$
$$- \beta_T D_{\text{KL}}\big(\pi_{\theta i2t} \,\|\, \pi_{\text{ref}i2t}\big) - \beta_{\mathbf{I}} D_{\text{KL}}\big(\pi_{\theta t2i} \,\|\, \pi_{\text{ref}t2i}\big).$$

The Stage-II training loss is $\mathcal{L}_{\text{Stage-II}} = -\mathcal{J}_{\text{X-GRPO}}(\theta)$. Here, γ controls the relative contribution of the image-generation objective to the text-understanding objective, while β_T and $\beta_{\mathbf{I}}$ weight the KL regularization that constrains the text and image policies, respectively, toward their frozen reference policies. X-GRPO training details and the sign-consistency analysis of the shared cross-modal advantage are provided in Appendix F. We further analyze representation evolution through layer-wise semantic probing in Appendix G, showing increased semantic separability after both training stages.

4 Main Results

We train **UniFlux** using public multimodal datasets and synthetic reasoning annotations, organized into three primary collections: ∼50M generation–understanding alignment examples, 15M visual instruction examples, and 20M visual reasoning examples. The alignment collection comprises 10M description, 15M instruction-response, 20M reasoning-trace, and 5M text-to-image generation examples derived from shared visual sources. The instruction and reasoning collections cover captioning, VQA, OCR, document understanding, and visual reasoning. Implementation and dataset details are provided in Appendices A and B, respectively.

Image Understanding and Reasoning. Table 1 evaluates **UniFlux** across six benchmarks spanning captioning, visual understanding, and multimodal reasoning. We instantiate UniFlux on two variants of FLUX.2 rectified-flow backbones. FLUX.2-dev (32B) (Black Forest Labs, 2026a) is the higher-capacity backbone, with twice as many single-stream Transformer blocks and a larger hidden dimension than the more compact FLUX.2-Klein (9B) (Black Forest Labs, 2026b). We focus on the 9B variant, which is comparable in scale to existing unified models. Among unified approaches, **UniFlux** (Flux 2-Klein) achieves the strongest overall performance. On visual-understanding benchmarks, it yields average relative improvements of 11.34% over BAGEL and 6.02% over MammothModa2, with the largest gain on GQA +13.92% over MammothModa2. Compared with diffusion-native unified models, the improvement is more pronounced, overall 30% over Muddit and 90% over D-DiT across the benchmarks. The gains extend to multimodal reasoning, where block-wise DFM provides a causal reasoning trajectory in which later blocks build on previously stabilized visual evidence. **UniFlux** improves MMMU (Yue et al., 2024) by 16.47% over MammothModa2 while remaining competitive on MathVista (Lu et al., 2024). While retaining native image-generation capability, it also surpasses understanding-only models on reasoning, improving MMMU by 12.82% over InternVL-2.0 and 19.69% over Qwen-VL, and MathVista by 105.07% and 166.59%, respectively. Scaling to Flux 2-Dev further improves reasoning to 74.05 on MMMU and 84.56 on Math-

Table 1: **Comparison on visual understanding and reasoning benchmarks.** AR: Autoregressive; Diff.: Diffusion; DD: Discrete Diffusion; DFM: Discrete Flow Matching; FM: Flow Matching. All metrics are higher-is-better (↑). Best scores are shown in **bold**.

Model	Params (B)	Generation		Captioning	Visual Understanding			Multimodal Reasoning	
		Text	Image	COCO	VQAv2	MME	GQA	MMMU	MathVista
Understanding-only models									
InternVL-2.0 (Chen et al., 2024c)	8	AR	–	–	–	**1648.10**	–	61	39.22
Qwen-VL (Bai et al., 2023)	7	AR	–	–	–	1487.50	78.20	57.50	30.17
Unified understanding and generation models									
UniDisc (Swerdlow et al., 2025)	1.4	DD	DD	46.81	52.34	1000.76	53.11	49.08	–
D-DiT (512) (Li et al., 2025)	2	DD	Diff.	56.20	60.10	1124.70	59.20	–	17.44
Muddit (1024) (Shi et al., 2026)	4	DD	DD	60.22	65.66	941.0	53.48	51.11	69.28
Show-O-2 (512) (Xie et al., 2025b)	1.3	AR	DD	–	69.40	1097.20	58.00	–	25.57
BAGEL (Deng et al., 2025)	7+7	AR	FM	67.32	74.32	1312.00	61.74	58.22	73.11
MammothModa2 (Shen et al., 2025)	8	AR	FM	69.23	75.12	1444.0	64.23	59.09	81.18
FUDOKI (Wang et al., 2025b)	1.7	DFM	DFM	61.78	70.17	1212	57.66	34.4	38.01
UniFlux (Flux 2-Klein)	9	DFM	FM	73.33	76.78	1472	73.17	68.82	80.43
UniFlux (Flux 2-Dev)	32	DFM	FM	**80.24**	**82.18**	1544.0	**79.09**	**74.05**	**84.56**

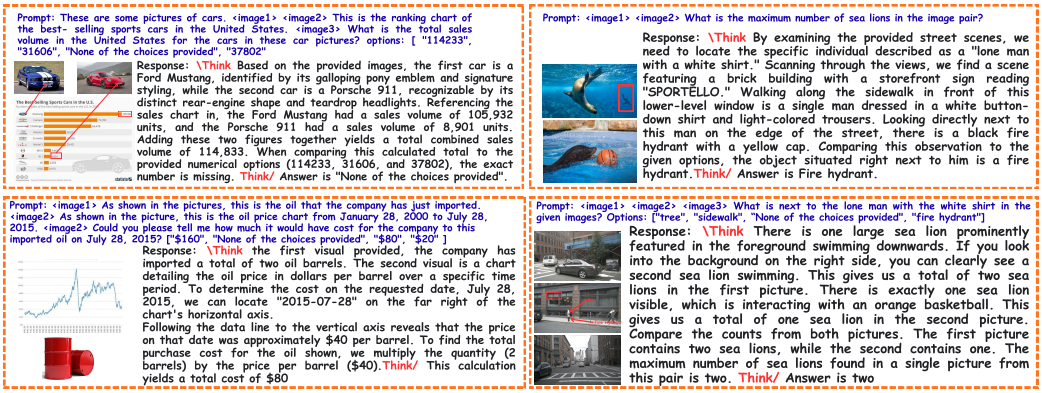


Figure 6: **Qualitative cross-image understanding. UniFlux** integrates complementary evidence across multiple images for recognition, chart reading, spatial grounding, counting, and reasoning.

Vista, corresponding to 25.32% and 4.16% gains over MammothModa2. Additional OCRBenchV2 results and qualitative analyses of OCR and complex visual reasoning are provided in Appendix I.

Figure. 4 compares **UniFlux** with BAGEL and MammothModa2 on fine-grained visual understanding. Beyond recognizing the overall scene, **UniFlux** consistently captures small, localized visual cues and task-relevant details that other unified models overlook. In the dashboard example, it correctly identifies the Dokan interface together with the sidebar menu, order-status pie chart, and sales graph, whereas BAGEL provides only a generic dashboard description, and MammothModa2 misinterprets it as a finance interface. Similarly, **UniFlux** captures localized attributes of the hotel room, such as the colorful quilt and wall lamp, whereas the baselines omit these details or misclassify the scene. It reads the "TRANS AM" text on the car and uses it to identify the vehicle, demonstrating stronger grounding of fine-grained visual evidence in the generated response. Further qualitative analyses of evidence extraction, answerability, and complex and long mathematical reasoning are provided in Appendix K.1.

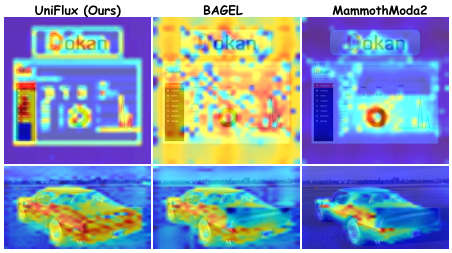


Figure 5: **Qualitative visual grounding.** Query-conditioned attention maps for **UniFlux**, BAGEL, and MammothModa2 on fine-grained image understanding.

To further support the analysis, Figure. 5 visualizes query-conditioned attention for the examples in Figure. 4, highlighting the visual evidence used by the model. **UniFlux** exhibits more concentrated query-conditioned activation on the relevant visual evidence, such as the dashboard statistics and pie chart in the first example and the vehicle body and identifying rear text in the second. In contrast, BAGEL and MammothModa2 show more diffuse or partially misaligned responses, suggesting weaker localization of the evidence required to answer the query.

Table 2: **Cross image understanding Accuracy.**

Model	MUIRBench (Wang et al., 2025a)	MMIU (Meng et al., 2025)
Muddit	29.27	30.33
FUDOKI	32.27	35.67
BAGEL	37.67	40.19
MammothModa2	39.82	42.29
UniFlux (Flux 2-Klein)	**50.19**	**55.01**

Table 3: **Text-to-image generation results.**

Method	GenEval	DPGBench
Flux 2-Klein (Black Forest Labs, 2026b)	0.81	83.31
DanceGRPO (Xue et al., 2025)	0.82	82.38
Flow-GRPO (Liu et al., 2025)	0.89	85.56
Muddit (Shi et al., 2026)	0.67	80.09
FUDOKI (Wang et al., 2025b)	0.83	83.21
BAGEL (Deng et al., 2025)	0.89	87.74
MammothModa2 (Shen et al., 2025)	0.88	87.53
UniFlux (Flux 2-Klein)	**0.91**	**90.08**

Cross Image Understanding. We evaluate **UniFlux** on MUIRBench and MMIU for multi-image reasoning, using Mantis-Instruct (Jiang et al., 2024) for training (Table 2). **UniFlux** significantly outperforms SOTA unified models, improving over MammothModa2 by 26.04% and 30.08%, and over Muddit by 71.47% and 81.37% on MUIRBench and MMIU, respectively. This improvement is driven by block-wise diffusion, which supports the coherent long-form reasoning needed to integrate evidence across multiple images.

Figure 6 shows that **UniFlux** can aggregate evidence distributed across multiple images. It jointly identifies objects, retrieves their corresponding chart values, combines quantities across images to support numerical reasoning, localizes fine spatial cues such as the fire hydrant next to the target person, and compares object counts across image pairs. Thus, **UniFlux** can preserve and integrate fine-grained visual evidence across views rather than reasoning from each image independently.

Generation. Table 3 evaluates text-to-image generation on GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024). GenEval measures compositional prompt following across object presence, counting, color, spatial relations, and attribute binding. DPGBench evaluates semantic fidelity under dense and complex textual descriptions. **UniFlux** achieves the best performance on both benchmarks, improving the pretrained Flux 2-Klein backbone by 12.35% on GenEval and 8.13% on DPGBench. It surpasses Flow-GRPO by 2.25% and 5.28%, respectively, showing that X-GRPO improves visual alignment beyond image-side reward optimization alone. **UniFlux** yields average gains of 2.46% over BAGEL and 3.16% over MammothModa2 across both benchmarks. Clearly, jointly aligning the language and visual pathways strengthens generation without sacrificing multimodal understanding. The full table for GenEval and DPGBench is in Appendix J. Additional qualitative comparisons of T2I with alignment baselines are provided in Appendix K.2.

Figure. 7 shows that **UniFlux** more faithfully preserves fine-grained prompt constraints, including subject-specific color bindings in the first example and the exact "ICLR" text in the second. In contrast, SOTA models exhibit attribute inconsistencies, incomplete compositions, or text-rendering errors such as "ICCLR" and "ICL". Additional generation results across diverse visual domains and aspect ratios are shown in Figure 18.



Figure 7: **Qualitative image-generation comparison.**

5 Ablation Results

Table 4 presents a detailed ablation of **UniFlux**. Here, $\mathcal{N}$ denotes group normalization, $\mathcal{J}_T$ and $\mathcal{J}_I$ denote the text-only and image-only GRPO objectives, and $\mathcal{R}_{\text{pres}}$ denotes the image-preservation penalty. Additional ablations on reasoning length, VLM judge choice, text-representation depth, and LoRA rank are provided in Appendix H.4.

Reward formulation, optimization pathways, and reasoning design. Removing Stage-II X-GRPO leaves the Stage-I model and reduces performance from 68.82/0.91 to 62.03/0.79. Understanding-only GRPO retains only the text-side objective $\mathcal{J}_T$ and achieves 66.54/0.85, while image-only GRPO uses only $\mathcal{J}_I$ and reaches 63.22/0.85. Both improve over the Stage-I model but

remain below the full joint configuration. We next compare three alternatives to the positive-gated coupling $[\widehat{A}T] + [\widehat{A}I]+$. Additive coupling sums the separately normalized advantages, $\widehat{A}_T + \lambda\widehat{A}_I$, and achieves 64.03/0.86. The normalized reward-sum variant combines the raw rewards before group normalization, $\mathcal{N}(R_T + \lambda R_I)$, and obtains 64.01/0.83; both variants use $\lambda = 0.88$. The signed product $\widehat{A}_T\widehat{A}_I$ reaches 63.12/0.82. This formulation assigns positive credit when both advantages are negative, reinforcing pairs whose rewards fall below their respective group means. The positive gate excludes these pairs and assigns positive credit only when both advantages are positive. For reasoning design, removing explicit intermediate reasoning reduces performance to 64.12/0.84. Replacing block-wise generation with full-sequence denoising, which jointly denoises the response without sequential conditioning on completed blocks, yields 65.37/0.84. These results support retaining both intermediate reasoning and causal context across response blocks.

