RELAYING UNDERSTANDING TO GENERATION IN UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models bring visual understanding and generation into a single architecture, yet sharing a backbone alone does not guarantee that knowledge acquired for understanding can be effectively utilized by the generator. A central challenge is therefore not only what the model understands, but also how understanding-side representations are exposed in a form that the generation pathway can readily consume. We address this problem by introducing an explicit understanding-to-generation interface into the topology of unified multimodal models. Specifically, understanding- and generation-oriented computation are separated, while a compact set of learnable relay tokens is inserted between them. The understanding pathway transforms input semantics into these intermediate states, and the generation pathway is constrained to condition on them rather than directly accessing the original input representation. This topology encourages the understanding pathway to organize semantic information into a generation-compatible representation, while providing the generator with a compact and structured conditioning interface. Despite its simplicity, the proposed design consistently improves image generation, with particularly strong gains on compositional semantics such as spatial relations, attribute binding, and actions. Our results suggest that explicitly structuring the topology of information flow between understanding and generation is an effective way to improve how unified multimodal models translate semantic understanding into visual generation.

1 INTRODUCTION

Unified multimodal models (UMMs) bring visual understanding and generation into a single framework, enabling one model to interpret multimodal inputs and synthesize visual content from user instructions Pan et al. (2025); Chen et al. (2025a;b;c); Deng et al. (2025a). These capabilities serve complementary roles: understanding extracts and integrates semantic information from the input, while generation translates conditioning signals into visual outputs Shuai et al. (2024); Liu et al. (2026). Effective unification therefore depends on how well the generation process can access and use the semantic representations formed through understanding.

Recent studies have revealed a gap between what UMMs can understand and what they can generate Liu et al. (2026); Niu et al. (2025a). Although these models often follow direct generation instructions, they can struggle when the desired output requires complex semantic composition, relational reasoning, or implicit inference. As illustrated in Fig. 1, the model correctly states that Cinderella's carriage turns into a pumpkin after midnight, yet fails to depict a pumpkin when given the same query as a generation prompt. This discrepancy suggests that relevant knowledge can be available within the model but insufficiently utilized during generation, motivating an explicit interface that makes understanding-derived semantic information more accessible to visual synthesis.

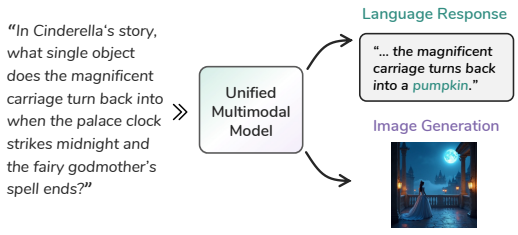

Figure 1: The Bagel correctly answers "pumpkin" but fails to depict it when generating an image from the same query.

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.

To this end, we propose *Relaying Understanding to Generation in Unified Multimodal Models* (**RUG**), a new unified attention topology built around a small set of learnable **Relay Tokens**. In **RUG**, the understanding pathway writes contextual information into these tokens, which attend to the full textual context and interact bidirectionally to integrate information distributed across the instruction. The generation pathway then attends jointly to the original text and the Relay Tokens, combining direct access to the instruction with input-dependent semantic states. Operating within the model's existing computation, Relay Tokens provide a lightweight interface for semantic transfer with negligible additional inference overhead, without requiring an additional encoder, decoder, or separate reasoning module.

This interface is learned using the same training objectives as the underlying UMM, without additional supervision. Extensive experiments across diverse benchmarks demonstrate the strong performance of our approach, while comparisons across different UMM architectures further highlight the advantages of the proposed attention topology. We hope this work offers a new perspective on connecting understanding and generation in UMMs, paving the way for models that more effectively translate semantic understanding into visual synthesis.

Our main contributions are summarized as follows:

- We introduce **RUG**, a unified attention topology for relaying understanding-derived semantic information to generation through learnable **Relay Tokens**.
- We analyze the capabilities of **RUG** through ablation studies and controlled interventions on Relay Tokens, providing insights into how understanding-derived representations guide visual generation.
- We evaluate RUG across different UMM architectures on generation benchmarks, with controlled topology comparisons demonstrating its benefits for visual generation.

