AIA: RETHINKING ARCHITECTURAL DECOUPLING IN UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) integrate image understanding and generation within a single architecture, yet the two tasks require different visual representations, creating inherent conflicts during joint training. Recent methods increasingly improve UMMs through architectural decoupling (e.g., decoupled visual encoders, MoE/MoT architectures, or frozen MLLMs), but why decoupling works remains poorly understood. In this work, we revisit architectural decoupling through layer-wise cross-modal attention interaction patterns. Our analysis reveals a previously overlooked phenomenon: as architectural decoupling becomes stronger, understanding and generation retain distinct interaction regimes, while each task's interaction pattern becomes increasingly similar to that of a strong task-specific reference model. This suggests that the benefit of architectural decoupling may stem from enabling **task-specific cross-modal interaction specialization**, rather than driving the two tasks toward a shared interaction behavior. Motivated by this finding, we propose **A**ttention **I**nteraction **A**lignment (**AIA**), which adaptively transfers task-specific reference interaction behavior without further architectural decoupling. Specifically, **stage-level interaction alignment** transfers the overall cross-layer interaction knowledge of the reference while preserving the student's own representation organization, whereas a **Huber loss** retains the model's ability to adaptively adjust interaction across individual layers. Applied to Emu3 during supervised fine-tuning and Janus-Pro during post-training, AIA consistently improves both understanding and generation, and further improves interleaved cross-task performance. These findings suggest that strong unified multimodal performance does not require understanding and generation to converge to a shared interaction regime; instead, each task can retain and exploit its own appropriate interaction behavior.

1 INTRODUCTION

Unified Multimodal Models (UMMs) aim to support visual understanding and generation within a single network. A fully unified architecture shares a latent space across tasks, supporting interleaved multimodal processing and avoiding unnecessary conversion between visual representations.

The central difficulty is that understanding and generation require different visual representations. Understanding mainly relies on high-level semantic information, whereas generation must preserve and reconstruct fine-grained visual details. This mismatch creates an inherent conflict when the two tasks are optimized within the same model. Early fully unified models (Wang et al., 2024; Wu et al., 2024b; Team, 2024), which share both the visual encoder and backbone, therefore lag behind strong task-specific systems. To alleviate this conflict and improve performance, recent work increasingly adopts architectural decoupling, including decoupled visual encoders or training objectives, MoE/MoT components, and frozen MLLMs with generation heads (Chen et al., 2025c; Wu et al., 2025b; Li et al., 2025; Deng et al., 2025). These designs are effective, and stronger decoupling has become a common route to stronger UMM performance. Despite its effectiveness, however, why architectural decoupling improves UMMs remains poorly understood.

To answer this question, we introduce a layer-wise cross-modal attention interaction intensity to characterize how strongly one modality attends to the other at each layer. We then use this indicator to analyze UMMs with different degrees of architectural decoupling. As shown in Fig 1, we obtain

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.

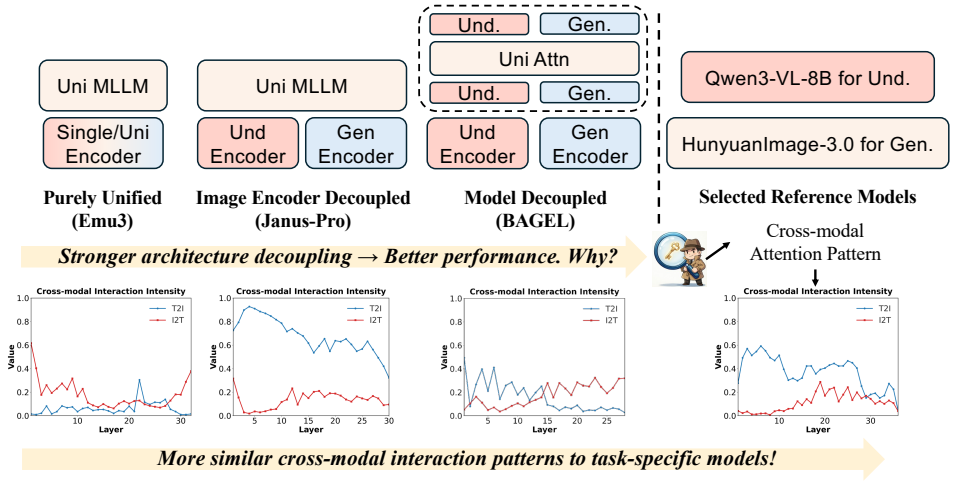

Figure 1: UMMs with different degrees of architectural decoupling and the cross-modal interaction patterns measured by our layer-wise interaction intensity. Models are arranged from less to more decoupled. Generation is shown in blue and understanding in red; higher values indicate stronger cross-modal interaction. Qwen3-VL-8B and HunyuanImage-3.0 serve as strong understanding and generation references, respectively. The negative correlation between understanding and generation persists as decoupling increases, while each task's pattern approaches its reference.