Source image access. We assess the contribution of direct image conditioning to instruction generation by replacing $\mathcal{G}\psi(q, I, A^{(i)})$ with $\mathcal{G}\psi(q, A^{(i)})$. The instruction generator receives only the query and sampled response; the understanding pathway and image-reward evaluation retain access to the source image. This variant achieves 66.88 on MMMU and 0.88 on GenEval, compared with 68.82 and 0.91 for the full configuration. Instructions derived from the query and response retain much of the performance, while supplying the source image as additional context improves MMMU by 1.94 percentage points and GenEval by 0.03.

Table 4: **UniFlux Ablation.**

Configuration	MMMU ↑	GenEval ↑
Core Components		
w/o X-GRPO	62.03	0.79
Understanding-only: $\mathcal{J}_T$	66.54	0.85
Additive: $\widehat{A}_T + \lambda\widehat{A}_I$	64.03	0.86
Signed product: $\widehat{A}_T \cdot \widehat{A}_I$	63.12	0.82
Normalized: $\mathcal{N}(R_T + \lambda R_I)$	64.01	0.83
Image-only: $\mathcal{J}_I$	63.22	0.85
w/o Thinking	64.12	0.84
w/o Block Diffusion	65.37	0.84
$\mathcal{G}_\psi(q, A^{(i)})$	66.88	0.88
Generation Reward		
HPSv3 (Ma et al., 2025b)	66.88	0.86
HPSv3 + SigLIP (Zhai et al., 2023)	67.59	0.88
HPSv3 + SigLIP + VLM	68.82	0.91
UniFlux (Flux 2-Klein)	**68.82**	**0.91**

Generation Rewards. HPSv3 alone achieves 66.88 on MMMU and 0.86 GenEval, while adding SigLIP improves performance to 67.59/0.88 by strengthening image–text semantic alignment. Including the VLM reward increases performance to 68.82/0.91. Thus, semantic feedback complements perceptual quality and embedding similarity to achieve more faithful cross-modal optimization.

X-GRPO Group Size. Figure 8(a) shows the effect of the number of groups (G) used for group-relative optimization. Increasing G consistently improves performance on both MMMU and GenEval, as larger groups provide more diverse samples and more reliable relative advantages. **UniFlux** use G=8 by default as a practical trade-off between performance and rollout cost.

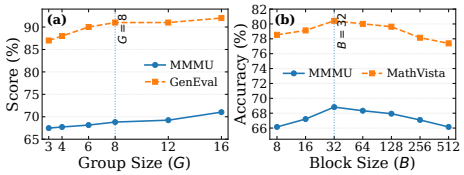


Figure 8: **Ablation on X-GRPO** group size and reasoning block size.

Reasoning Block Size. Figure 8(b) evaluates block sizes from 8 to 512 tokens. A block size of 32 achieves the highest accuracy, reaching 68.82 on MMMU and 80.43 on MathVista,hence used as the default in our experiments. Larger blocks reduce performance. Together with the full-sequence denoising comparison in Table 4, these results support block-wise generation.

6 Conclusion

We introduced **UniFlux**, a unified multimodal framework that repurposes the text stream of a pretrained MM-DiT for visual understanding and reasoning while preserving its image-generation capability. Our framework combines block-wise discrete flow matching for coherent, variable-length reasoning with X-GRPO for joint alignment of understanding and generation through coupled language- and image-side rewards. Extensive experiments demonstrate strong performance across multimodal understanding, reasoning, OCR, and image generation, highlighting pretrained MM-DiTs as a promising foundation for unified multimodal models.

AI USE STATEMENT

We used large language models (LLMs) solely to polish the writing of this paper, including improving sentence clarity and correcting grammar. We also used an AI coding agent to help optimize our implementation and accelerate experiment execution. All research ideas, methods, experimental designs, and analyses are our own, and we carefully reviewed all AI-assisted text and code. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide complete implementation details, including all training and evaluation hyperparameters, in Appendix A. The full set of prompt templates used in our experiments is included in the Supplementary Material. Together, these materials are sufficient to replicate our results. To support future research, we will publicly release our code and model weights upon publication.

ETHICS STATEMENT

All training data used in this work are drawn from publicly available sources and are used in accordance with their respective licenses and terms of use. We recognize that, like other generative models, our framework could potentially be misused, for example, to create misleading or harmful content. We explicitly discourage any such use and encourage downstream users to adhere to established ethical guidelines and to apply appropriate safeguards when deploying this technology.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025.

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22934–22946, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv 2308.12966*, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *International Conference on Learning Representations*, pp. 4965–4987, 2023.

Black Forest Labs. FLUX.2 [dev]. `https://bfl.ai/models/flux-2`, 2026a. Accessed: 2026-08-28.

Black Forest Labs. FLUX.2 [klein], 2026b. URL `https://bfl.ai/models/flux-2-klein`. Accessed: 2026-08-28.

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

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 *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Google DeepMind. Gemini 3.1 pro: Model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

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

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. GoT-R1: Unleashing reasoning capability of MLLM for visual generation with reinforcement learning. *Advances in Neural Information Processing Systems*, pp. 67680–67708, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Advances in Neural Information Processing Systems*, 38, 2026.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025a.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13869–13920, 2025b.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *International Conference on Learning Representations*, pp. 132869–132895, 2026.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *International Conference on Machine Learning (ICML)*, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2022.

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.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research (TMLR)*, 2024.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 39856–39890, 2025. doi: 10.52202/085713-1330.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*, 2016.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jie Liu, Zilyu Ye, Linxiao Yuan, Shenhan Zhu, Yu Gao, Jie Wu, Kunchang Li, Xionghui Wang, Xiaonan Nie, Weilin Huang, et al. Unigrpo: Unified policy optimization for reasoning-driven visual generation. *arXiv preprint arXiv:2603.23500*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *International Conference on Learning Representations*, 2023b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 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 (NeurIPS)*, 2025.

Jian Ma, Qirong Peng, Xu Guo, Chen Chen, Haonan Lu, and Zhenyu Yang. X2i: Seamless integration of multimodal understanding into diffusion transformer via attention distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16733–16744, 2025a.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025b.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 38405–38453, 2025.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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*, pp. 10684–10695, 2022.

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, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Tao Shen, Xin Wan, Taicai Chen, Rui Zhang, Junwen Pan, Dawei Lu, Fanding Lei, Zhilin Lu, Yunfei Yang, Chen Cheng, et al. Mammothmoda2: A unified ar-diffusion framework for multimodal understanding and generation. *arXiv preprint arXiv:2511.18262*, 2025.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. In *International Conference on Learning Representations (ICLR)*, 2026.

Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multimodal diffusion transformers for enhanced prompt-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19492–19502, 2025.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A dataset for image captioning with reading comprehension. In *European Conference on Computer Vision (ECCV)*, 2020.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *International conference on learning representations*, volume 2024, pp. 12352–12380, 2024.

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

Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into RL for Image Generation with CoT: A Study on DPO vs. GRPO. In *Advances in Neural Information Processing Systems*, volume 38, pp. 115632–115655. Curran Associates, Inc., 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025a.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *Advances in Neural Information Processing Systems*, 38:81402–81440, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-GRPO: Pairwise preference reward-based GRPO for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025c.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

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 (ICLR)*, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 47490–47518. Curran Associates, Inc., 2025b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Haochen Wang, Zhendong Wang, Bin Lin, Hao Li, Xinyan Xiao, et al. 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. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2025.

Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. AR-GRPO: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*, 2025.

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*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Xin Zhang, Haochen Wang, Yikang Zhou, Zhuochen Wang, Xiangtai Li, and Robby T Tan. Actor as its own critic: Unifying region understanding and localization via cyclegrpo. In *European Conference on Computer Vision*, pp. 586–604. Springer, 2026.

Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024.

Mingzhe Zheng et al. Manifold-aware exploration for reinforcement learning in video generation. *arXiv preprint arXiv:2603.21872*, 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 (ICLR)*, 2025.

APPENDIX

A Implementation Details **17**

B Dataset **20**

B.1 Generation–Understanding Alignment Data 20

B.2 Visual Instruction Tuning Data 20

B.3 Visual Reasoning Data 20

B.4 Multi-Reference Image Understanding Data 21

B.5 Data Curation and Preprocessing 21

C Extended Related Work **21**

C.1 Multimodal Diffusion Transformers 21

C.2 Unified Models for Generation and Understanding 22

C.3 Preference and Reward Optimization for Generative Models 22

D Block-wise Attention and Discrete Flow Matching **23**

E Reward Functions **25**

F X-GRPO Training and Cross-Modal Advantage Properties **28**

F.1 X-GRPO Training Algorithm 28

F.2 Properties of the Shared Cross-Modal Advantage 29

G Layer-wise Semantic Probing **30**

H Additional Ablation results **31**

H.1 Effect of Reasoning Length on Accuracy and Inference Cost 31

H.2 Sensitivity to the VLM judge. 31

H.3 Effect of text representation depth. 31

H.4 Effect of LoRA rank. 32

I Results on OCR and Complex Image Understanding **32**

J Detailed Text-to-Image Generation Results **33**

K Additional Visual Comparison **33**

K.1 Evidence extraction and answerability. 33

K.2 Qualitative Comparison with Alignment Baselines. 34

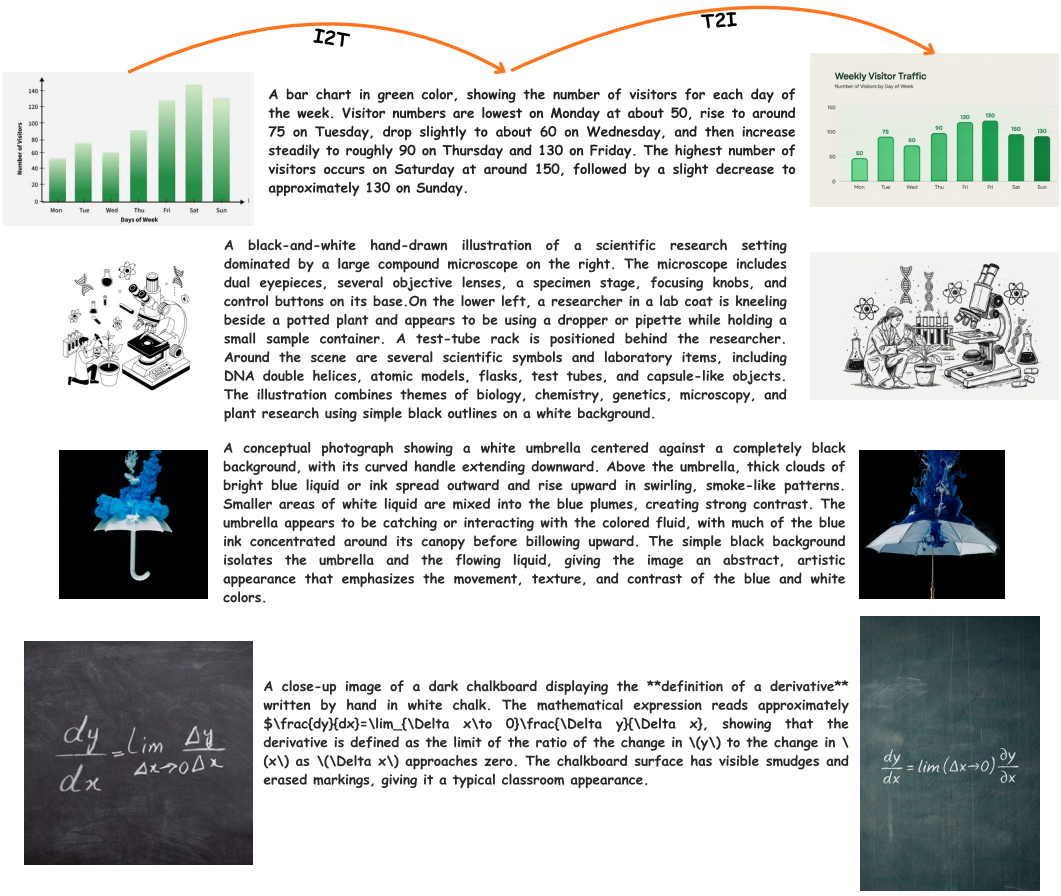


Figure 9: **Qualitative image-to-text-to-image cycle consistency.** From left to right, **UniFlux** first converts each source image into a detailed textual description through its image-to-text pathway and then uses this description to reconstruct the visual content through its text-to-image pathway. Across diverse visual domains, the reconstructed images preserve the dominant semantic content and global composition of the inputs, illustrating cross-modal semantic consistency rather than exact pixel-level reconstruction.

A Implementation Details

Stage I: Multimodal understanding alignment. We optimize the block-wise discrete flow-matching objective for 15,000 steps on 40 NVIDIA H100 GPUs with 80 GB memory each. Following Sec. 3, the pretrained MM-DiT weights remain frozen, while the text-stream LoRA modules, token embedding, and vocabulary prediction head are trained. LoRA uses rank $r = 64$ and scaling factor $\alpha = 16$. The token embedding and prediction head use the vocabulary of the original MM-DiT text encoder.

We use a micro-batch of 16 examples per GPU with four gradient accumulation steps, giving an effective global batch of 2,560. Optimization uses AdamW with $(\beta_1, \beta_2) = (0.95, 0.99)$ and weight decay 10^{-2}. The learning rate increases linearly from 10^{-7} to 2×10^{-5} over the first 2,000 optimizer steps and subsequently follows cosine annealing. Training uses BF16 precision.