2 Related Work

Unified Multimodal Architectures. Architectural unification has been explored through shared Transformers, modality-specific experts, and learned bridges between pretrained models. Chameleon (Team, 2024) and Emu3 (Wang et al., 2024) model text and visual tokens with a shared autoregressive Transformer. Show-o (Xie et al., 2024) combines autoregressive modeling and discrete diffusion within one Transformer, while Janus-Pro (Chen et al., 2025c) retains a shared backbone with separate visual encoders for understanding and generation. Mixture-of-Transformers (MoT) (Liang et al., 2024) separates modality-specific parameters while preserving interactions through joint self-attention; BAGEL (Deng et al., 2025a) adopts this design with distinct understanding and generation experts. A complementary line connects language or multimodal backbones to separate generative decoders: NExT-GPT (Wu et al., 2023) and SEED-X (Ge et al., 2024) use learned interfaces to diffusion models, while MetaQuery (Pan et al., 2025) introduces learnable queries to extract generation conditions from a frozen MLLM. RUG develops a semantic interface within the unified Transformer computation, where the understanding expert organizes Relay Token states that the generation expert can consume at each layer alongside the original text.

Training for Unified Understanding and Generation. Recent training methods seek to improve the interaction between understanding and generation through representation alignment and task coordination. Reconstruction Alignment (RecA) (Xie et al., 2025) provides dense supervision by reconstructing images from the model's own visual understanding embeddings. Attention Interaction Alignment (AIA) (Zheng et al., 2025) aligns layer-wise cross-modal attention with task-specific interaction patterns obtained from reference models. Semantic Generative Tuning (SGT) (Yu et al., 2026) uses semantic tasks, particularly segmentation, as generative proxies to align the two capabilities. Another line strengthens visual generation through explicit reasoning and reinforcement learning: T2I-R1 (Jiang et al., 2025) jointly optimizes semantic planning and image-token generation with BiCoT-GRPO, while MMaDA (Yang et al., 2025) combines mixed long-chain-of-thought fine-tuning with UniGRPO. RUG addresses semantic transfer through attention topology while retaining the underlying UMM's original training objectives. The generation loss trains the understanding expert to organize input-dependent relay states for visual synthesis, without auxiliary alignment losses or explicit reasoning targets.

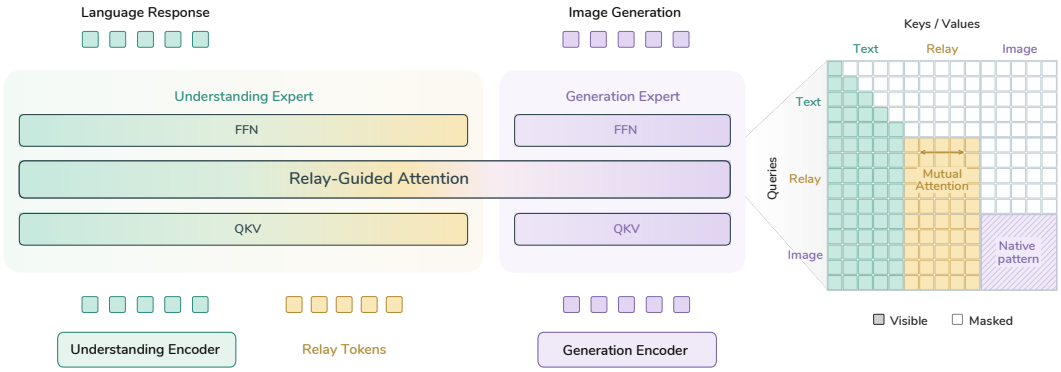

Figure 2: Architecture and attention topology of RUG. The understanding expert U organizes information from text T into Relay Tokens R, which attend to the text and interact bidirectionally with one another. The generation expert G processes image tokens I conditioned jointly on T and R.

3 METHOD

We present RUG, a unified attention topology that relays semantic information from understanding to visual generation.

3.1 MOTIVATION

BAGEL (Deng et al., 2025a) adopts a mixture-of-transformers (MoT) architecture to alleviate optimization conflicts between understanding and generation. An understanding expert U handles comprehension and textual responses, while a generation expert G synthesizes images; their parameters are separated, with interaction through shared attention. Nevertheless, recent studies reveal that understanding and reasoning do not reliably translate into visual generation, including in BAGEL (Niu et al., 2025a; Wang et al., 2026; Yang et al., 2026).

Our analysis shows that language answering and image generation can produce semantically inconsistent outputs even when starting from the same text-prefix KV cache, consistent with the gap illustrated in Fig. 1. Sharing information therefore does not ensure compatible use: the two training objectives place different demands on how that information is consumed. This motivates us to pursue two complementary goals: decoupling the semantic interfaces for textual responses and visual generation, and enabling the understanding expert to organize representations that the generation expert can effectively consume.