two observations. First, understanding and generation exhibit consistently negative correlations in their interaction patterns, and this behavior persists even in more strongly decoupled models, indicating that the two tasks retain distinct interaction regimes as decoupling becomes stronger. Second, as decoupling increases, the interaction pattern of each task becomes increasingly similar to that of a strong reference model for the same task. Together, these observations reveal a previously overlooked phenomenon: stronger architectural decoupling is accompanied by increasingly task-specific cross-modal interaction. This suggests that the benefit of architectural decoupling may stem from enabling task-specific cross-modal interaction specialization.

This finding suggests a direct alternative. If the useful behavior is reflected in the interaction pattern, we can guide that behavior without further decoupling the architecture. We therefore propose Attention Interaction Alignment (AIA), which explicitly regularizes layer-wise cross-modal interaction toward task-dependent reference behavior while preserving the original architecture. Two design choices are critical. First, exact layer-wise alignment is undesirable because reference and student models may organize their representations differently across depth. We therefore introduce stage-level interaction alignment, which preserves only the coarse shallow-to-deep interaction structure rather than imposing one-to-one layer correspondence. Second, rigidly fixing the interaction intensity of individual layers can suppress the model's ability to adapt internally to the competing requirements of understanding and generation. We therefore adopt a Huber loss to provide soft interaction guidance while retaining layer-wise flexibility within each stage. We use Qwen3-VL-8B (Bai et al., 2025a) as the understanding reference and HunyuanImage-3.0 (Cao et al., 2025) as the generation reference.

We evaluate AIA in two representative settings: Emu3, a purely unified model, during supervised fine-tuning, and Janus-Pro, which uses decoupled visual encoding, during post-training. AIA consistently improves both understanding and generation benchmarks. We further compare AIA with feature- and output-level distillation to verify that the gain comes from interaction-pattern guidance rather than generic teacher transfer. Finally, we validate the effectiveness of AIA on interleaved cross-task reasoning without specific tuning.

We summarize our contributions as follows:

- We introduce a layer-wise cross-modal attention interaction intensity to revisit architectural decoupling in UMMs. Our analysis reveals a previously overlooked phenomenon: as decoupling becomes stronger, understanding and generation retain distinct interaction

regimes, while each task increasingly approaches the interaction pattern of a strong task-specific reference.

- We propose Attention Interaction Alignment (AIA) to directly shape task-specific cross-modal interaction without further architectural decoupling. AIA combines **stage-level interaction alignment** to avoid unreliable layer-wise correspondence between reference and student models with a **Huber-based flexible constraint** that preserves layer-wise adaptation within each stage.
- We validate AIA on Emu3 and Janus-Pro across different architectures and training stages. AIA consistently improves both understanding and generation, outperforms feature- and output-level distillation, generalizes across model and data scales, and further improves interleaved cross-task performance.