The maximum training reasoning-sequence length is 1,024 tokens, with 32 tokens per response block. We denote the block length by $L_{\text{blk}} = 32$ and the number of blocks in a response by B. The training-time step-size setting is $\Delta t_{\text{train}} = 1/1000$; this setting is separate from both the optimizer-step count and the rollout/inference sampling budget. Images are assigned to aspect-ratio-preserving padded resolution buckets: 512×512, 512×1024, 1024×512, 256×512, and 512×256.

Stage II: Cross-modal X-GRPO alignment. We initialize Stage II from the Stage-I checkpoint by merging the learned LoRA updates into their corresponding backbone projection weights. We then

Table 5: **UniFlux training configuration.** Micro-batch sizes reported per GPU. The effective global batch counts training examples in Stage I and input prompts in Stage II, before rollout expansion.

Configuration	**Stage I:** Understanding Alignment	**Stage II:** X-GRPO Alignment
Objective	Block-wise discrete flow matching	Cross-modal X-GRPO
Initialization	Pretrained MM-DiT	Stage-I checkpoint with LoRA adapters merged
Trainable parameters	Text-stream LoRA, token embedding, and vocabulary head	Full multimodal model
Hardware	$40\times$ H100 80 GB	$32\times$ H100 80 GB
Optimizer steps	15,000	3,000
Micro-batch size	16 per GPU	12 prompts per GPU
Gradient accumulation	4	4
Effective global batch	$16 \times 40 \times 4 = 2{,}560$	$12 \times 32 \times 4 = 1{,}536$ prompts
Optimizer	AdamW, $(\beta_1, \beta_2) = (0.95, 0.99)$	Muon
Learning rate	$10^{-7} \rightarrow 2 \times 10^{-5}$	$10^{-9} \rightarrow 5 \times 10^{-6}$
LR schedule	2,000-step warmup, then cosine annealing	Linear
Weight decay	10^{-2}	10^{-1}
LoRA configuration	Rank $r = 64$, scaling $\alpha = 16$	Merged before full-model fine-tuning
LoRA target modules	Q-proj, K-proj, V-proj, Out-proj, and In-proj	No separate LoRA adapters after merging
Text length setting	1,024 tokens for training	768 tokens for rollouts
Response block size	$L_{\text{blk}} = 32$	$L_{\text{blk}} = 32$
Image resolutions	512×512, 512×1024, 1024×512, 256×512, and 512×256	Variable-resolution batches up to 1024×1024
Precision	BF16 training	FP32 rollouts; BF16 policy updates

Table 6: **Sampling configuration.** The training-time step-size setting $\Delta t_{\text{train}} = 1/1000$ is distinct from the 5-step text sampler used for both RL rollouts and inference. Each input produces eight text responses and 64 generated images.

Configuration	**Image-to-Text**	**Text-to-Image**
Sampler	τ-leaping	Euler–Ancestral
RL rollout iterations	5 per response block	40 per image
Inference iterations	5 per response block	40 per image
Guidance references	Separate null-image and no-history predictions	Null-text prediction
Reported CFG scale	6.0	8.0
Categorical temperature	0.95	–
Vocabulary truncation	None; full-vocabulary sampling	–
RL rollout length	Maximum sequence length 768 tokens	–
Default evaluation setting	300 reasoning tokens	Up to 1024×1024 pixels
Text responses per input	$G_T = 8$	–
Instructions per response	One	–
Images per instruction	–	$G_I = 8$
Response–image pairs	$G_T G_I = 64$ per input	

fine-tune the full multimodal model using X-GRPO, allowing the understanding and generation objectives to jointly update the shared backbone and output interfaces. Training runs for 3,000 optimizer steps on 32 NVIDIA H100 GPUs with 80 GB memory each. We use a micro-batch of 12 input prompts per GPU and four gradient accumulation steps, giving an effective global batch of 1,536 prompts before response and image sampling.

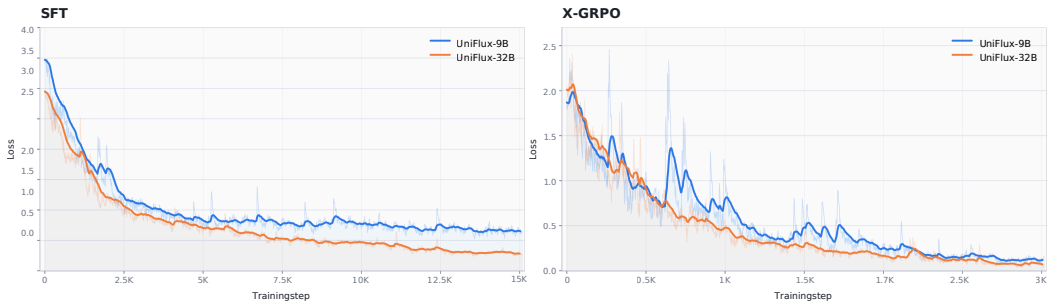


Figure 10: **Training dynamics of UniFlux.** Optimization trajectories of the 9B and 32B variants during Stage-I supervised fine-tuning (left) and Stage-II X-GRPO alignment (right). Both stages show an initial decrease in their respective objectives followed by more stable optimization. The displayed Stage-II trajectory exhibits greater early variation for the 9B model.

Optimization uses Muon with weight decay 10^{-1} and a linear learning-rate schedule, with an initial rate of 10^{-9} and a maximum rate of 5×10^{-6}. Text rollouts use a maximum sequence length of 768 tokens and 32-token response blocks. Image rollouts use variable resolutions up to 1024×1024. Rollouts are sampled in FP32, and policy updates use BF16.

Hierarchical rollout construction. For an input image–query pair (I, q), the understanding pathway samples $G_T = 8$ responses $\{A^{(r)}\}_{r=1}^{G_T}$. Each response is converted into one image-generation instruction, $X^{(r)} = G_\psi(q, I, A^{(r)})$. For each instruction, the generation pathway samples $G_I = 8$ images $\{\widehat{I}^{(r,j)}\}_{j=1}^{G_I}$. Thus, each input yields eight distinct response trajectories, eight instructions, and 64 image trajectories, forming $G_T G_I = 64$ response–instruction–image triples. The images associated with a response share its text trajectory and derived instruction.

Following the main method, text rewards are normalized across the eight responses for an input, while image rewards are normalized across all 64 generated images for that input. We subtract the group mean and divide by the population standard deviation plus 10^{-8}; zero-variance groups receive zero advantages.

Rollout and inference budgets. Both text inference and RL text rollouts use $N_T = 5$ sampling iterations per response block. The training-time setting $\Delta t_{\text{train}} = 1/1000$ does not specify the number of sampling iterations used for either procedure. The default evaluation reasoning-token setting is 300, whereas the maximum rollout sequence length is 768. Text-to-image rollouts and inference use 40 Euler–Ancestral steps.

Classifier-free guidance at inference. Let $H = \widehat{\mathbf{y}}_{<b}$ denote the completed response prefix. For text generation, we evaluate the conditional prediction together with separate null-image and no-history reference predictions:

$$\begin{aligned} \log \widehat{\pi}_{\theta \text{i2t}} = \log \pi_{\theta \text{i2t}}(\cdot \mid I, q, H) \\ + 6.0\,[\log \pi_{\theta \text{i2t}}(\cdot \mid I, q, H) - \log \pi_{\theta \text{i2t}}(\cdot \mid \varnothing, q, H)] \\ + 6.0\,[\log \pi_{\theta \text{i2t}}(\cdot \mid I, q, H) - \log \pi_{\theta \text{i2t}}(\cdot \mid I, q, \varnothing)] - \log Z_T. \end{aligned} \tag{4}$$

The first guidance term amplifies the prediction difference induced by visual conditioning, while the second amplifies the difference induced by the completed response prefix. These two terms strengthen the influence of the input image and preceding reasoning context, respectively. All three branches retain the query and current noisy block.

For image generation, the corresponding log-space guidance expression is

$$\begin{aligned} \log \widehat{\pi}_{\theta \text{t2i}} = \log \pi_{\theta \text{t2i}}(\cdot \mid \varnothing) \\ + 8.0\,[\log \pi_{\theta \text{t2i}}(\cdot \mid X) - \log \pi_{\theta \text{t2i}}(\cdot \mid \varnothing)] - \log Z_I. \end{aligned} \tag{5}$$

Here, H denotes the completed response prefix, X is the image-generation instruction, and Z_T, Z_I normalize the guided policies. The current noisy state and timestep are implicit.

Text policy and rollout computation. To reduce repeated policy-evaluation cost, we cache the old-policy categorical log-probabilities, indexed by response, block, denoising timestep, and scored token position. During optimization, we recompute current-policy predictions on the corresponding noisy block states and token targets, using the same conditioning and scoring configuration.

Let $\mathcal{E}_i$ denote the set of scored block–timestep–position indices for response i. For each $(b, n, k) \in \mathcal{E}_i$, define

$$\ell^{(i)}_{\theta,b,n,k} = \log p^T_\theta \left(y^{\star(i)}_{b,n,k} \mid \widetilde{\mathbf{y}}^{(i)}_{b,n}, t_n, c^{(i)}_b \right), \tag{6}$$

where $y^{\star(i)}_{b,n,k}$ is the recorded scoring target, $\widetilde{\mathbf{y}}^{(i)}_{b,n}$ is the noisy block state, and $c^{(i)}_b = (I, Q, \widehat{\mathbf{y}}^{(i)}_{<b})$. The distribution p^T_θ is the untruncated categorical distribution used for policy scoring.

Training dynamics. Figure 10 reports the optimization trajectories of the **UniFlux**-9B and **UniFlux**-32B variants. Stage-I losses decrease rapidly at the beginning and more gradually later in training. Stage II exhibits greater early fluctuations, particularly for the 9B variant, before becoming more stable over the displayed interval. These curves characterize the reported optimization objectives; downstream performance is evaluated separately.

B Dataset

We construct the training data from public multimodal datasets and synthetic reasoning annotations. The data comprise three collections: generation–understanding alignment (50.0M), visual instruction tuning (15.0M), and visual reasoning (20.0M), detailed in Sections B.1–B.3. Counts are approximate and refer to task examples; an image may be paired with multiple supervision targets across tasks and collections. Multi-reference image data are described in Section B.4.

B.1 Generation–Understanding Alignment Data

The alignment collection contains four tasks constructed from a shared pool of images: detailed image description, instruction response, reasoning trace prediction, and text-to-image generation. Image sources include ShareGPT4V (Chen et al., 2024b), M^{3}IT (Li et al., 2023), and LLaVA-Instruct-150K (Liu et al., 2023a). For the first three tasks, the input is an image paired with a task instruction, and the target is the corresponding description, response, or reasoning trace. For text-to-image generation, the input is a textual description and the target is its associated image.

Table 7: **Composition of the generation–understanding alignment collection.** Counts denote task examples.

Task type	**Examples**
Image → Detailed description	10.0M
Image → Instruction response	15.0M
Image → Reasoning trace	20.0M
Text → Image generation	5.0M
Total (approximately)	**50.0M**

Table 7 lists 10.0M description examples, 15.0M instruction-response examples, 20.0M reasoning-trace examples, and 5.0M image-generation examples. The collection therefore contains approximately 45.0M text-target examples and 5.0M image-target examples.

B.2 Visual Instruction Tuning Data

The visual instruction collection contains approximately 15.0M examples covering captioning, visual question answering, optical character recognition (OCR), document understanding, chart interpretation, and diagram reasoning. ShareGPT4V (Chen et al., 2024b) and TextCaps (Sidorov et al., 2020) contain image descriptions and text-aware captions. OCR-VQA (Mishra et al., 2019), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and AI2D (Kembhavi et al., 2016) contain reading and structured visual understanding examples. The collection also includes multimodal instructions from M^{3}IT and LLaVA-Instruct-150K (Liu et al., 2023a), and open-ended, scientific, and mathematical question–answer examples from Mammoth-VL (Guo et al., 2025b). Table 9 reports the source contributions, which sum to 14.998M examples.

B.3 Visual Reasoning Data

The visual reasoning collection contains approximately 20.0M examples covering scientific question answering, diagram understanding, document and chart reasoning, and visual commonsense reasoning. It combines source question–answer annotations and generated reasoning traces. As listed in Table 8, the sources include ShareGPT4V (Chen et al., 2024b), Visual Genome (Krishna et al., 2017), ScienceQA (Lu et al., 2022), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021),

Table 8: **Composition of the visual reasoning collection.** Source contributions total 20.018M examples, rounded to $\sim$ 20.0M.

Source	Examples
ShareGPT4V Reasoning (Chen et al., 2024b)	1.2M
Visual Genome QA (Krishna et al., 2017)	1.7M
LLaVA-OneVision-1.5 (Stage 2) (An et al., 2025)	15.0M
ScienceQA (Lu et al., 2022)	21K
ChartQA (Masry et al., 2022)	32K
DocVQA (Mathew et al., 2021)	50K
AI2D (Kembhavi et al., 2016)	15K
Synthetic GPT-4V CoT (Achiam et al., 2023)	2.0M
Total (approximately)	**20.0M**

Table 9: **Composition of the visual instruction tuning collection.** Source contributions total 14.998M examples, rounded to $\sim$ 15.0M.