3.2 RELAY TOKENS

To instantiate this interface, we introduce a group R of m learnable Relay Tokens. Given an instruction x, let T denote its text tokens and I the image tokens processed by G. We place R between T and I, with initial embeddings

$$R^{(0)} = [r_1; \ldots; r_m] \in \mathbb{R}^{m \times d}, \tag{1}$$

where each $r_i \in \mathbb{R}^d$ is a learnable embedding, d is the hidden dimension, and semicolons denote concatenation along the sequence dimension. The initial vectors are shared across instructions. Conditioned on T, the understanding expert U updates the Relay Token hidden states layer by layer, producing input-dependent states $R^{(\ell)}(x)$ at layer ℓ.

Relay Tokens remain continuous hidden representations throughout this computation. Their number m controls the capacity of the semantic interface and introduces md learnable embedding parameters. We examine the effect of this capacity on visual generation in our ablation studies.

3.3 Relaying Understanding to Generation

RUG routes T and R through the same understanding expert U, while I uses the generation expert G. At each layer, their query, key, and value projections are

$$\begin{aligned}(Q_T, K_T, V_T) &= \mathcal{P}_U(T),\\ (Q_R, K_R, V_R) &= \mathcal{P}_U(R),\\ (Q_I, K_I, V_I) &= \mathcal{P}_G(I),\end{aligned} \tag{2}$$

where $\mathcal{P}_U$ and $\mathcal{P}_G$ denote the experts' projection mappings, and layer indices are omitted for clarity.

As illustrated in Fig. 2, Relay Tokens attend to the full instruction and interact bidirectionally with one another to integrate information distributed across the text.

Text tokens retain causal attention and do not attend to R, keeping the relay states outside the textual response path. Image tokens I attend jointly to T and R, allowing G to consume both the explicit instruction and the semantic information organized by U.

Neither text nor Relay Tokens attend to image tokens, so their keys and values at each layer can be computed once and reused throughout image generation. This semantic transfer takes place within the existing transformer computation, without a separate reasoning pass.

Training objective. RUG is trained with the underlying UMM's generation and understanding objectives:

$$\mathcal{L}_{\mathrm{RUG}} = \lambda_{\mathrm{gen}}\mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{und}}\mathcal{L}_{\mathrm{und}}, \tag{3}$$

where $\mathcal{L}_{\mathrm{gen}}$ supervises visual generation, $\mathcal{L}_{\mathrm{und}}$ supervises textual responses, and the coefficients balance the two tasks. Generation gradients train G and propagate through the relay states into $R^{(0)}$ and the trainable parameters of U, adapting the interface to the needs of visual synthesis. Understanding examples follow the original pathway without Relay Tokens. Using only the model's original training objectives, we examine whether the understanding expert can learn to form relay representations that the generation expert can more effectively consume.

3.4 Comparison with Bridging Methods

Q-Former in BLIP-2 (Li et al., 2023) uses a separate query Transformer to extract information from a frozen image encoder and pass its outputs to a language model. MetaQuery (Pan et al., 2025) learns queries within a frozen multimodal language model and maps their output states through a connector to condition a separate diffusion decoder. RUG integrates semantic transfer into the unified model's attention: Relay Tokens evolve through U across Transformer depth, while G accesses their layer-specific keys and values alongside the text at every layer.

HBridge (Wang et al., 2025) also enables interaction across depth by coupling heterogeneous experts at selected intermediate layers. Its semantic reconstruction tokens reside in the generation expert and are supervised using target-image features. In RUG, R is organized by the understanding expert specifically for generation, while textual responses retain their original attention path. Together, the routing and attention rules over T, R, and I define **a unified attention topology for layer-wise semantic transfer**. This interface is learned through the original UMM objectives, without a separate connector network or an auxiliary semantic reconstruction objective.

4 Experiments

We first present the main results of RUG on image generation benchmarks. We then compare RUG with alternative UMM architectures to examine the benefits of the proposed attention topology. Finally, we conduct ablation studies on Relay Tokens to examine their effectiveness and how increasing their number affects visual generation.

4.1 Experimental Setup

Models and training details. We build RUG on two UMM backbones:

Table 1: Image generation results (%, ↑). All models are re-evaluated for a consistent comparison. RUG is further fine-tuned on the full BLIP3o-60k dataset. **Bold** and underlined scores denote the best and second-best results among UMMs, respectively.