2 Related Works

2.1 Unified Multimodal Models (UMMs)

With the rapid progress of LLMs and MLLMs (Bai et al., 2025b; Team et al., 2024; Yang et al., 2025; Wu et al., 2024d; Liu et al., 2025), unified multimodal models have been widely explored to integrate visual understanding and generation within a single framework. Early attempts, such as Liquid (Wu et al., 2024b), Emu3 (Wang et al., 2024), and Chameleon (Team, 2024), use highly unified autoregressive architectures but rely on VAE/VQ tokenizers (Kingma & Welling, 2013; Esser et al., 2021; Van Den Oord et al., 2017), which limits understanding performance. Unified-tokenizer methods including TokenFlow (Qu et al., 2025), Unitok (Ma et al., 2025), UniLip (Tang et al., 2025), and VTP (Yao et al., 2025) improve tokenizer compatibility and scalability. Another line of work widely adopts architectural decoupling to alleviate the understanding–generation conflict. Janus (Chen et al., 2025c; Wu et al., 2024a; Ma et al., 2024), Show-o (Xie et al., 2025b;c), and Transfusion-style models (Zhou et al., 2024; Cao et al., 2025) decouple visual encoders or training objectives; BAGEL (Deng et al., 2025) and OneCat (Li et al., 2025) adopt partially decoupled MoT/MoE architectures; and Metaquery (Pan et al., 2025), OmniGen2 (Wu et al., 2025b), and BLIP3-o (Chen et al., 2025a) keep the MLLM fixed and mainly optimize the generation branch. These methods demonstrate the practical effectiveness of architectural decoupling. In contrast, our work investigates why architectural decoupling improves UMMs by using cross-modal attention interaction to study what changes as models become increasingly decoupled.

2.2 Reference Models

Strong understanding and generation models provide useful reference behaviors for our analysis. For understanding, models such as Qwen (Bai et al., 2025b;a) and LLaVA (Li et al., 2024) use a semantic visual encoder with an autoregressive LLM. For generation, the field spans CLIP-conditioned diffusion models such as SDXL (Podell et al., 2023), SD3 (Esser et al., 2024), and FLUX (Labs, 2024), autoregressive models such as SimpleAR (Wang et al., 2025b), and MLLM-conditioned diffusion models such as Qwen-Image (Wu et al., 2025a), Longcat-Image (Team et al., 2025), Z-Image (Cai et al., 2025), and HunyuanImage-3.0 (Cao et al., 2025). We use Qwen3-VL-8B as the understanding reference and HunyuanImage-3.0 as the generation reference. HunyuanImage-3.0 is itself a unified model; here we use only its generation-side interaction behavior as a reference.

2.3 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher model to a student model, while FitNets (Romero et al., 2014) extends this idea by aligning intermediate features. **AIA differs from conventional output- or feature-level distillation** in two ways. First, it uses only pre-computed layer-wise cross-modal interaction statistics and therefore requires no concurrent teacher model during training. Second, AIA does not ask the student to reproduce teacher outputs or internal features. It provides coarse, task-dependent interaction targets while leaving the model free to optimize its own representations. As shown in Table 5, the tested feature- and output-level distillation baselines do not provide the same consistent gains as interaction-pattern guidance.

Table 1: Standard deviation across 1000 und. and 1000 gen. samples for each model.

Method	Emu3	Janus-Pro	BAGEL	Reference
Std	0.11	0.04	0.07	0.09

3 Attention Interaction Alignment

This section first introduces our cross-modal attention interaction intensity and uses it to examine how task-dependent interaction changes under different degrees of architectural decoupling. We find that understanding and generation retain distinct interaction regimes as decoupling becomes stronger, while each task increasingly approaches a strong task-specific reference. Based on this observation, we introduce Attention Interaction Alignment (AIA), which transfers this task-specific interaction behavior through stage-level alignment while retaining flexible layer-wise adaptation with a Huber loss.

3.1 Analysis of Cross-Modal Attention Interaction Pattern in Various Model Architectures

Method. We introduce cross-modal attention interaction intensity to measure how strongly one modality attends to the other at each layer. Taking text-to-image generation as an example, as shown in Fig 2, we first obtain the attention map after softmax. For each image-query row, we sum the attention weights assigned to text-key tokens, and then average over all image-query tokens and attention heads. This gives the cross-modal interaction intensity at layer l:

$$I_l = \frac{1}{H \times Q} \sum_{h=1}^{H} \sum_{q=1}^{Q} \sum_{k \in \mathcal{K}_{\text{text}}} \text{Attn}_l(h, q, k), \quad (1)$$

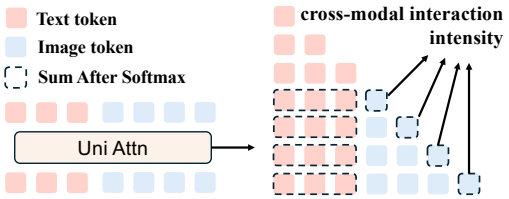

Figure 2: Cross-modal interaction intensity for text-to-image generation. For each image-query row, we sum the attention weights assigned to text-key tokens and then average over image queries and attention heads.

where H is the number of attention heads, Q is the number of image-query tokens, and $\mathcal{K}_{\text{text}}$ denotes the set of text-key token indices. For understanding, we analogously measure the attention mass from text-query tokens to image-key tokens. We compute the layer-wise interaction patterns over 1,000 understanding and generation samples respectively (detail in Table A2), and average them to obtain the curves in Fig 1. Table 1 reports the standard deviation between these samples, showing that the overall pattern is stable between different input types.