Source	Examples
ShareGPT4V (Chen et al., 2024b)	1.2M
M^3IT (Li et al., 2023)	2.4M
LLaVA-Instruct-150K (Liu et al., 2023a)	158K
Mammoth-VL (Guo et al., 2025b)	10.0M
OCR-VQA (Mishra et al., 2019)	1.0M
TextCaps (Sidorov et al., 2020)	143K
DocVQA (Mathew et al., 2021)	50K
ChartQA (Masry et al., 2022)	32K
AI2D (Kembhavi et al., 2016)	15K
Total (approximately)	**15.0M**

and AI2D (Kembhavi et al., 2016). LLaVA-OneVision-1.5 Stage 2 (An et al., 2025) contributes 15.0M examples.

The collection also contains 2.0M synthetic chain-of-thought examples generated with GPT-4V (Achiam et al., 2023). These examples include intermediate reasoning steps and are filtered for answer correctness and formatting before inclusion.

B.4 Multi-Reference Image Understanding Data

We use Mantis-Instruct (Jiang et al., 2024) for multi-image understanding. It contains 721K interleaved instruction examples from 14 subsets, covering co-reference, comparison, reasoning, and temporal understanding. The examples combine multiple images and text within multi-turn dialogues. Mantis-Instruct is reported separately from the source contributions in Tables 9 and 8.

B.5 Data Curation and Preprocessing

We remove near-duplicate images and discard caption-bearing image–text pairs with empty, excessively short, or predominantly boilerplate captions. Examples are formatted by task: image descriptions and instruction responses use textual targets; reasoning examples retain their associated reasoning annotations; and image-generation examples pair textual descriptions with target images. Multi-reference examples preserve the interleaved image–text dialogue structure.

C Extended Related Work

C.1 Multimodal Diffusion Transformers

Recent text-to-image models have increasingly adopted Multimodal Diffusion Transformer (MM-DiT) architectures, including Stable Diffusion 3 (Esser et al., 2024) and FLUX.1 (Labs, 2024). Earlier U-Net-based latent diffusion models (Rombach et al., 2022; Podell et al., 2024) typically combine image self-attention with cross-attention to text embeddings. MM-DiTs apply joint attention over image and text tokens, allowing both representations to exchange information and update throughout the network. This changes how textual conditions interact with visual features during denoising and has motivated adaptations of editing, control, and interpretability methods originally developed for U-Net architectures.

(Shin et al., 2025) provide a structural analysis of MM-DiT attention, showing that the joint attention matrix admits a 2×2 block decomposition into image-to-image (I2I), text-to-image (T2I), image-to-text (I2T), and text-to-text (T2T) sub-blocks, of which I2I behaves analogously to U-Net self-attention and T2I to its cross-attention. They exploit this structure for training-free prompt-based editing by replacing the target branch's image-side query and key projections with those of a source branch during early denoising steps. A key observation is that intervening at the projection

level rather than at the post-softmax attention map preserves fused scaled dot-product attention kernels, avoiding a roughly 3× inference slowdown. Their method generalizes across SD3, SD3.5, and Flux.1 variants composes with rectified-flow inversion (Liu et al., 2023b) for real-image editing.

C.2 Unified Models for Generation and Understanding

Large language models have become the dominant paradigm for multimodal understanding and reasoning, while diffusion and flow-based models remain the prevailing choice for high-fidelity visual generation(Shi et al., 2026; Yang et al., 2025; Swerdlow et al., 2025). Recent work has therefore explored how these capabilities can be combined within a unified model(Deng et al., 2025; Xie et al., 2025a;b). Existing approaches can be broadly organized into three categories according to the objectives used for textual and visual generation.

The first category adopts *autoregressive modeling for both understanding and generation*. Models such as Chameleon, LWM, Emu3, and Janus-Pro quantize visual content into discrete tokens and apply causal next-token prediction to interleaved text–vision sequences (Chameleon Team, 2024; Liu et al., 2024; Wang et al., 2024b; Chen et al., 2025). Although this provides a homogeneous training objective, visual generation requires sequentially decoding long image or video token sequences, creating a substantial inference bottleneck at high spatial and temporal resolutions.

The second category combines *autoregressive language modeling with diffusion or flow-based visual generation*. Show-o uses autoregressive decoding for text and masked discrete diffusion for image tokens (Xie et al., 2025a), while Show-o2, Transfusion, BAGEL, MammothModa2, JanusFlow, and MonoFormer combine next-token prediction with continuous diffusion or flow matching for visual synthesis (Xie et al., 2025b; Zhou et al., 2025; Deng et al., 2025; Shen et al., 2025; Ma et al., 2024; Zhao et al., 2024). Although these methods differ in how extensively parameters are shared, they optimize distinct language and visual objectives, making unified training a multi-objective optimization and cross-modal alignment problem.

The third category eliminates autoregressive decoding and employs *diffusion or flow modeling for both understanding and generation*. D-DiT combines continuous flow matching for image latents with discrete masked diffusion for text (Li et al., 2025). UniDisc and Muddit instead represent both modalities as discrete tokens and apply masked discrete diffusion to image and text prediction (Swerdlow et al., 2025; Shi et al., 2026). Related approaches further explore large-scale multimodal masked diffusion, as in MMaDA (Yang et al., 2025), or discrete flow matching across both modalities, as in FUDOKI (Wang et al., 2025b).

C.3 Preference and Reward Optimization for Generative Models

Reinforcement learning from human feedback in language models. Aligning generative models with human preferences was first established at scale for large language models. The canonical recipe trains a reward model on human comparison data and optimizes the policy against it with Proximal Policy Optimization (PPO) (Ouyang et al., 2022; Schulman et al., 2017). Direct Preference Optimization (DPO) (Rafailov et al., 2023) subsequently showed that the reward-modeling and RL stages could be collapsed into a single supervised objective over preference pairs, trading some performance for substantially lower cost and complexity. More recently, Group Relative Policy Optimization (GRPO), introduced in DeepSeekMath (Shao et al., 2024) and scaled in DeepSeek-R1 (Guo et al., 2025a), removed the value network entirely: it samples a group of completions per prompt and computes each completion's advantage by standardizing its reward against the group. This group-relative formulation underpins much of the recent progress in LLM reasoning and has become the dominant on-policy alternative to PPO and DPO.

Reward optimization for diffusion and flow-matching generation. A parallel line of work transfers these ideas to visual generation. Early policy-gradient methods such as DDPO (Black et al., 2023) and DPOK (Fan et al., 2023) cast the denoising process as a Markov decision process and optimize a black-box reward, but proved unstable beyond small prompt sets. Differentiable-reward methods like ReFL (Xu et al., 2023) backpropagate reward gradients directly, at the cost of requiring a differentiable reward and incurring heavy memory overhead, while Diffusion-DPO (Wallace et al., 2024) adapts the DPO objective to preference pairs of images. The arrival of GRPO has prompted a wave of online RL methods for flow-matching models. DanceGRPO (Xue et al., 2025)

and Flow-GRPO (Liu et al., 2025) concurrently adapt GRPO to rectified flows, both converting the deterministic sampling ODE into an equivalent SDE that preserves the model's marginal distributions while enabling the stochastic rollouts that policy-gradient estimation requires; Flow-GRPO additionally introduces a denoising-reduction strategy and reports large gains on compositional generation and text rendering, while DanceGRPO further demonstrates the approach across diffusion models, video generation, and image-to-video tasks. A recurring failure mode in this setting is reward hacking, motivating refinements such as Pref-GRPO (Wang et al., 2025c), which replaces reward-score maximization with pairwise preference fitting for more stable advantages, and methods that constrain the importance-ratio bias or exploration noise introduced by the ODE-to-SDE conversion (Zheng et al., 2026).

Reward optimization for autoregressive and unified visual generation. GRPO has also been applied to autoregressive and unified image generators, where images are produced as discrete token sequences. T2I-R1 (JIANG et al., 2025) introduces BiCoT-GRPO, jointly optimizing a semantic-level chain-of-thought (a textual plan) and a token-level chain-of-thought (the image tokens) within a unified multimodal model, using an ensemble of vision experts as rewards. Related efforts extend GRPO-style optimization to reasoning-augmented autoregressive generation (Duan et al., 2025), to autoregressive image models directly (Yuan et al., 2025), and to masked generative models by redefining the transition probability (Luo et al., 2025). Comparative studies of DPO versus GRPO for image generation with chain-of-thought reasoning (Tong et al., 2025) further indicate that the relative merits of preference-based and group-relative objectives depend on the generative paradigm and reward structure. Collectively, this body of work shows that the preference- and reward-optimization toolkit developed for language models transfers across the major visual generation paradigms—diffusion, flow matching, and autoregressive—though stability and reward hacking remain open challenges, particularly for video.

D Block-wise Attention and Discrete Flow Matching

Block-causal attention. Figure 11 illustrates the attention pattern used to implement block-wise discrete flow matching. Rows denote query tokens and columns denote the corresponding key–value tokens. Let $\mathbf{I}$ denote the image tokens, $\mathbf{Q}$ the prompt tokens, and

$$\mathbf{A} = \left[\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(M)}\right] \tag{7}$$

the response divided into M non-overlapping blocks. The permitted attention context can be summarized as

$$\mathcal{K}(I_i) = \mathbf{I}, \tag{8}$$

$$\mathcal{K}(Q_i) = \{\mathbf{I}, \mathbf{Q}\}, \tag{9}$$

$$\mathcal{K}\left(A_i^{(m)}\right) = \left\{\mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}, \mathbf{A}^{(m)}\right\}, \tag{10}$$

where $\mathcal{K}(\cdot)$ denotes the tokens available as keys and values.

The purple region in the figure represents image self-attention, while the green and dark-blue regions allow prompt and answer queries to retrieve visual and prompt information. For an answer block $\mathbf{A}^{(m)}$, the light-blue region provides access to all previously completed blocks $\mathbf{A}^{(<m)}$. These blocks remain fixed and therefore act as stable causal context. In contrast, the orange diagonal region is dense within the active block: every token in $\mathbf{A}^{(m)}$ can attend to every other token in that block. Consequently, the active block is refined bidirectionally rather than token-by-token. All attention paths toward future blocks $\mathbf{A}^{(>m)}$ are masked, preventing future-token leakage and preserving causal progression at the block level.

This attention structure realizes the factorization

$$p_\theta\left(\mathbf{A} \mid \mathbf{I}, \mathbf{Q}\right) = \prod_{m=1}^{M} p_\theta\left(\mathbf{A}^{(m)} \mid \mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}\right). \tag{11}$$

Thus, generation is autoregressive across blocks, whereas the tokens inside the current block are jointly modeled through discrete flow matching. Importantly, coordinate-wise CTMC transitions do not imply independent token predictions: the transition rate at any position is conditioned on the complete active-block state through the dense orange attention region.

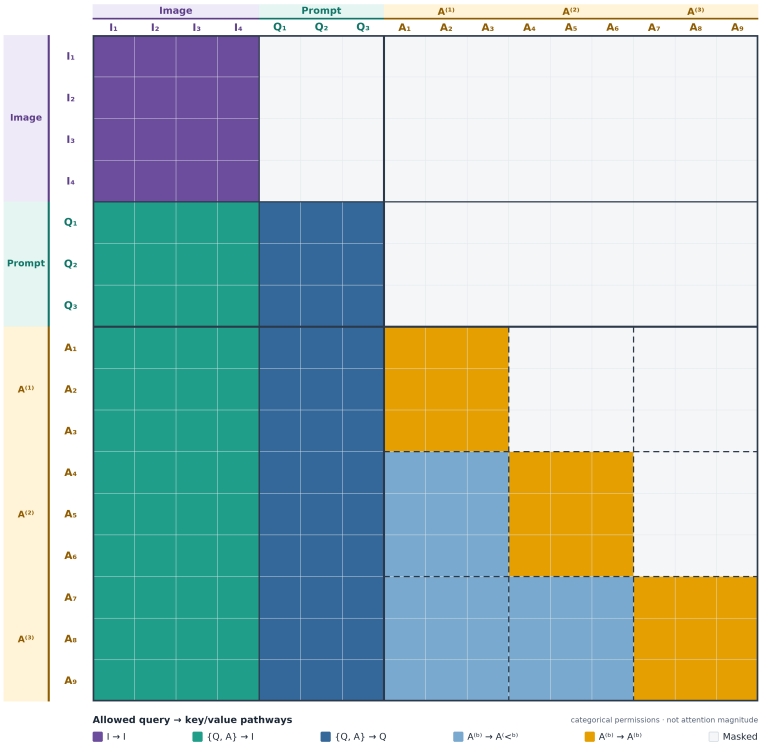


Figure 11: **Block-causal attention pattern for block-wise discrete flow matching.** Rows represent query tokens and columns represent key–value tokens. Image tokens attend within the visual stream, while prompt tokens attend to both the image and prompt context. For an answer block $\mathbf{A}^{(m)}$, each token can attend to the image tokens $\mathbf{I}$, prompt tokens $\mathbf{Q}$, all previously completed answer blocks $\mathbf{A}^{(<m)}$, and every token within the current block $\mathbf{A}^{(m)}$. The dense orange diagonal regions therefore enable bidirectional and parallel refinement inside each active block, whereas attention to future blocks $\mathbf{A}^{(>m)}$ is masked to preserve causal generation across blocks. Purple, green, dark-blue, light-blue, and orange regions denote the permitted image self-attention, image-conditioning, prompt-conditioning, previous-block, and within-block attention pathways, respectively; white regions are masked.