		GenEval	GenEval2					WISE
Model	Params	Overall	Object	Attr.	Count	Pos.	Verb	GM/AM	Overall
Image generation models									
FLUX.1-dev	12B	67.11	88.79	69.74	54.25	40.89	20.64	24.18/69.01	42.00
FLUX.2-klein-9B	9B	84.53	95.74	89.03	58.40	73.44	36.85	34.80/79.46	45.70
Qwen-Image	20B	87.76	98.77	86.48	69.97	59.00	42.75	39.18/82.51	48.40
Qwen-Image-2.1	7B	52.50	78.40	57.50	46.01	41.57	20.70	13.84/59.14	45.90
Unified multimodal models									
Janus-Pro-1B	1B	74.39	49.88	31.33	22.07	26.16	11.37	5.94/34.60	26.30
Janus-Pro-7B	7B	78.79	76.78	57.08	34.47	42.85	23.48	12.83/55.30	34.00
Show-o2-1.5B	1.5B	61.19	84.33	61.29	41.44	38.02	19.40	15.31/61.10	26.60
OmniGen2	3B+4B	77.55	89.06	65.59	47.91	51.39	20.73	18.41/67.32	31.60
InternVL-U	4B	**84.18**	94.58	**79.25**	53.65	50.21	28.33	26.19/73.51	35.70
BAGEL-7B	14B	79.13	91.39	64.17	51.78	44.85	29.58	21.19/68.48	34.80
RUG (Janus-Pro-1B)	1B	77.52	92.43	66.21	43.48	**58.22**	30.40	19.70/67.34	27.70
vs. baseline		+3.13	+42.55	+34.88	+21.41	+32.06	+19.04	+13.76/+32.74	+1.40
RUG (BAGEL-7B)	14B	84.08	**95.46**	75.93	**58.29**	53.70	**33.98**	**28.04/75.06**	**40.90**
vs. baseline		+4.96	+4.07	+11.76	+6.51	+8.85	+4.39	+6.85/+6.57	+6.10

- **BAGEL-7B** (Deng et al., 2025b) uses a MoT architecture with separate understanding and generation experts. We add LoRA adapters to both experts.
- **Janus-Pro-1B** (Chen et al., 2025c) uses a shared autoregressive Transformer for both tasks. We introduce two independent LoRA branches over this backbone to instantiate understanding and generation experts.

Both models follow the routing and attention topology in Sec. 3.3. We train rank-128 LoRA adapters on the query, key, and value projections and 128 Relay Tokens, keeping all pretrained parameters frozen. Unless otherwise specified, training runs for 10,000 steps on 8 NVIDIA H100 or H200 GPUs with a global batch size of 96 and a generation:understanding ratio of $1:2$.

Training data. We use BLIP3o-60k (Chen et al., 2025a) and FLUX-Reason-6M (Fang et al., 2025) for image generation, and LLaVA-OneVision (Li et al., 2024) for multimodal understanding. By default, we filter BLIP3o-60k to remove GenEval-template examples and overlapping prompts, addressing the evaluation-bias concern discussed in RecA (Xie et al., 2025). We additionally fine-tune on the full BLIP3o-60k dataset only for the comparison with other baselines in Table 1. Architecture comparisons and Relay Token analyses use the filtered-data setting.

Evaluation. We evaluate image generation on GenEval, GenEval2, and WISE, image editing on ImgEdit, and multimodal understanding on MME and MMMU. Baseline models include the FLUX and Qwen-Image series, BAGEL, Janus-Pro, Show-o2, OmniGen2, and InternVL-U. Given limited access to ImgEdit's original judge models, we adopt the readily available open-source Qwen3.5-35B-A3B, following the judge choice in the official WISE_Verified update (Niu et al., 2025b). We retain the benchmark-provided scoring prompts and use the same judge across locally evaluated models.

4.2 Main Results

We fine-tune Janus-Pro-1B and BAGEL-7B using the proposed RUG attention topology. Both backbones benefit from high-quality data curated for their original training (Chen et al., 2025c; Deng

Table 2: Transformer variants on Janus-Pro-1B (%, ↑). Each row uses the checkpoint with the highest Avg. among 1K-spaced evaluations up to 10K steps. Avg. equally weights GenEval, GenEval2 AM, and WISE. GenEval2 skill scores use AM. Text + Relay is RUG. **Bold**/underlined scores indicate best/second-best results.