Reference interaction patterns. Fig 1 shows that Qwen3-VL (Bai et al., 2025a) consistently assigns relatively low attention mass to image tokens, which is consistent with the observation behind visual-token pruning methods (Wang et al., 2025a; Zhang et al., 2024) for understanding. For generation, HunyuanImage-3.0 (Cao et al., 2025) maintains around 40% attention to text through most early and middle layers, followed by a sharp decline in the final layers. This behavior is consistent with the common view in generative models (Yu et al., 2025): earlier layers rely more on high-level semantic conditions, whereas later layers increasingly focus on fine-grained visual reconstruction. This decline in text attention also aligns with RecA's motivation (Xie et al., 2025a). We use Qwen3-VL-8B for understanding and HunyuanImage-3.0 for generation as strong references.

What does architectural decoupling actually change? Architectural decoupling is commonly introduced to alleviate the conflict between understanding and generation. Our interaction analysis shows that the negatively correlated patterns of the two tasks persist across all examined architectures, indicating that understanding and generation retain distinct interaction regimes even as decoupling becomes stronger. What changes more systematically is the pattern of each individual task: as decoupling increases, its cross-modal interaction becomes increasingly similar to the corresponding reference behavior. Janus-Pro, for example, is closer to the understanding reference than Emu3, while Emu3 deviates substantially from both references. **These observations reveal that stronger**

architectural decoupling is accompanied by increasingly task-specific cross-modal interaction, suggesting that its performance benefit may stem from enabling task-specific interaction specialization. This motivates us to directly optimize the interaction pattern itself.

3.2 Attention Interaction Alignment Loss

Based on the above observation, we propose **Attention Interaction Alignment (AIA)** to guide task-dependent cross-modal interaction toward strong reference behavior without further architectural decoupling. Rather than rigidly matching the reference interaction curve layer by layer, AIA transfers its coarse interaction structure while preserving the student's layer-wise flexibility.

Stage-level interaction alignment. Exact layer-wise alignment assumes that reference and student models organize representations similarly across depth, which may not hold across different architectures. We therefore align only coarse interaction stages, encouraging the student to preserve the broad shallow-to-deep interaction pattern rather than an exact layer correspondence. After normalizing layer indices by model depth and interpolating the reference curve to the student depth, we partition the curve according to its major turning points. For each stage, the mean interaction intensity defines the target T_l. The final stage boundary is denoted by τ, with $\tau = 30$ for Emu3 and $\tau = 29$ for Janus-Pro.

Table 2: Configuration of Huber loss in Emu3 and Janus-Pro.

Layer Range	**Generation** (δ_l, T_l)	**Understanding** (δ_l, T_l)
$0 \le l < 10$	$(0.2, 0.4)$	$(0.05, 0.1)$
$10 \le l < 20$	$(0.1, 0.4)$	$(0.05, 0.15)$
$20 \le l < 25$	$(0.1, 0.4)$	$(0.05, 0.3)$
$25 \le l \le \tau$	$(0.05, 0.2)$	$(0.05, 0.3)$
$l > \tau$	$(0.05, 0.2)$	$(0.05, 0.2)$

Flexible intra-stage alignment. Even within the same stage, rigidly constraining every layer can suppress the model's ability to adapt its interaction behavior during joint understanding–generation training. We therefore do not impose any reference-derived ordering among layers within a stage, and use a Huber loss to provide soft guidance around the stage target:

$$\mathcal{L}_{\text{AIA}} = \frac{1}{L} \sum_{l=1}^{L} \begin{cases} \frac{1}{2}(I_l - T_l)^2, & |I_l - T_l| \le \delta_l, \\ \delta_l |I_l - T_l| - \frac{1}{2}\delta_l^2, & \text{otherwise}, \end{cases} \tag{2}$$

where I_l is the interaction intensity at layer l, and δ_l is obtained from the within-stage standard deviation. This constrains the overall interaction regime while retaining fine-grained layer-wise adaptation.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda \mathcal{L}_{\text{AIA}}, \tag{3}$$

where λ is set to 40 by default. Table 2 summarizes the stage-wise (δ_l, T_l) configurations.