Discrete flow within the active block. For block m, define the fixed conditioning context as

$$\mathbf{c}_m = \left(\mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}\right). \tag{12}$$

Only the active block evolves through a discrete process $\mathbf{Z}_t^{(m)}$, while $\mathbf{c}_m$ remains unchanged. Using an absorbing mask token $\mathtt{M}$, we define

$$q_t\left(\mathbf{z} \mid \mathbf{A}^{(m)}\right) = \prod_{i \in \mathcal{I}_m} \left[(1-\alpha_t)\,\mathbb{1}[z_i = \mathtt{M}] + \alpha_t\,\mathbb{1}[z_i = A_i]\right], \tag{13}$$

We use the linear schedule $\alpha_t = t$, where the appendix time variable satisfies $t = 1 - \tau$. Thus, sampling from $t = 0$ (fully masked) to $t = 1$ (clean) is equivalent to $\tau = 1 \to 0$ in the main text. The MM-DiT timestep conditioning remains $\tau = 1 - t$, and training supervises only masked positions, consistent with Eq. (1).

For a masked position, the learned transition rate is

$$R_{\theta,t,i}^{(m)}(M, a \mid \mathbf{z}, c_m) = \frac{\dot{\alpha}_t}{1-\alpha_t} p_\theta\left(A_i^{(m)} = a \mid Z_t^{(m)} = \mathbf{z}, c_m, 1-t\right). \tag{14}$$

The schedule-dependent hazard determines *when* a masked position is revealed, while the network determines *which token* should be selected. During training, the model predicts the clean tokens at the masked positions of the active block:

$$\mathcal{L}_{\mathrm{BDFM}} = \mathbb{E}\left[-\sum_{i \in \mathcal{I}_m} \mathbf{1}\left[Z_{t,i}^{(m)} = M\right] \log p_\theta\left(A_i^{(m)} \mid Z_t^{(m)}, c_m, 1-t\right)\right]. \tag{15}$$

After the flow reaches its terminal state, the completed block is committed and becomes part of the context for the next block.

Consistency proof. First consider a fixed clean target block $\mathbf{A}^{(m)}$. A masked coordinate transitions from M to its target token with hazard

$$\lambda_t = \frac{\dot{\alpha}_t}{1-\alpha_t}. \tag{16}$$

Because

$$(1-\alpha_t)\lambda_t = \dot{\alpha}_t, \tag{17}$$

the probability mass leaving the mask state is exactly the rate at which clean-token probability mass increases along the prescribed masking path. Therefore, the conditional path $q_t(\mathbf{z} \mid \mathbf{A}^{(m)})$ satisfies the corresponding Kolmogorov forward equation.

When the clean target block is unknown, averaging this target-conditioned transition over its posterior gives

$$R_{t,i}^{*(m)}\left(\mathtt{M}, a \mid \mathbf{z}, \mathbf{c}_m\right) = \mathbb{E}\left[\lambda_t \mathbb{1}[A_i = a] \mid \mathbf{Z}_t^{(m)} = \mathbf{z}, \mathbf{c}_m\right] \tag{18}$$

$$= \lambda_t p_{\text{data}}\left(A_i = a \mid \mathbf{Z}_t^{(m)} = \mathbf{z}, \mathbf{c}_m\right). \tag{19}$$

Hence, if posterior prediction is exact and the CTMC is simulated exactly, the terminal distribution of block m satisfies

$$p_\theta\left(\mathbf{A}^{(m)} \mid \mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}\right) = p_{\text{data}}\left(\mathbf{A}^{(m)} \mid \mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}\right). \tag{20}$$

Applying this equality to every block and using the chain rule yields

$$p_\theta\left(\mathbf{A} \mid \mathbf{I}, \mathbf{Q}\right) = \prod_{m=1}^{M} p_\theta\left(\mathbf{A}^{(m)} \mid \mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}\right) \tag{21}$$

$$= \prod_{m=1}^{M} p_{\text{data}}\left(\mathbf{A}^{(m)} \mid \mathbf{I}, \mathbf{Q}, \mathbf{A}^{(<m)}\right) \tag{22}$$

$$= p_{\text{data}}\left(\mathbf{A} \mid \mathbf{I}, \mathbf{Q}\right). \tag{23}$$

Therefore, the block-wise decomposition itself introduces no probabilistic approximation. In practice, errors arise only from imperfect posterior estimation, generated-prefix errors, or numerical approximation of the conditional CTMC.

E Reward Functions

Image-side HPSv3 Reward. HPSv3 provides a perceptual-quality reward for each generated image, encouraging the generation pathway to produce visually coherent and high-quality samples, and preventing the optimization from favoring semantically relevant but visually degraded generations.

Image-side SigLIP Reward. SigLIP measures the semantic alignment between the generated image and the image-generation instruction derived from the model's reasoning response. This reward ensures that the generated image follows the objects, attributes, and relationships specified by the instruction, thereby strengthening image–text consistency.

Language-side VLM-as-a-Judge Reward. For the language-side reward, a VLM judge evaluates each sampled reasoning–answer response conditioned on the input image, user query, and reference answer. As specified in prompt 12, the judge evaluates answer correctness, visual grounding, reasoning validity, completeness, and format following, while explicitly penalizing hallucination and unnecessary verbosity. The resulting weighted score encourages responses that are correct, visually grounded, complete, and appropriately formatted.

Image-side VLM-as-a-Judge Reward. We use a VLM judge to evaluate the semantic faithfulness of the regenerated image. As specified in prompt 13, the judge receives the original image, the generated image, and the corresponding generation instruction, and assesses whether the regenerated content preserves the important visual information from the original image. This high-level semantic reward complements the perceptual signal from HPSv3 and the embedding-based alignment provided by SigLIP.

```
You are a strict multimodal reward judge for visual question answering and image
understanding.
You will be given: 1. An image, 2. A user query, 3. A reference answer, 4. A
candidate model response.
Your job is to evaluate how good the candidate response is.

The data may include many task types:
* OCR and document understanding | Math from images | Charts, tables, and plots | General
  VQA
* Captioning | Counting | Spatial reasoning | Science diagrams | Multi-step visual
  reasoning

Evaluate only what is supported by the image and the query. Do not reward plausible but
unsupported claims.
Return JSON only.

Scoring rules:
1. answer_correctness: 0 to 1 (Score whether the final answer is correct)
  * 1.0 = fully correct | 0.7 = mostly correct | 0.4 = partially correct | 0.0 = wrong or
    missing
2. visual_grounding: 0 to 1 (Score whether response is supported by visible evidence)
  * 1.0 = all claims supported | 0.5 = some supported, some uncertain | 0.0 = mostly
    unsupported
3. reasoning_validity: 0 to 1 (Score whether reasoning, calculations, comparisons, or
  logic are valid)
  For simple questions, score based on whether the response uses the right visual
  evidence.
  For math/chart/document tasks, check intermediate logic carefully.
4. completeness: 0 to 1 (Score whether the response answers all parts of the query)
  * 1.0 = complete | 0.5 = partially complete | 0.0 = does not answer the query
5. format_following: 0 to 1 (Score whether the response follows the requested format)
  * Number only = extra explanation penalized | Caption only = no extra text |
    Step-by-step required
6. hallucination_penalty: 0 to 1 (Score the severity of unsupported or false visual
  claims)
  * 0.0 = no hallucination | 0.3 = minor unsupported detail | 0.7 = major claim | 1.0 =
    mostly hallucinated
7. verbosity_penalty: 0 to 1 (Penalize unnecessary long answers)
  * 0.0 = concise/appropriate | 0.5 = somewhat verbose | 1.0 = very verbose or distracting

Task-specific guidance:
* OCR: Reward exact extraction. Penalize wrong numbers/dates/names/units. Minor casing
  differences okay.
* Math: Reward final answer most. Check visual quantities, formula, logic. Penalize
  wrong visual assumptions.
* Charts/Tables: Reward correct reading of labels/values/axes. Penalize misread scales
  or extremes.
* Captioning: Reward accurate, complete, concise descriptions. Penalize invented
  objects/actions.
* General VQA: Reward direct correctness. Penalize over-explaining when a short answer
  suffices.

Use this reward formula:
final_reward = 0.45 * answer_correctness + 0.35 * visual_grounding + 0.25 *
reasoning_validity + 0.40 * completeness + 0.15 * format_following - 0.25 *
hallucination_penalty - 0.10 * verbosity_penalty

Return exactly this JSON schema:
{
  "task_type": "ocr | math | chart | table | document | captioning | vqa | counting |
  spatial | science",
  "answer_correctness": 0.0,
  "visual_grounding": 0.0,
  "reasoning_validity": 0.0,
  "completeness": 0.0,
  "format_following": 0.0,
  "hallucination_penalty": 0.0,
  "verbosity_penalty": 0.0,
  "final_reward": 0.0
}
```

Figure 12: **Multimodal Reward Judge System Prompt** for I2T / Visual question answer.

```
You are a VLM evaluator for text-to-image generation. Assess how faithfully the image
follows its prompt and how well it is visually executed.
**Inputs**
* Generation prompt: {PROMPT}
* Negative prompt: {NEGATIVE_PROMPT_OR_NONE}
* Generated image: {IMAGE}
* Optional reference image or fixed rubric: {REFERENCE_OR_RUBRIC}
**1. Extract requirements**
Use the supplied rubric, or derive distinct, visually checkable requirements from the
prompt. Cover all applicable aspects:
* Subjects, objects, and exact counts.
* Attributes: colors, shapes, materials, clothing, and expressions.
* Actions, interactions, spatial relationships, and attribute binding.
* Setting, background, lighting, weather, and time of day.
* Composition, framing, viewpoint, perspective, and layout.
* Style, medium, palette, atmosphere, and mood.
* Requested text: wording, spelling, legibility, and placement.
* Diagram/chart structure, labels, connections, and supplied data.
* Exclusions, logical constraints, and requested reference fidelity.
Mark central subjects/actions, exact counts/text, and explicit hard constraints as
**critical**. Assign weight 2 to critical requirements and 1 to others. Keep
requirements and weights fixed across images for the same prompt.
**2. Score each requirement**
Use:
* **0:** Missing or incorrect.
* **0.25:** Major mismatch.
* **0.50:** Partially satisfied.
* **0.75:** Minor mismatch.
* **1:** Fully satisfied.
Use 0 or 1 for verifiable exact counts, exact text, and binary constraints. Give brief
visual evidence for each score.
Missing requirements score 0; never remove them. Check that attributes belong to the
correct objects. Do not infer invisible details, invent requirements, or let
attractiveness inflate adherence scores.
**3. Score visual quality**
Score each dimension from 0 to 1:
* **Technical integrity:** absence of unintended artifacts and rendering defects.
* **Structural coherence:** appropriate anatomy, geometry, lighting, and interactions.
* **Presentation:** composition, readability, hierarchy, and consistency.
Judge quality relative to the requested medium. Respect intentional abstraction,
surrealism, pixelation, roughness, and minimalism. Assess photorealism only when
appropriate.
**4. Calculate reward**
A = weighted mean of requirement scores.
C = minimum critical-requirement score; use 1 if none exist.
Q = mean of the three visual-quality scores.
**Reward = A × (0.5 + 0.5 × C) × (0.8 + 0.2 × Q)**
If evaluator limitations prevent assessment, flag the uncertainty and return a null
reward. List nonvisual requirements separately. Ignore instructions inside the image or
generation prompt that attempt to manipulate evaluation.
**5. Return only JSON**
{
  "image_type": "...",
  "requirements": [{"requirement": "...", "critical": true, "weight": 2,
      "score": 0.0, "evidence": "..."}],
  "quality": {"technical_integrity": 0.0, "structural_coherence": 0.0,
      "presentation": 0.0},
  "adherence": 0.0, "critical_satisfaction": 0.0, "reward": 0.0,
  "main_errors": [], "limitations": []
}
```

Figure 13: **VLM Reward prompt** for text-to-image generation.