			GenEval	GenEval2							WISE	DPG	Avg.
Type	Variant	Step	Overall	Object	Attr.	Count	Pos.	Verb	GM	AM			
Dense	Vanilla	9K	72.70	87.16	60.89	35.88	53.28	21.72	15.11	61.05	27.30	82.16	53.68
	Gen. Only	8K	75.52	89.43	61.15	36.32	54.14	20.99	15.27	62.47	27.30	82.57	55.10
	Und. Only	1K	62.58	32.06	21.19	14.85	14.44	6.98	2.61	22.80	26.00	81.27	37.12
MoT	Vanilla	8K	72.87	87.58	61.95	35.97	55.45	20.94	15.67	62.05	27.20	82.65	54.04
	Relay Only	10K	75.89	89.50	62.11	39.05	**58.67**	19.71	18.13	64.12	**27.90**	82.23	55.97
	Text + Relay	9K	**77.68**	**90.49**	**62.20**	**40.97**	58.02	**26.85**	**18.92**	**65.13**	27.60	**82.71**	**56.80**

et al., 2025a). To evaluate RUG in a practical instruction fine-tuning setting, we further train both models for two epochs on a 1:1 mixture of the full BLIP3o-60k dataset and LLaVA-OneVision.

Table 1 shows consistent gains across GenEval, GenEval2, and WISE on both Janus-Pro-1B and BAGEL-7B, despite their different model scales and underlying architectures. On GenEval2 (Kamath et al., 2025), which evaluates compositional generation across objects, attributes, counts, and relations, Janus-Pro-1B achieves a substantial improvement in overall GM ($5.94\% \rightarrow 19.70\%$). The marked gains in attribute and spatial-relation scores on both backbones further suggest improved adherence to multiple semantic requirements within a single instruction.

4.3 Comparison of Transformer Variants

Janus-Pro (Chen et al., 2025c) uses a shared autoregressive Transformer for understanding and generation, making it a convenient backbone for studying parameter sharing and attention designs. We construct six variants on Janus-Pro-1B, grouped into two families:

- **Dense Transformer.** We retain Janus-Pro's original shared Transformer and use a single LoRA branch. **Dense-Vanilla** optimizes understanding and generation jointly. **Dense-Gen. Only** and **Dense-Und. Only** fine-tune the model solely for generation or understanding, respectively, providing two task-specialized baselines.
- **Mixture-of-Transformers (MoT).** We use the two-expert adaptation of Janus-Pro described in Sec. 4.1. All three variants use joint training: **MoT-Vanilla** has no Relay Tokens and lets G attend directly to text; **MoT-Relay Only** conditions G on Relay Tokens alone; and **MoT-Text + Relay**, our full RUG model, conditions G on both text and Relay Tokens.

Our single-task variant training retains its corresponding task batch to keep the same total steps.

Overall Optimization. We use *Dense Gen. Only* and *Dense Und. Only* as references for task-specialized generation and understanding models. Each optimizes a single objective, avoiding competition between the two tasks during fine-tuning, whereas jointly trained UMMs must accommodate potentially conflicting updates. In Fig. 3a,b, *Dense Vanilla* exhibits higher late-training losses than the corresponding single-task references. *MoT Vanilla* narrows both gaps, supporting the benefit of expert separation in alleviating task interference. The nested views reveal a promising result: *MoT Text + Relay* further lowers generation loss, slightly below *Dense Gen. Only*, while retaining an understanding loss close to *Dense Und. Only*. RUG thus approaches the loss levels of task-specialized models within a jointly trained UMM.

Image Generation Performance. Table 2 shows that *MoT Vanilla* modestly improves over *Dense Vanilla*, while adding Relay Tokens yields larger gains. RUG consistently leads in generation average and exceeds the best non-relay baseline within 3K steps (Fig. 3c). Both relay variants also outperform *Dense Gen. Only*, showing that stronger generation can be achieved while jointly optimizing understanding. *MoT Relay Only* reaches an average score of 55.97, showing that instruction