3.3 Training Details Under Different Architectures

We test whether AIA can induce the desired task-dependent interaction patterns without further architectural decoupling, using supervised fine-tuning for Emu3 and post-training for Janus-Pro.

Emu3. We initialize from the pretrained Emu3 weights and perform unified SFT on our training data. The original Emu3 paper reports task-specific SFT results, whereas our setting trains understanding and generation jointly. As shown in Fig 5, the NTP loss curves remain similar across the tested AIA coefficients, indicating that the additional interaction guidance does not disrupt the main language-modeling objective when the target is chosen appropriately.

Janus-Pro. Janus-Pro provides only its final SFT checkpoint, so we apply AIA as post-training. This is a stricter setting because the model's interaction pattern is already relatively stable, as also reflected by the lower variance in Table 1. Fig 5 shows that Janus-Pro is more sensitive to the AIA coefficient, but an appropriate λ still moves the interaction pattern toward the desired reference and improves downstream performance. We analyze this sensitivity further in the ablation study.

Table 3: System-level comparison on image understanding and generation benchmarks. † denotes results re-implemented by us. "Types" indicates diffusion, autoregressive, or masked-prediction training. Gray values come from the original papers and were not reproduced in our setup.

Method	Params	Types	Image Understanding						Image Generation	
			MMMU	MMBench	MMVP	MMVet	POPE	MME-P	GenEval	DPG
Gen. Only										
SD3-medium (Esser et al., 2024)	2B	Diff			-				0.74	84.08
Infinity (Han et al., 2025)	8B	VAR			-				0.79	86.60
FLUX.1-dev (Labs, 2024)	12B	Diff			-				0.82	84.00
Emu3-Gen (Wang et al., 2024)	8B	AR			-				0.66	80.60
Qwen-Image (Wu et al., 2025a)	7B+20B	AR+Diff			-				0.87	88.32
Und. Only										
Emu3-Chat (Wang et al., 2024)	8B	AR	31.6	58.5	36.6	37.2	85.2	1244	-	-
InternVL3 (Zhu et al., 2025)	8B	AR	62.7	83.4	-	81.3	91.1	-	-	-
Qwen3-VL (Bai et al., 2025a)	8B	AR	69.6	85.0	-	-	-	-	-	-
Uni. Double Image Encoders / Training Objectives										
Show-o (Xie et al., 2025b)	1.3B	AR+Mask	26.7	-	-	-	80.0	1097	0.69	67.27
ARM (Wang et al., 2026)	7B	AR+Diff	40.2	-	-	-	87.3	1463	0.79	84.48
Janus-Pro (Chen et al., 2025c)	7B	AR	41.0	65.54 (79.2)	47.3	50.0	87.4	1567	0.80	84.19
Janus-Pro + AIA (Ours)	7B	AR	**42.1**	**75.6**	**48.0**	**51.2**	**89.8**	**1656**	**0.81**	**84.49**
Uni. Purely										
Chameleon (Team, 2024)	7B	AR	28.4	35.7	-	8.3	-	-	0.39	-
VILA-U (Wu et al., 2024c)	7B	AR	32.2	**66.6**	**22.0**	**27.7**	**83.9**	**1336**	0.39	72.48
Emu3 (Wang et al., 2024) †	8B	AR	31.6	61.4	8.7	15.1	77.3	910	0.60	79.24
Emu3 + AIA (Ours)	8B	AR	**35.7**	64.8	10.8	18.7	82.7	1084	**0.67**	**81.20**

Table 4: Controlled study of architectural decoupling and AIA. All variants use the same backbone, model scale, training data, and optimization recipe.