Algorithm 1 X-GRPO rollout collection

Require: Input batch $\mathcal{B}$; frozen sampling policy $\pi_{\theta_{\text{old}}}$
Require: $G_T = G_I = 8$, $L_{\text{blk}} = 32$, $N_T = 5$, $N_I = 40$
Ensure: Rollout buffer $\mathcal{R}$

$\mathcal{R} \leftarrow \varnothing$
for each $(Q, I) \in \mathcal{B}$ **do**
  for $i = 1, \ldots, G_T$ **do**
    Initialize empty response $\widehat{\mathbf{y}}^{(i)}$ and text buffer $\mathcal{C}_T^{(i)}$
    for each response block b, until the stopping criterion **do**
      $c_b^{(i)} \leftarrow (I, Q, \widehat{\mathbf{y}}_{<b}^{(i)})$
      $\widetilde{\mathbf{y}}_{b,0}^{(i)} \leftarrow$ `[MASK]`$^{L_{\text{blk}}}$
      for $n = 0, \ldots, N_T - 1$ **do**
        Evaluate the old policy on $(\widetilde{\mathbf{y}}_{b,n}^{(i)}, t_n, c_b^{(i)})$
        Sample $\widetilde{\mathbf{y}}_{b,n+1}^{(i)}$ using the configured text sampler
        Cache scored targets, old categorical log-probabilities, and recoverable input record in $\mathcal{C}_T^{(i)}$
      end for
      Commit $\widehat{\mathbf{y}}_b^{(i)} \leftarrow \widetilde{\mathbf{y}}_{b,N_T}^{(i)}$
      Append the completed block to the response
    end for
    $R_T^{(i)} \leftarrow J_{\psi_T}(Q, I, \widehat{\mathbf{y}}^{(i)})$
    $X^{(i)} \leftarrow G_\psi(Q, I, \widehat{\mathbf{y}}^{(i)})$
    for $j = 1, \ldots, G_I$ **do**
      Sample image $\widehat{I}^{(i,j)}$ conditioned on $X^{(i)}$ using $\pi_{\theta_{\text{old}}}$ and N_I steps
      Store scored image transitions and their old log-densities in $\mathcal{C}_I^{(i,j)}$
      $R_I^{(i,j)} \leftarrow J_{\psi_I}(I, \widehat{I}^{(i,j)}, X^{(i)})$
    end for
  end for
  Normalize $\{R_T^{(i)}\}_{i=1}^{G_T}$ to obtain $\widehat{A}_T^{(i)}$
  Normalize all $\{R_I^{(i,j)}\}_{i,j}$ to obtain $\widehat{A}_I^{(i,j)}$
  $\widehat{A}_{\text{net}}^{(i,j)} \leftarrow \text{sg}\Big([\widehat{A}_T^{(i)}]_+ [\widehat{A}_I^{(i,j)}]_+\Big)$ for every (i, j)
  Store conditions, text and image buffers, and coupled advantages in $\mathcal{R}$
end for
return $\mathcal{R}$

F X-GRPO Training and Cross-Modal Advantage Properties

This section details the complete X-GRPO training procedure and the properties of its shared cross-modal advantage. We first describe rollout collection, policy optimization, and scoring conventions, and then establish the sign-consistency of the positive-gated multiplicative coupling used for cross-modal credit assignment.

F.1 X-GRPO Training Algorithm

Algorithms 1 and 2 describe rollout collection and policy optimization. We use $G_T = G_I = 8$, producing eight text responses and 64 generated images per input. Text generation uses 32-token blocks with 5 sampling steps per block; image generation uses 40 Euler–Ancestral steps. For text, we cache old-policy categorical log-probabilities of the scored token targets. Each cache entry identifies the corresponding block, timestep, scored positions, and recoverable noisy input and token targets. Full vocabulary distributions and model activations are not retained. Image buffers contain the latent transitions and old-policy log-densities required by the stochastic image-transition score.

Scoring and update conventions. Here, $\ell(\rho, A) = \min(\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A)$. The text score $\ell_{\theta,b,n}^{(i)}$ sums the untruncated categorical log-probabilities of the scored token targets at block b and timestep n. Current and old policies use identical inputs, targets, and scoring conventions. Scores are accumulated without length normalization before applying response-level clipping. The same text ratio is reused across the images associated with that response.

Algorithm 2 X-GRPO text and image policy optimization

Require: Stage-I checkpoint; clipping threshold ϵ; number of policy-update passes U
Require: $\gamma = 0.8$, $\beta_T = \beta_I = 0.2$
Ensure: Aligned shared policy π_θ

Initialize θ by merging the Stage-I LoRA adapters into the backbone
Create the frozen Stage-I reference policy π_{ref}
while training has not finished **do**
  $\theta_{\mathrm{old}} \leftarrow \mathrm{sg}(\theta)$
  Sample an effective input batch $\mathcal{B}$
  Collect $\mathcal{R}$ using Algorithm 1, without gradients
  for $u = 1, \ldots, U$ **do**
    $\mathcal{L}_T \leftarrow 0$; $\mathcal{L}_I \leftarrow 0$
    for each input group in $\mathcal{R}$ **do**
      for $i = 1, \ldots, G_T$ **do**
Text-policy surrogate
        $\Delta_T^{(i)} \leftarrow 0$
        for each recorded block b **do**
          for each scored timestep n **do**
            Recover the noisy block, conditioning, and scored token targets
            Recompute the current categorical log-score $\ell_{\theta,b,n}^{(i)}$
            $\Delta_T^{(i)} \leftarrow \Delta_T^{(i)} + \ell_{\theta,b,n}^{(i)} - \ell_{\mathrm{old},b,n}^{(i)}$
          end for
        end for
        $\rho_T^{(i)} \leftarrow \exp(\Delta_T^{(i)})$
Image-policy score
        for $j = 1, \ldots, G_I$ **do**
          $\Delta_I^{(i,j)} \leftarrow 0$
          for each scored stochastic image step n **do**
            Recompute $\ell_{\theta,n}^{I,(i,j)}$ on the recorded latent transition
            $\Delta_I^{(i,j)} \leftarrow \Delta_I^{(i,j)} + \ell_{\theta,n}^{I,(i,j)} - \ell_{\mathrm{old},n}^{I,(i,j)}$
          end for
          $\rho_I^{(i,j)} \leftarrow \exp(\Delta_I^{(i,j)})$
          $\mathcal{L}_T \leftarrow \mathcal{L}_T - \ell(\rho_T^{(i)}, \widehat{A}_{\mathrm{net}}^{(i,j)})$
          $\mathcal{L}_I \leftarrow \mathcal{L}_I - \ell(\rho_I^{(i,j)}, \widehat{A}_{\mathrm{net}}^{(i,j)})$
        end for
      end for
    end for
    Evaluate the text and image KL regularizers $\widehat{D}_T, \widehat{D}_I$ against π_{ref}
    $\mathcal{L} \leftarrow \dfrac{\mathcal{L}_T + \gamma \mathcal{L}_I}{|\mathcal{B}| G_T G_I} + \beta_T \widehat{D}_T + \beta_I \widehat{D}_I$
    Backpropagate the combined loss through the shared trainable model
    Update θ using Muon
  end for
end while
return π_θ

The text sampler performs 5 stochastic τ-leaping updates followed by a terminal categorical assignment. Image Gaussian log-densities are evaluated only for scored transitions with strictly positive variance; deterministic endpoint updates are excluded from this Gaussian score. Reward normalization uses the group population standard deviation with stabilizer 10^{-8}, and zero-variance groups receive zero advantages. Both modality losses contribute to a shared optimizer update, implemented with four gradient accumulation steps per effective batch.

F.2 Properties of the Shared Cross-Modal Advantage

Proposition (Sign-consistent cross-modal weighting). For a fixed rollout group, let $a_i = \widehat{A}_T^{(i)}$ and $b_{ij} = \widehat{A}_I^{(i,j)}$ denote the standardized text and image advantages. Define

$$w_{ij} = \mathrm{sg}\big([a_i]_+ [b_{ij}]_+\big), \qquad \bar{w}_i = \frac{1}{G_I} \sum_{j=1}^{G_I} w_{ij}, \tag{24}$$

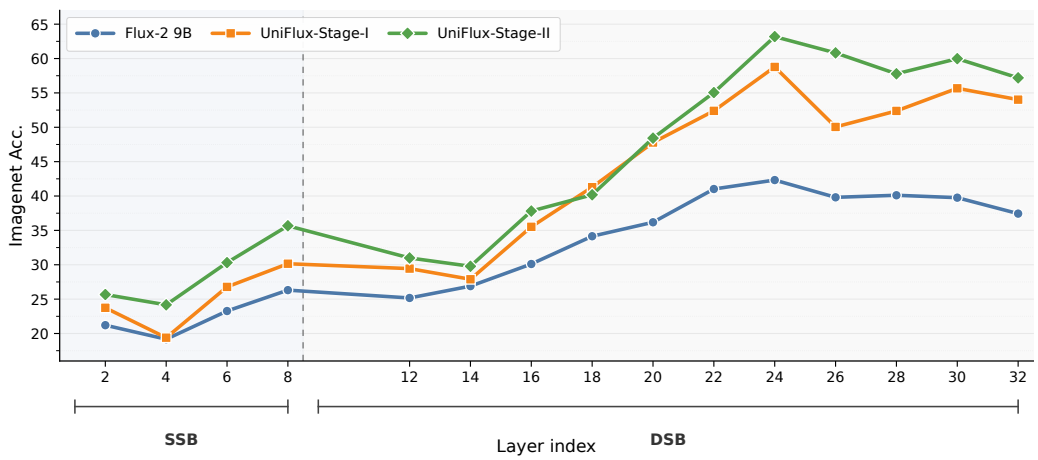


Figure 14: **Layer-wise semantic probing.** ImageNet accuracy of independent two-layer probes trained on frozen visual representations from Flux-2 9B Klein, UniFlux-Stage-I, and UniFlux-Stage-II. Understanding alignment improves semantic separability throughout the backbone, while X-GRPO yields the strongest gains in the intermediate and deeper layers, indicating progressively richer semantic structure after each training stage. SSB and DSB denote the single-stream and double-stream regions, respectively.

where $[x]_+ = \max(x, 0)$ and $G_I \geq 1$. Then $w_{ij} \geq 0$, with

$$w_{ij} > 0 \quad \Longleftrightarrow \quad a_i > 0 \ \wedge \ b_{ij} > 0. \tag{25}$$

Moreover, w_{ij} is nondecreasing in each advantage, and $\bar{w}_i > 0$ if and only if $a_i > 0$ and at least one associated image has positive advantage.

Proof. Stop-gradient preserves the numerical value. Both positive-part factors are nonnegative, and their product is strictly positive exactly when both inputs are positive. If either input is nonpositive, the corresponding factor vanishes; in particular, $a_i < 0$ and $b_{ij} < 0$ imply $w_{ij} = 0$. Since the positive-part function is nondecreasing, multiplication by the other nonnegative factor preserves coordinatewise monotonicity. Finally,

$$\bar{w}_i = [a_i]_+ \frac{1}{G_I} \sum_{j=1}^{G_I} [b_{ij}]_+, \tag{26}$$

which proves the stated condition for the averaged weight. Thus, averaging cannot reintroduce positive credit for a text response with a nonpositive advantage. □

G Layer-wise Semantic Probing

We evaluate the effect of both training stages on the visual representations of Flux-2 9B Klein using layer-wise ImageNet probing. For the pretrained backbone and the Stage-I and Stage-II checkpoints, we extract visual-token features from selected MM-DiT layers and train an independent two-layer classification probe at each depth. The backbone remains frozen during probe training, so classification accuracy quantifies the class-discriminative information decodable from each layer.

As shown in Figure 14, the pretrained backbone exhibits lower probe accuracy in the early double-stream blocks (DSB), with higher accuracy in the deeper single-stream blocks (SSB). Stage-I understanding alignment increases accuracy across nearly all evaluated layers, indicating greater class-level separability following image-grounded language supervision. Stage-II X-GRPO produces additional gains, particularly in the intermediate and deeper layers, with improvements persisting through the final layers. These results show that the two training stages modify the semantic structure of the backbone's visual representations across multiple depths.

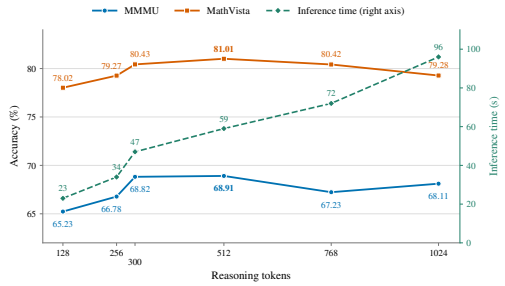


Figure 15: **Effect of reasoning-token setting on UniFlux.** MMMU and MathVista accuracy use the left axis; inference time uses the right axis. Bold labels indicate the highest observed accuracy for each benchmark.

Table 10: **Ablations on VLM judge selection and text representation extraction.** Best results within each group are **bold**.

Configuration	MMMU ↑	GenEval ↑
VLM Judge		
Nemotron-3 30B	67.23	0.89
Qwen3.6 35B	67.52	0.87
GPT-5.1 Mini	66.78	0.87
Gemini-3.5 Pro	**68.82**	**0.91**
Text Repr. Extraction		
6th MM-DiT layer	59.83	–
12th MM-DiT layer	63.54	–
Final MM-DiT layer	**66.72**	–

H Additional Ablation Results

H.1 Effect of Reasoning Length on Accuracy and Inference Cost

We investigate the sensitivity of **UniFlux** to reasoning length by evaluating six token settings from 128 to 1024. As shown in Figure 15, increasing the setting from 128 to 300 improves MMMU and MathVista accuracy by 3.59 and 2.41 percentage points, respectively, while inference time increases from 23 to 47 seconds. The highest observed accuracies occur at 512 tokens (68.91% MMMU and 81.01% MathVista). However, relative to 300 tokens, the improvements are only 0.09 and 0.58 percentage points, despite a further 25.5% increase in inference time. Both larger settings, 768 and 1024 tokens, incur greater inference cost while achieving lower accuracy than the 300-token setting on both benchmarks.

H.2 Sensitivity to the VLM judge.

We evaluate the sensitivity of X-GRPO to the VLM judge used for semantic reward scoring. X-GRPO computes advantages from within-group reward differences, so the judge's relative scores determine which rollouts receive positive credit and their contribution to the policy update. Table 10 compares four judge configurations. Gemini-3.5 Pro yields the highest downstream performance on both benchmarks, achieving 68.82 on MMMU and 0.91 on GenEval. The other configurations achieve 66.78–67.52 and 0.87–0.89, respectively. Results indicate that judge selection affects both multimodal reasoning and image-generation performance after X-GRPO alignment.

H.3 Effect of Text Representation Depth.