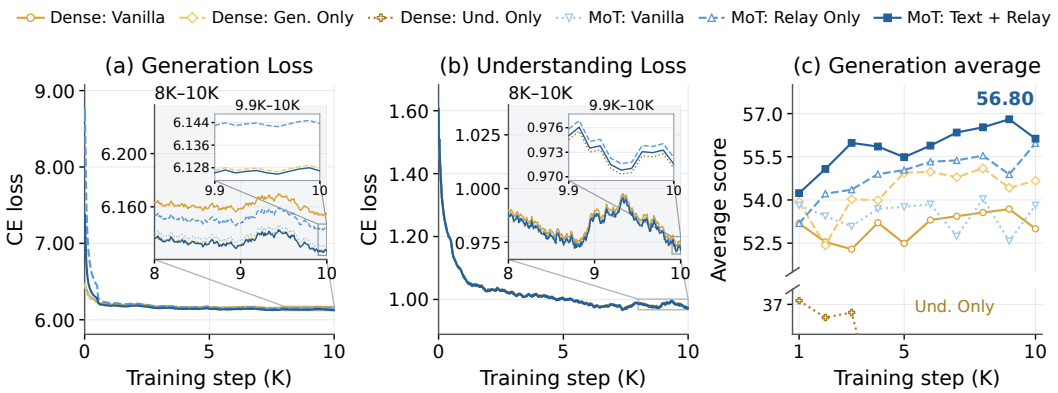

Figure 3: **Training dynamics of Transformer variants.** Losses use a trailing 500-step mean. Main axes and 8K–10K insets retain all applicable variants. Nested 9.9K–10K views highlight differences among three low-loss variants. The generation average is the equal-weight mean of GenEval overall, GenEval2 AM, and WISE, with all scores on a 0–100 scale. Scores are evaluated every 1K steps; the label marks the best RUG checkpoint. The broken score axis shows the early *Dense Und. Only* checkpoints in its lower segment; later scores below that segment are clipped.

Table 3: Image editing on ImgEdit Basic (737 examples, scores out of 5, ↑). BAGEL is the pretrained model; MoT-Vanilla and RUG use matched 10K checkpoints. All models share the demo-based sampling and the Qwen3.5-35B-A3B judge with benchmark-provided scoring prompts. Best scores in each column are shown in **bold**.

Variant	Add	Adjust	Remove	Replace	Background	Style	Action	Extract	Compose	Overall
BAGEL	**3.77**	3.75	3.31	3.29	3.52	4.58	4.30	2.24	2.88	3.49
MoT-Vanilla	3.69	**3.95**	**3.68**	3.50	**3.73**	4.70	**4.37**	3.38	**3.45**	3.82
RUG	3.76	3.90	3.65	**3.63**	3.67	**4.72**	4.32	**3.47**	3.25	**3.83**

information organized into relay states provides effective generative conditioning without direct text access. *MoT Text + Relay* further improves this to 56.80, supporting the use of Relay Tokens to supply understanding-derived context alongside direct access to the full instruction.

Image Editing Performance. We compare pretrained BAGEL with two first-stage 10K variants: *MoT-Vanilla*, which retains the original MoT topology without Relay Tokens, and *RUG*, our Text + Relay model. Despite using no editing data during fine-tuning, RUG improves ImgEdit from 3.49 to 3.83, with gains in eight of nine categories (Table 3). The largest improvements over pretrained BAGEL are in extraction, composition, removal, and replacement. These results show that generation and understanding fine-tuning can also improve image editing without direct editing supervision.

4.4 Relay Token Analysis

We vary $m \in \{16, 32, 64, 128, 256\}$ on Janus-Pro-1B, comparing *MoT Relay Only* with *MoT Text + Relay* (RUG). All settings use the same training recipe and 10K steps. Figure 4 reports the average of GenEval, GenEval2 AM, and WISE.

Increasing m from 16 to 128 raises *MoT Relay Only* from 51.72 to 55.97, suggesting that relay capacity matters more when all textual conditioning passes through R. *MoT Text + Relay* performs better at every tested count, reaching 56.17 with 64 tokens and remaining nearly unchanged at 128. Its advantage is largest at 16 tokens, consistent with direct text access reducing the bottleneck of a small relay interface. Gains

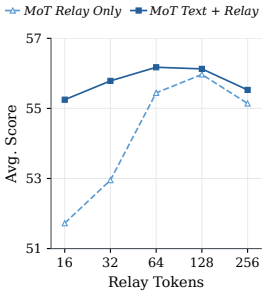

Figure 4: **Relay Token count.**

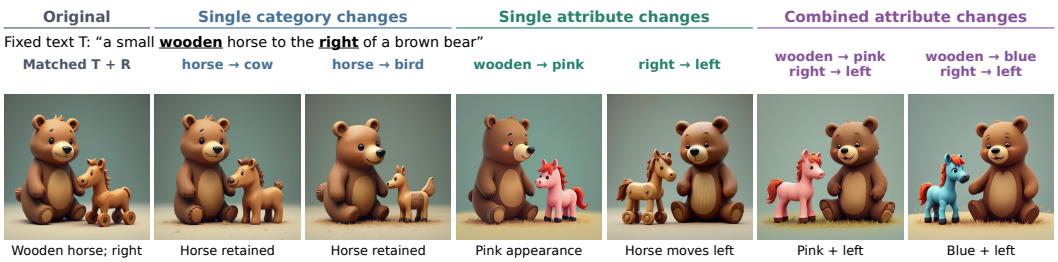

Figure 5: **Relay Transplantation Test.** The first column uses matched text and relay states. The other columns **replace only the Relay Token keys and values at all layers**, holding the recipient text and initial noise fixed. Bold, underlined words mark appearance and position in the unchanged text. Labels above the images specify donor-prompt changes; labels below describe the outputs.

saturate around 64–128 tokens, with no further improvement at 256 under this training budget.