Architecture	AIA	Corr.(Und., Ref.)	Corr.(Gen., Ref.)	MMMU	MME-P	GenEval	DPG
Purely Unified	×	-0.39	-0.02	29.8	1249.0	0.48	75.3
Purely Unified	✓	0.47	0.48	37.4	1445.9	0.61	80.5
Double Encoder	×	0.36	0.43	35.2	1374.2	0.58	79.8
MOT	×	0.45	0.52	38.5	1520.0	0.65	82.5

4 Experiment

4.1 Experimental Setup

Datasets. We use open-source datasets for both image generation and understanding. The generation data include ShareGPT-4o-Image (Chen et al., 2025b), BLIP3-o (Chen et al., 2025a), and OpenSora (Zheng et al., 2024); the understanding data include LLaVA-OneVision-1.0 (Li et al., 2024) and Mammoth-VL (Guo et al., 2024). We sample 1.5M examples for each task. Because the public data do not match the quality of the original Janus-Pro training data, we additionally include 200K internal samples when post-training Janus-Pro.

Implementation Details. For Emu3, images are proportionally resized to approximately 720×720 for both tasks, and understanding and generation data are sampled at a 1:1 ratio. Training the 8B model with DeepSpeed ZeRO-3 takes about 10 days on 64 NVIDIA H800 (80GB) GPUs. For Janus-Pro, images are resized to approximately 384×384 and the two tasks are also sampled at a 1:1 ratio. Post-training the 7B model with FSDP takes about one day on 64 NVIDIA H800 (80GB) GPUs.

4.2 Evaluation on the Benchmarks

Multimodal Understanding. We evaluate on six widely used benchmarks: MME (Fu et al., 2023), MMBench (1.0-EN) (Liu et al., 2024), MMVet (Yu et al., 2023), MMMU (Yue et al., 2024), POPE (Li et al., 2023), and MMVP (Tong et al., 2024). As shown in Table 3, AIA improves both Emu3 and Janus-Pro across the evaluated understanding metrics.

Table 5: Ablations on data quality, AIA design, reference-pattern selection, data sampling, and λ.

Method	Image Understanding						Image Generation	
	MMMU	MMBench	MMVP	MMVet	POPE	MME-P	GenEval	DPG
Data Quality and Effectiveness of AIA								
Emu3 + AIA (Final)	**35.7**	**64.8**	**10.8**	**18.7**	**82.7**	**1084**	**0.67**	**81.20**
w/o AIA (baseline)	31.6	61.4	8.7	15.1	77.3	910	0.60	79.24
Janus-Pro + AIA (Final)	**42.1**	**75.6**	**48.0**	**51.2**	**89.8**	**1656.4**	**0.81**	**84.49**
w/o stage-level intensity	40.2	67.4	43.1	44.0	87.6	1543.2	0.79	83.79
w/o Huber	41.2	73.1	47.6	47.3	88.2	1613.9	0.80	84.25
w/ random AIA	33.5	61.4	41.3	38.6	83.6	1459.2	0.73	81.79
w/o AIA (baseline)	40.7	71.5	47.3	49.2	88.1	1593.1	0.80	84.19
AIA vs. Knowledge Distillation (Janus-Pro)								
AIA (ours)	**42.1**	**75.6**	**48.0**	**51.2**	**89.8**	**1656.4**	**0.81**	**84.49**
Feature-level	34.2	62.8	42.7	40.3	85.5	1485.4	0.75	82.19
Output-level	40.2	68.4	47.3	46.9	87.7	1603.7	0.79	83.97
Reference Pattern Selection (Janus-Pro)								
Qwen3-VL+FLUX	40.2	72.1	45.5	46.0	87.5	1555.1	0.76	83.56
Qwen3-VL+SimpleAR	40.3	72.5	44.1	46.6	88.0	1546.7	0.75	82.89
Qwen3-VL+HunyuanImage-3.0	**42.1**	**75.6**	**48.0**	**51.2**	**89.8**	**1656.4**	**0.81**	**84.49**
Qwen3-VL+Qwen-Image	41.5	74.3	47.3	49.6	89.5	1623.4	0.80	84.12
λ *Selection (Janus-Pro)*								
NTP:AIA: 1:1	37.2	62.9	44.5	44.0	87.3	1498.9	0.77	83.35
NTP:AIA: 10:1	40.5	71.1	46.6	48.2	87.2	1545.6	0.80	83.26
NTP:AIA: 50:1	**42.1**	**75.6**	**48.0**	**51.2**	**89.8**	**1656.4**	**0.81**	**84.49**
NTP:AIA: 100:1	41.3	71.3	47.3	50.5	87.9	1595.2	0.80	84.11

Text-to-Image Generation. Following Janus-Pro (Chen et al., 2025c), we report GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024). Table 3 shows that AIA also improves generation performance for both backbones, without introducing additional architectural decoupling.