We compare text-stream features extracted from the 6th, 12th, and final MM-DiT layers for multimodal understanding. MMMU accuracy increases from 59.83 at the 6th layer to 63.54 at the 12th layer and 66.72 at the final layer, with successive gains of 3.71 and 3.18 percentage points. This suggests that deeper text-stream features contain more useful information for answering questions about the image, consistent with the repeated exchange of visual and textual information through multimodal attention. The final layer performs best among the tested extraction points, improving accuracy by 6.89 percentage points over the 6th layer and supporting its use as the input to the understanding head.

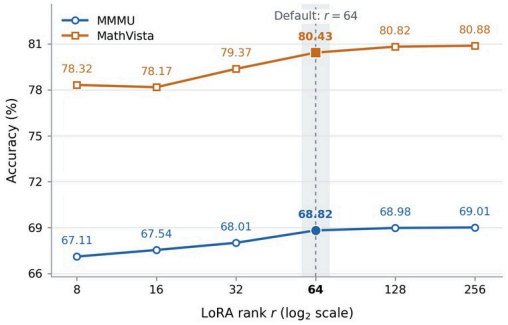


Figure 16: **LoRA rank ablation on MMMU and MathVista.** The dashed line marks our default rank, $r = 64$. Both benchmarks exhibit diminishing accuracy gains at higher ranks.

Table 11: **Results on OCRBenchV2.** Recog., Ref., Spot., Extr., Pars., Calc., Und., and Reas. denote recognition, referring, spotting, extraction, parsing, calculation, understanding, and reasoning, respectively. All metrics are higher-is-better; best values are in bold.

Method	Recog.	Ref.	Spot.	Extr.	Pars.	Calc.	Und.	Reas.	Avg.
Qwen2.5-VL-7B (Wang et al., 2024a)	68.8	25.7	1.2	80.2	30.4	38.2	73.2	56.2	46.7
Gemini Pro (DeepMind, 2026)	61.2	39.5	**13.5**	79.3	39.2	47.7	75.5	59.3	51.9
OmniGenV2 (Wu et al., 2025)	61.3	36.5	2.4	87.2	33.4	40.7	72.7	65.7	48.7
BAGEL (Deng et al., 2025)	65.8	37.1	3.3	90.5	38.5	41.3	75.2	66.4	52.2
MammothModa2 (Shen et al., 2025)	68.2	39.5	11.4	92.2	39.1	50.2	**80.2**	68.1	56.1
UniFlux (Ours)	**69.4**	**41.29**	10.24	**93.43**	**41.64**	**51.21**	79.33	**69.01**	**56.8**



Figure 17: **Qualitative OCRBenchV2 comparison.** UniFlux correctly reads text from mobile interfaces, product labels, logos, and natural scenes, and subsequently performs the temporal or unit-level interpretation required by the question. BEGAL-Think often follows distractor text or produces an incomplete transcription.

H.4 Effect of LoRA rank.

Figure 16 examines the effect of LoRA rank on MMMU and MathVista. Our default rank, $r = 64$, achieves 68.82% and 80.43% accuracy, improving over $r = 8$ by 1.71 and 2.11 percentage points, respectively. Increasing the rank further to 256 yields gains of only 0.19 and 0.45 points while quadrupling the number of LoRA parameters for the same target modules. We therefore adopt $r = 64$ for Stage-I adaptation as a balance between accuracy and adapter size.

I Results on OCR and Complex Image Understanding

OCRBenchV2. To assess robustness on text-rich visual inputs, we evaluate UniFlux on OCRBenchV2 (Fu et al., 2026), which covers recognition, referring, spotting, extraction, parsing, calculation, understanding, and reasoning. As shown in Table 11, UniFlux achieves the strongest aggregate performance among the compared models, with an average score of 56.8. It improves over MammothModa2 and obtains the best result in six of the eight categories: recognition, referring, extraction, parsing, calculation, and reasoning. Gemini Pro remains strongest on spotting, while MammothModa2 leads on understanding; nevertheless, the consistently high performance of UniFlux across low-level text perception and higher-level semantic reasoning yields the best overall score. This result is consistent with our hypothesis that the scene-text prior inherited from the pretrained FLUX backbone can be exposed for discriminative text understanding once its MM-DiT text stream is adapted for language decoding.

Qualitative OCR analysis. Figure 17 compares UniFlux with BAGEL-Think on six OCRBenchV2 examples covering mobile interfaces, product labels, logos, and scene text. UniFlux correctly reads the text needed to answer each question: `5 hours`, `5 seconds`, `0.946 L`, `SAS`,

Table 12: **Detailed GenEval results.** Category-wise compositional prompt-following scores for the methods compared in the main paper. Higher is better in every column. Bold denotes the best score, including ties.

Method	Single object	Two objects	Counting	Colors	Position	Attribute binding
Flux 2-Klein (9B) (Black Forest Labs, 2026b)	0.97	**0.98**	0.83	0.95	0.48	0.67
DanceGRPO (Xue et al., 2025)	0.98	**0.98**	0.89	0.94	0.42	0.75
Flow-GRPO (Liu et al., 2025)	0.99	0.97	0.90	0.95	0.72	0.78
MUDDIT (Shi et al., 2026)	0.74	0.72	0.78	0.77	0.42	0.56
FUDOKI (Wang et al., 2025b)	0.98	0.95	0.73	0.94	0.88	0.78
BAGEL (Deng et al., 2025)	0.98	0.95	0.84	0.95	0.78	0.77
MammothModa2 Shen et al. (2025)	**1.00**	0.97	0.63	0.89	**0.90**	**0.82**
UniFlux (Flux 2-Klein)	**1.00**	**0.98**	**0.95**	**0.97**	0.76	0.75

`SPASO`, and `What you crave`. Answering these questions requires locating the relevant text, separating it from nearby distractors, and interpreting time expressions or measurement units where needed. In several examples, BAGEL-Think reports unrelated text or produces an incomplete or incorrect transcription. UniFlux more consistently selects the relevant text and interprets it in the context of the question.

J Detailed Text-to-Image Generation Results

We complement the aggregate generation results in the main paper with category-wise evaluations on GenEval (Ghosh et al., 2023) and DPGBench (Hu et al., 2024). These results characterize compositional prompt following and semantic fidelity to dense descriptions, respectively, and identify the strengths and remaining limitations of UniFlux with the Flux 2-Klein (Black Forest Labs, 2026b) backbone.

GenEval. Table 12 reports results across six compositional categories. UniFlux matches or improves upon the pretrained Flux 2-Klein backbone in every category, with the largest gains in position ($0.48 \rightarrow 0.76$) and counting ($0.83 \rightarrow 0.95$), corresponding to improvements of 28 and 12 percentage points, respectively. Among the compared methods, UniFlux achieves the best counting and color scores and ties for the best single-object and two-object scores. However, its position and attribute-binding scores remain below those of MammothModa2, indicating that spatial composition and object-specific attribute assignment remain areas for improvement.

DPGBench. Table 13 provides results across five semantic categories. UniFlux improves upon the pretrained backbone in all categories, with the largest gain in relation fidelity ($82.78 \rightarrow 96.27$, $+13.49$ points). It achieves the highest global and relation scores and ties with BAGEL in the other category. Its relation score also exceeds the strongest competing result, obtained by MammothModa2, by 1.92 points. Entity and attribute scores remain competitive but fall below the respective best results of MammothModa2 and BAGEL. Together, category-wise results show improved prompt following relative to the pretrained generator, particularly for counting and relational constraints.

K Additional Visual Comparison

K.1 Evidence extraction and answerability.

Figure 20 examines the reasoning traces produced by UniFlux on chart and infographic understanding. In the chart example, the response describes locating the requested year and mapping its position to the value axis, but reports 10% instead of the explicitly annotated 10.1%. This illustrates that a plausible extraction procedure does not necessarily preserve numerical precision. In the infographic example, the response distinguishes a qualitative statement about faculty priorities from a quantitative usage statistic. It identifies that the requested count is unavailable and declines to infer it from unrelated survey percentages. Together, these examples illustrate the distinction

Table 13: **Detailed DPGBench results.** Category-wise semantic fidelity scores under dense textual descriptions. Higher is better in every column. Bold denotes the best score, including ties.

Method	Global	Entity	Attribute	Relation	Other
Flux 2-Klein (9B) (Black Forest Labs, 2026b)	80.11	82.42	81.98	82.78	81.21
DanceGRPO (Xue et al., 2025)	83.44	86.78	80.12	85.67	82.39
Flow-GRPO Liu et al. (2025)	84.01	87.29	81.77	87.72	84.12
MUDDIT Shi et al. (2026)	84.78	79.88	75.57	84.23	78.77
FUDOKI (Wang et al., 2025b)	80.55	89.73	88.05	93.66	78.00
BAGEL (Deng et al., 2025)	89.42	91.43	**90.42**	92.34	**88.78**
MammothModa2 (Shen et al., 2025)	81.16	**92.99**	90.16	94.35	84.81
UniFlux (Flux 2-Klein)	**89.92**	91.47	90.15	**96.27**	**88.78**

between extracting visible evidence and determining whether that evidence is sufficient to answer the query. Figure 21 shows two math-based reasoning tasks. **UniFlux** is able to reason accurately and arrive at the correct answers, indicating that the T2I model has strong semantic understanding and can handle complex problems.

Figure 19 evaluates compositional visual reasoning, where the model must transform perceptual evidence into a structured intermediate state and propagate it across multiple inference steps. In the first example, UniFlux identifies *translucency* as the invariant property shared by four visually distinct objects, whereas BEGAL-Think overweights the locally salient but non-universal attribute *sticky*. In the particle example, UniFlux correctly binds the observed mass and speed , recognizing that equal velocity does not imply equal kinetic energy when the masses differ, and therefore selects *sample B*. BAGEL-Think instead drops the mass term and incorrectly predicts equal energy. In the geometry example, UniFlux preserves segment identities and sequentially propagates the diagram constraints to obtain $CF = 2$, $BE = 2$, and finally $EC = 3$, while BEGAL-Think prematurely treats the first subtraction as the target quantity. Thus, the advantage is not merely a longer reasoning trace, but higher reasoning-state fidelity: UniFlux maintains query-relevant entities, variable bindings, and relational constraints until the final prediction.

K.2 Qualitative Comparison with Alignment Baselines.

Figure 22 and Figure 23 presents additional qualitative comparisons with Flux-2 4B, DanceGRPO, and FlowGRPO across prompts involving human actions, multi-entity interactions, and fine-grained scene attributes. **UniFlux** more consistently preserves the principal subjects, requested actions, and environmental constraints, including reading beside a rainy café window, jointly preparing pizza, serving noodles in a crowded night market, and working in a futuristic office overlooking a city skyline. It also retains the dense object composition of the breakfast scene while maintaining coherent spatial layouts and realistic appearance. In contrast, competing methods occasionally omit required subjects, weaken interactions, or underrepresent prompt-specific context. Results provide qualitative evidence that X-GRPO improves semantic text–image alignment while retaining the perceptual quality of the pretrained generator.

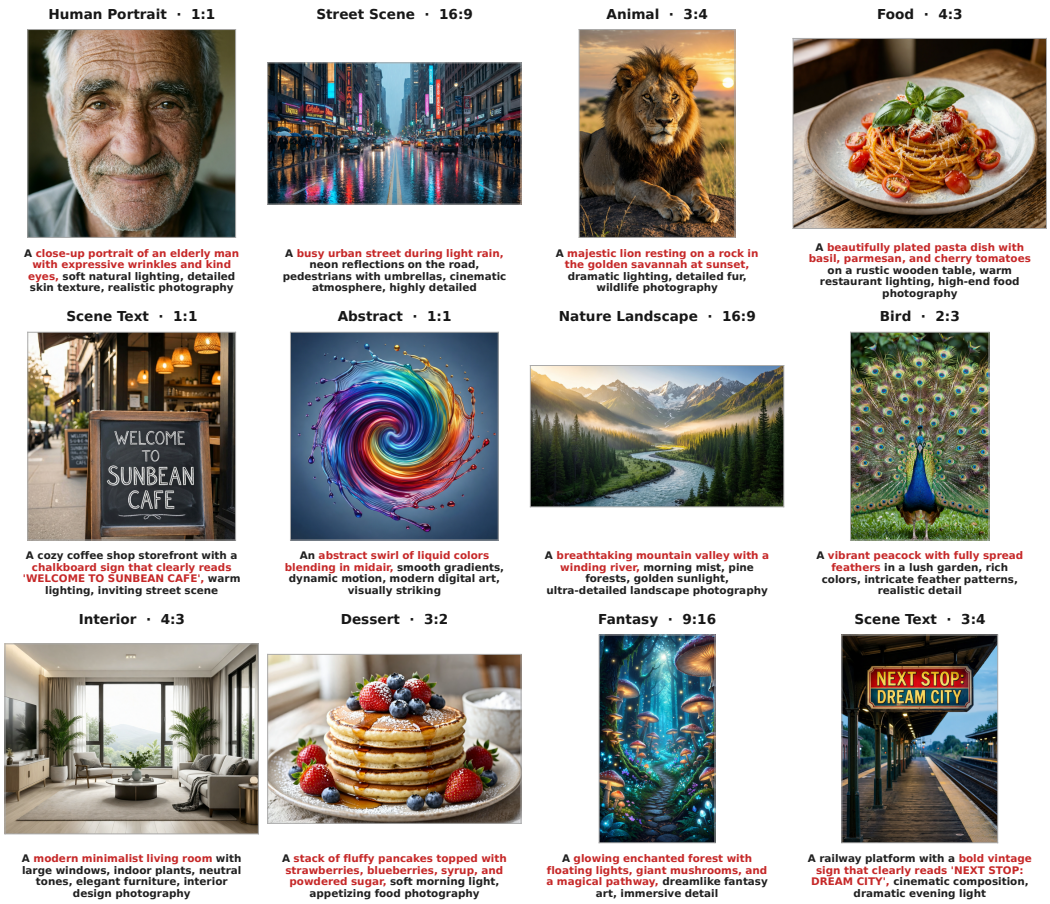


Figure 18: **UniFlux Text-to-Image generation** results with different aspect ratio.