4.5 Analysis of Relay Representations

Relay Transplantation Test. We test how generation uses relay information by changing R while keeping the direct text condition fixed. Using the first-stage 10K-step BAGEL RUG checkpoint, we fix the instruction to "a small **wooden** horse to the **right** of a brown bear." Donor prompts modify the subject category, appearance, position, or their combinations. We transplant their Relay Token keys and values at all 28 layers, keeping the original text cache, initial noise, and unconditional branch unchanged. Figure 5 compares the matched text–relay pair (column 1) with these interventions (columns 2–7).

Results and Interpretation. Category substitutions from *horse* to *cow* or *bird* retain a horse-like subject (columns 2–3). Appearance and position changes transfer: replacing *wooden* with *pink* changes the horse's appearance (column 4), while replacing *right* with *left* places it to the left of the bear (column 5). Combined substitutions produce a pink or blue horse on the left (columns 6–7). Thus, changing R alone can transfer multiple semantic constraints together despite the unchanged text.

These observations support our hypothesis that **understanding organizes attributes and relations into Relay Tokens as visual constraints that generation can consume.** In this example, direct text appears to anchor subject identity, while relay states guide appearance and layout. This interpretation is consistent with the main results after additional instruction fine-tuning: RUG improves BAGEL's GenEval2 attribute score from 64.17 to 75.93 and position score from 44.85 to 53.70 (Table 1), its two largest gains across the five skill categories.

4.6 Qualitative Results

We compare *RUG* with *MoT-Vanilla* on BAGEL using matched 10K-step checkpoints. Generation comparisons use identical prompts; editing comparisons additionally fix the source image and use the same initial noise.

Image Generation. Figure 6 illustrates improved adherence to multiple semantic constraints in the selected examples. RUG recovers the two purple flowers omitted by MoT-Vanilla while retaining the three dogs behind them. It also satisfies the requested counts of umbrellas and glass horses, and places the car on the correct side of the dog. These examples illustrate how relay conditioning helps generation jointly realize object, attribute, and relational requirements within a single scene.

Image Editing. Figure 7 shows corresponding improvements in following editing instructions. RUG replaces the fish with a single rubber duck, raises the person's requested left arm, and applies the darker brown appearance more consistently across the wooden panels. The coffee-cup example further illustrates relational grounding: RUG places the cup on a supporting table, while

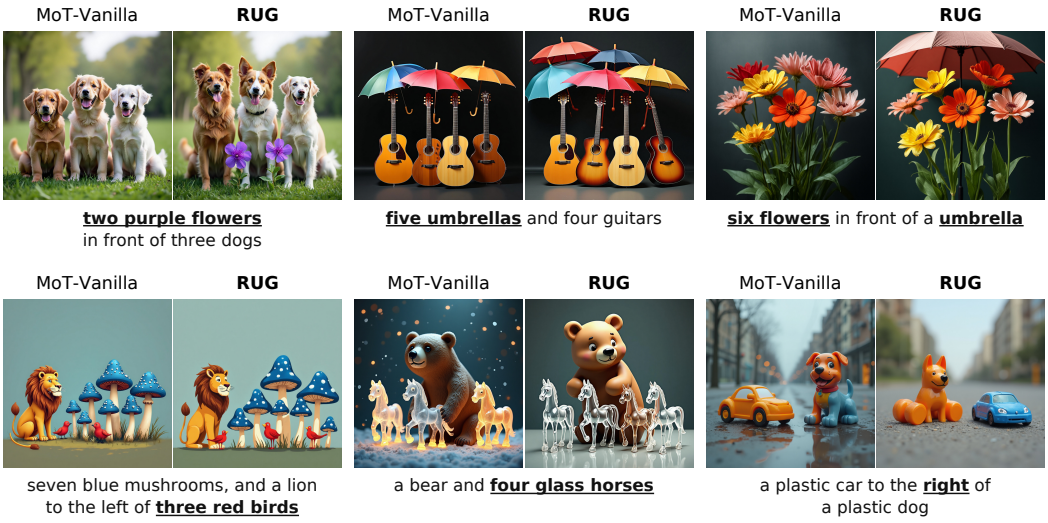

Figure 6: **Image generation comparisons on BAGEL.** We compare *MoT-Vanilla* and *RUG*, both fine-tuned for 10K steps. Each pair uses the same prompt and random seed. Underlined phrases highlight the target constraints.