Interleaved Cross-task Evaluation. Beyond evaluating understanding and generation independently, we report RealUnify's "Understanding aids Generation" metric (Shi et al., 2025) as a targeted interleaved cross-task evaluation. This metric measures one direction of cross-task information flow—using understanding information to support subsequent generation—so we use it only as a targeted interleaved setting rather than as a general measure of cross-task interaction. As shown below, the result provides additional evidence that the gains from AIA extend beyond isolated understanding and generation benchmarks to a setting that requires the two capabilities to interact.

Metric	Original Janus-Pro	Post-train w/o AIA	Post-train w AIA
Understanding aids Generation	25.2	25.4	**29.6**

4.3 Controlled Study of Architectural Decoupling

Our analysis in Fig. 1 suggests that stronger architectural decoupling is accompanied by increasingly reference-like task-specific interaction. Since the compared UMMs differ in more than architecture, we further conduct a controlled study to isolate the effect of decoupling. We construct variants with the Qwen3-1.7B (Yang et al., 2025), internal training data, and optimization recipe (**from scratch**), while varying only the degree of architectural decoupling. Due to the limited training compute, we conduct this controlled analysis on a smaller model as a targeted validation rather than reproducing the full-scale training setting. For the interaction similarity, we normalize layer positions by model depth and linearly interpolate the model and reference interaction curves to a shared length. We then report their Pearson correlation, denoted as Corr.(Und., Ref.) and Corr.(Gen., Ref.).

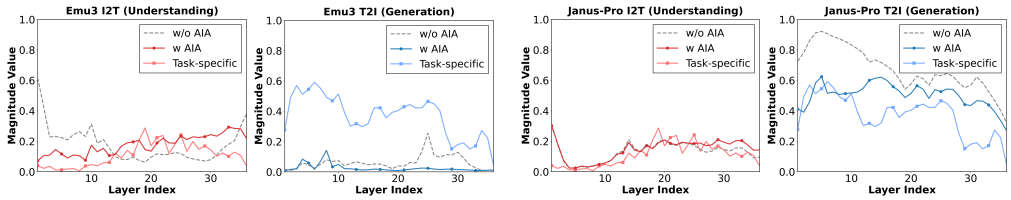

Figure 3: Cross-modal interaction patterns before and after AIA training. References: Qwen3-VL-8B (understanding) and HunyuanImage-3.0 (generation). Understanding is red; generation is blue.

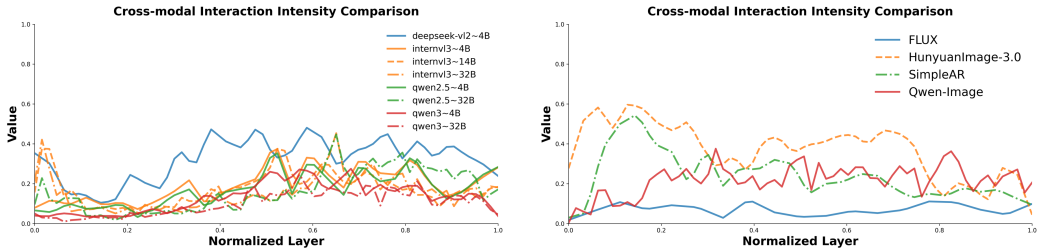

Figure 4: Cross-modal interaction patterns of different reference models.

As shown in Table 4, increasing architectural decoupling consistently moves both understanding and generation interaction patterns closer to their corresponding task-specific references, together with improved downstream performance. This controlled result confirms that the trend observed across existing UMMs is **not merely caused by differences in training data or optimization.**

4.4 Ablation Study

Because Emu3 is substantially slower to train, we conduct most ablations on Janus-Pro. This also provides a stricter test because AIA is applied only after the model has completed SFT.

Data Quality Analysis. The original Janus-Pro training data are not publicly available, so we first isolate the effect of continued post-training from the effect of AIA. As shown by the "w/o AIA" baseline in Table 5, post-training Janus-Pro on the same additional data does not reproduce the gains obtained with AIA, particularly on the understanding benchmarks. We therefore use this model as the controlled Janus-Pro baseline for all AIA comparisons. For Emu3, the "w/o AIA" result corresponds to our unified SFT setting and should not be directly compared with the task-specific SFT results reported in the original Emu3 paper.