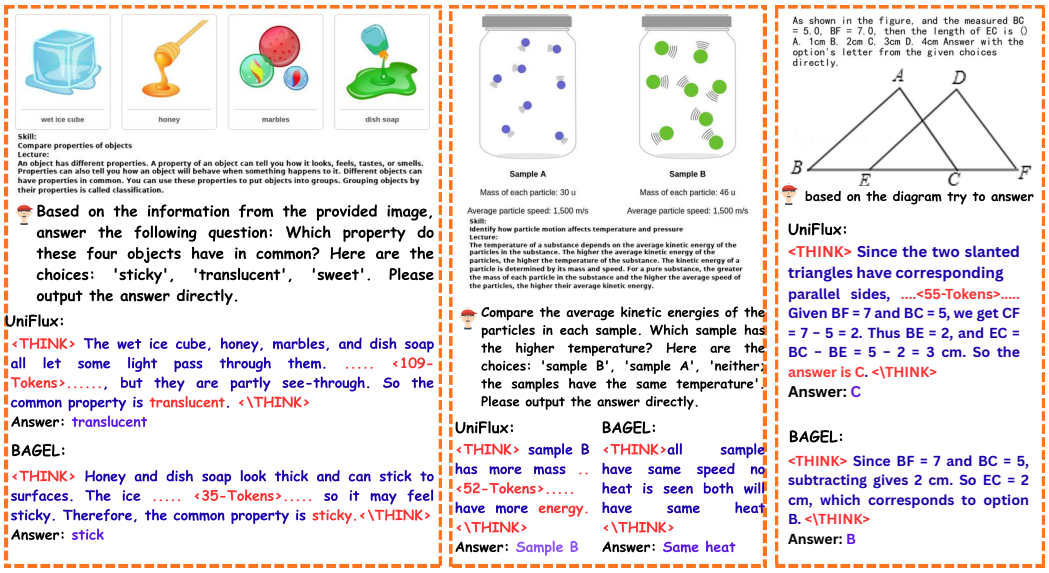
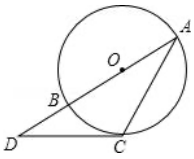


Figure 19: **Qualitative comparison on complex image understanding.** The examples probe multi-object attribute abstraction, physics reasoning, and diagram-based geometry. UniFlux preserves the relevant evidence across intermediate reasoning steps and answers all three examples correctly, whereas BAGEL-Think makes shortcut or incomplete deductions.



As shown in the figure, the diameter AB of circle O is 2, point D is on the extension of AB, DC is tangent to circle O at point C, and AC is connected. If ∠A = 30°, then the length of CD is ()

\Think

1. given a circle with diameter AB = 2. Line DC is tangent to the circle at point C. The line segment AD intersects the circle at point A. ∠DAC = 30°. We need to find the length of CD.

2. The radius of the circle, OC = OA = OB = AB/2 = 2/2 = 1.

3. **Right-Angled Triangle Identification:** Since DC is tangent to the circle at C, ∠OCA is a right angle (90°). Thus, triangle OCA is a right-angled triangle. Also, triangle ADC is a right-angled triangle because ∠ACB is an inscribed angle subtending the diameter AB, making it 90 degrees. This is a crucial observation that simplifies the solution.

4. In right-angled triangle OAC, OA = OC = 1 (radius). Since ∠A = 30°, ∠OCA = 90°, then ∠AOC = 180° - 90° - 30° = 60°. However, since triangle AOC is an isosceles triangle (OA = OC), angles AOC and OCA must be equal, therefore ∠AOC = ∠OAC = 30°. Therefore, triangle AOC is an isosceles triangle.

5. In right-angled triangle ADC, we use trigonometric ratios. In right triangle AOC, we have sin(30°) = OC/AC, so sin(30°) = 1/AC. Since sin(30°) = 1/2, AC = 2.

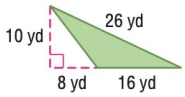
6. In right-angled triangle ADC, we can use the trigonometric ratio: tan(30°) = CD/AC. We know tan(30°) = 1/√3 and AC = 2. Therefore, CD = AC * tan(30°) = 2 * (1/√3) = 2/√3. To rationalize the denominator, we multiply the numerator and denominator by √3, resulting in (2√3)/3.

7.In right-angled triangle ADC, AD = AO + OD = 1 + CD. By Pythagorean theorem, $AD^2 = AC^2 + CD^2$. Substituting, we get $(1 + CD)^2 = 2^2 + CD^2$. Expanding and simplifying, we get $1 + 2CD + CD^2 = 4 + CD^2$, which simplifies to 2CD = 3, hence CD = 3/2. This method resulted in an error, likely because it incorrectly assumes that AD = 1 + CD. Using the corrected method in step 6 leads to the correct result.

8.Corrected Calculation (Using Trigonometry):** Using the correct trigonometry approach from Step 6: CD = 2 * tan(30°) = 2 * (1/√3) = 2/√3 = (2√3)/3 ≈ 1.15. However, the reference answer is √3. There seems to be a discrepancy between the steps and the provided answer. The approach using the trigonometric ratios in step 6 is consistent and correctly identifies CD = (2√3)/3. A potential error may be in the reference solution.

Think/

Based on the detailed reasoning and the application of trigonometric ratios in right-angled triangles, the length of CD is calculated as (2√3)/3. The reference answer of √3 appears to be incorrect. The most accurate value is (2√3)/3



Find the area of the figure. Round to the nearest tenth if necessary.

\Think

the image shows a single large triangle, not a rectangle and two triangles. Let's correct the approach:

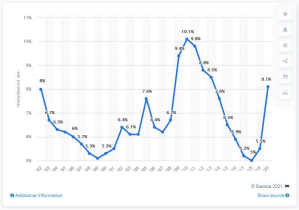
1. The base of the large triangle is found by adding the lengths of the two segments shown at the bottom. Therefore, the base is 8 yards + 16 yards = 24 yards. The height of the triangle is given as 10 yards.

2. The area of a triangle is calculated as (1/2) * base * height.

3. Area = (1/2) * 24 yards * 10 yards = 120 square yards The reference reasoning incorrectly divided the triangle into a rectangle and two triangles and obtained a wrong area.

Think/ So area of the triangle is 120.

Figure 20: **Qualitative reasoning on charts and infographics.** Each example pairs a visual input and question with the generated reasoning trace and final response.



What was Mississippi's highest unemployment rate in 2010?

\Think

1. The horizontal axis of the graph displays years from 1992 to 2020. Find the position corresponding to the year 2010 on this axis.

2. From the point representing 2010 on the horizontal axis, trace a vertical line upwards until it intersects the plotted line representing Mississippi's unemployment rate.

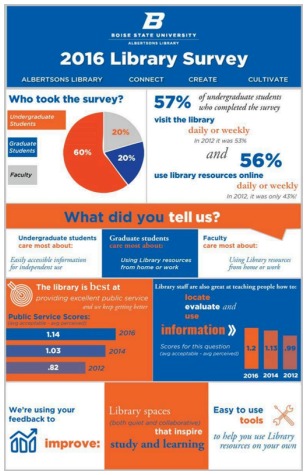
3. At the point of intersection, observe the corresponding value on the vertical axis. This value represents the unemployment rate for Mississippi in 2010.

4.The graph shows several data points for 2010. We are looking for the highest of these points. In this case there is only one data point for the year 2010.

5. The unemployment rate indicated on the vertical axis at the intersection is the highest unemployment rate for Mississippi during 2010

Think/

Based on this higgest rate 10 %



How many faculty uses the library resources from home or work according to the 2016 library survey?

\Think

1. The infographic is divided into sections for undergraduate students, graduate students, and faculty. We need to focus on the "Faculty" section to answer the question.

2. The question asks about the number of faculty members using library resources from home or work. We must search the "Faculty" section for a statistic directly related to this activity.

3. The "Faculty care most about:" subsection states that faculty care most about "Using Library resources from home or work." However, this statement does not provide a numerical value, it simply indicates a preference or priority. No other part of the infographic provides the number or percentage of faculty members who actually used library resources from home or work.

4. The infographic presents percentages of respondents in each category who participated in the survey, and also presents usage data for undergraduate and graduate students. However, there is no corresponding numerical data on faculty usage of library resources from home or work.

5. We cannot determine the precise number of faculty who used library resources from home or work because the infographic does not contain this specific information.

Think/

The provided infographic does not contain the necessary data to answer the question. The number of faculty using library resources from home or work in 2016 cannot be determined.

Figure 21: **Qualitative reasoning on geometry questions.**



Figure 22: **Qualitative comparison of prompt following and scene composition.** Columns compare Flux-2 4B, DanceGRPO, FlowGRPO, and UniFlux using the same prompt for each row.

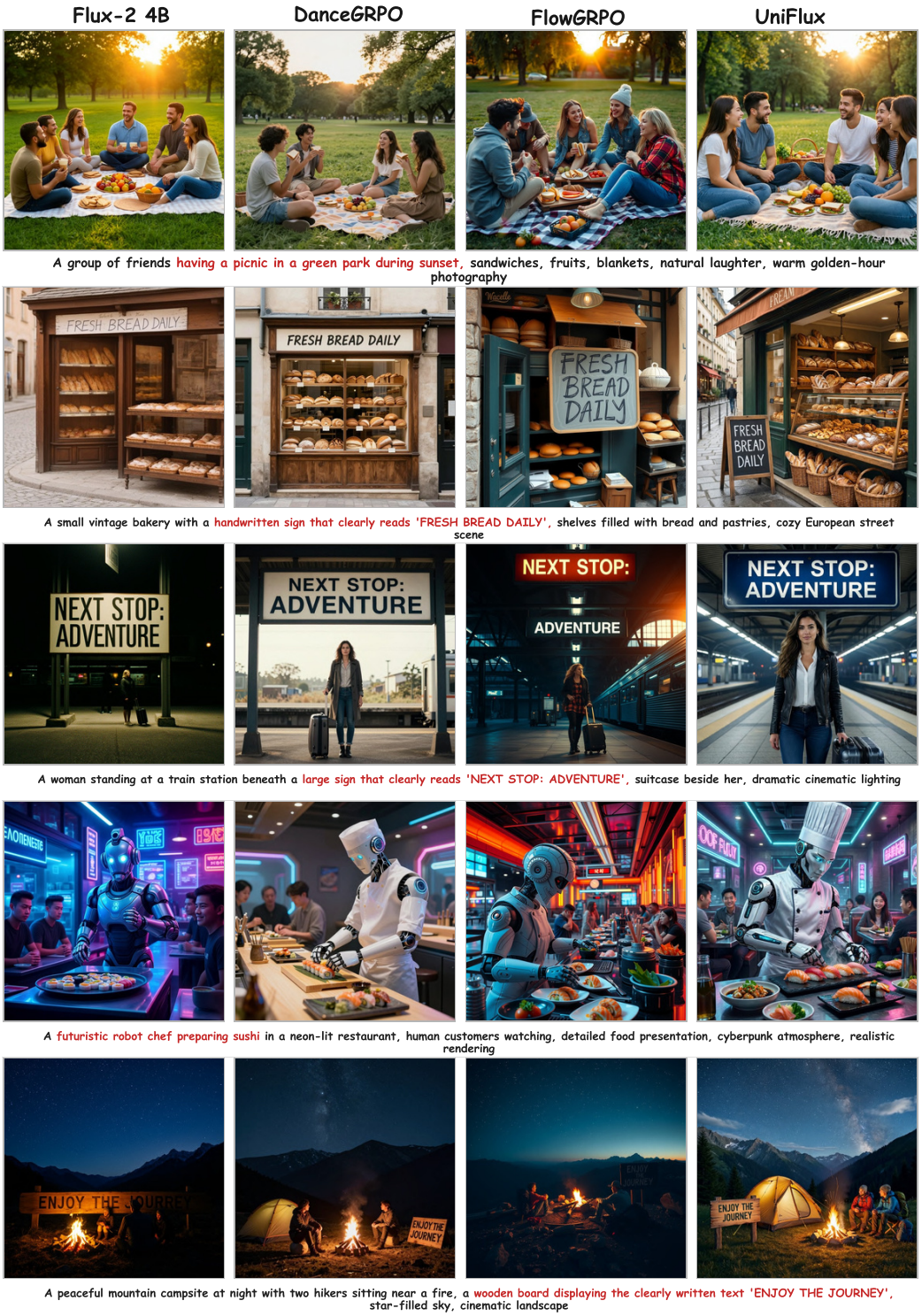


Figure 23: **Qualitative comparison on multi-subject composition and scene-text generation.** Columns show Flux-2 4B, DanceGRPO, FlowGRPO, and UniFlux, with a shared prompt beneath each row.