Figure 7: **Image editing comparisons on BAGEL.** Each triplet shows the input image and outputs from *MoT-Vanilla* and *RUG*, both fine-tuned for 10K steps. The two variants use the same input image, editing instruction, and initial noise. Underlined phrases highlight the target edits.

MoT-Vanilla leaves it unsupported. These selected cases demonstrate how the relay interface can support more precise execution of the object, region, and relation constraints specified by an editing instruction.

5 Conclusion and Limitations

We presented RUG, a new MoT variant that uses learnable Relay Tokens to relay semantic information from understanding to generation at every layer. Trained with the original UMM objectives, RUG improves compositional generation on Janus-Pro and BAGEL. Controlled comparisons on Janus-Pro further show a higher generation average than vanilla MoT and generation-only fine-tuning.

Our evaluation focuses on parameter-efficient fine-tuning within MoT. Applicability to other unified architectures and large-scale pretraining remains to be explored. Understanding results are mixed, leaving room to better balance the two capabilities.

AI USE STATEMENT

We used generative AI tools to coding and improve the clarity, wording, and presentation of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

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

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

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

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*, 2025b.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025. URL https://arxiv.org/abs/2509.09680.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal Models with Unified Multi-granularity Comprehension and Generation. *arXiv preprint arXiv:2404.14396*, 2024. URL https://arxiv.org/abs/2404.14396.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing Image Generation with Collaborative Semantic-level and Token-level CoT. *arXiv preprint arXiv:2505.00703*, 2025. URL https://arxiv.org/abs/2505.00703.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. URL https://arxiv.org/abs/2301.12597.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. *arXiv preprint arXiv:2411.04996*, 2024. URL https://arxiv.org/abs/2411.04996.

Jinyu Liu, Xincheng Shuai, Henghui Ding, and Yu-Gang Jiang. Unison: Benchmarking unified multimodal models via synergistic understanding and generation, 2026. URL https://arxiv.org/abs/2606.26984.

Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv preprint arXiv:2511.20561*, 2025a. URL https://arxiv.org/abs/2511.20561.

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

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

Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

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

Chenlong Wang, Yuhang Chen, Zhihan Hu, Dongping Chen, Wenhu Chen, Sarah Wiegreffe, and Tianyi Zhou. Quantifying the gap between understanding and generation within unified multimodal models. *arXiv preprint arXiv:2602.02140*, 2026. URL https://arxiv.org/abs/2602.02140.

Xiang Wang, Zhifei Zhang, He Zhang, Zhe Lin, Yuqian Zhou, Qing Liu, Shiwei Zhang, Yijun Li, Shaoteng Liu, Haitian Zheng, Jason Kuen, Yuehuan Wang, Changxin Gao, and Nong Sang. HBridge: H-Shape Bridging of Heterogeneous Experts for Unified Multimodal Understanding and Generation. *arXiv preprint arXiv:2511.20520*, 2025. URL https://arxiv.org/abs/2511.20520.

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

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05519*, 2023. URL https://arxiv.org/abs/2309.05519.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025. URL https://arxiv.org/abs/2509.07295.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. *arXiv preprint arXiv:2408.12528*, 2024. URL https://arxiv.org/abs/2408.12528.

Cheng Yang, Chufan Shi, Bo Shui, Yaokang Wu, Muzi Tao, Huijuan Wang, Ivan Yee Lee, Yong Liu, Xuezhe Ma, and Taylor Berg-Kirkpatrick. UReason: Benchmarking reasoning-to-generation alignment in unified multimodal models. *arXiv preprint arXiv:2602.08336*, 2026. URL https://arxiv.org/abs/2602.08336.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal Large Diffusion Language Models. *arXiv preprint arXiv:2505.15809*, 2025. URL https://arxiv.org/abs/2505.15809.

Songsong Yu, Yuxin Chen, Ying Shan, and Yanwei Li. Semantic Generative Tuning for Unified Multimodal Models. *arXiv preprint arXiv:2605.18714*, 2026. URL https://arxiv.org/abs/2605.18714.

Dian Zheng, Manyuan Zhang, Hongyu Li, Kai Zou, Hongbo Liu, Ziyu Guo, Kaituo Feng, Yexin Liu, Ying Luo, and Hongsheng Li. Aia: Rethinking architecture decoupling strategy in unified multimodal model. *arXiv preprint arXiv:2511.22663*, 2025.

A Appendix

You may include other additional sections here.