Effectiveness of AIA. Table 5 shows that AIA improves both understanding and generation over the corresponding no-AIA baseline, while Fig. 3 shows that it moves each task's interaction pattern toward its selected reference. Importantly, the two tasks remain distinct after AIA training, indicating that AIA guides task-specific interaction rather than making their behaviors more similar. For Emu3, especially on generation, AIA better captures the reference trend than its absolute intensity. **Random interaction targets substantially degrade performance, ruling out gains from generic attention regularization.** The controlled study in Table 4 further shows that AIA consistently improves both reference-pattern similarity and downstream performance across different degrees of architectural decoupling.

AIA Design. We first remove the stage-level formulation and use finer-grained interaction targets. The resulting performance drop supports our motivation that exact layer-wise correspondence between reference and student models is unnecessarily restrictive, and that transferring the coarse interaction structure is more effective. We then remove the Huber formulation, which also leads to consistent degradation. This indicates that even within the desired interaction stage, individual layers benefit from retaining flexibility to adapt their interaction behavior during joint understanding–generation training. Together, these results validate the two complementary principles of AIA: coarse cross-model interaction alignment and flexible intra-stage adaptation.

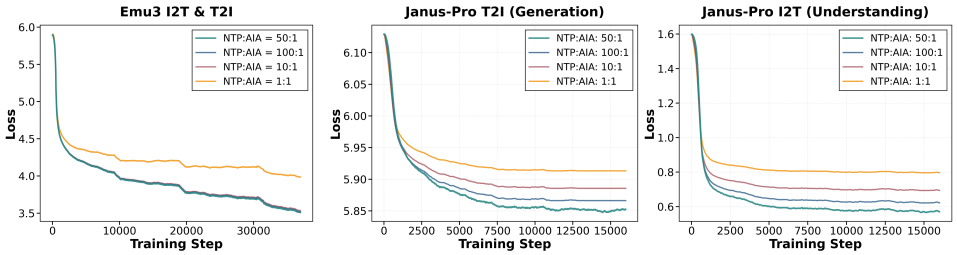

Figure 5: Training NTP loss curves of Emu3 and Janus-Pro under different AIA coefficients. The plotted curves omit the AIA term and reflect the main next-token-prediction objective.

Performance Gain From Knowledge Distillation or AIA? Because AIA uses statistics extracted from reference models, we compare it with feature- and output-level distillation to determine whether the gain comes from generic teacher transfer or specifically from interaction-pattern guidance. Table 5 shows that the tested feature- and output-level distillation baselines do not provide the same consistent gains. Among the compared distillation baselines, AIA is the only approach that improves both understanding and generation, supporting the role of the interaction pattern itself as the useful supervision signal in our setting.

Reference Interaction Pattern Selection. To identify suitable interaction patterns for AIA, we compare Deepseek-VL2 (Wu et al., 2024d), the InternVL (Chen et al., 2024; Zhu et al., 2025) series, and the Qwen (Bai et al., 2025b;a) series for understanding, together with FLUX (Labs, 2024), SimpleAR (Wang et al., 2025b), Qwen-Image (Wu et al., 2025a), and HunyuanImage-3.0 (Cao et al., 2025) for generation. Fig 4 shows that understanding patterns are highly consistent across model families, whereas generation patterns vary more substantially. We therefore use Qwen3-VL-8B as the understanding reference and compare four alternative generation references in Table 5. HunyuanImage-3.0 gives the strongest downstream result, while SimpleAR and FLUX are less effective. This indicates that the choice of reference pattern matters, especially for generation, and that the best target is not determined only by whether the underlying attention is causal or bidirectional. HunyuanImage-3.0 should not be viewed as an optimal target by definition; it is simply the strongest reference among the models tested here.

λ **Selection.** We vary the AIA coefficient during Janus-Pro post-training. In Table 5, NTP:AIA denotes the observed loss-magnitude ratio $\mathcal{L}_{\text{NTP}} : \mathcal{L}_{\text{AIA}}$, not the coefficient ratio. Janus-Pro is sensitive to the strength of the added AIA term, consistent with Fig 5. The λ selected on Janus-Pro also works well for Emu3, suggesting that the same coefficient transfers reasonably across the two settings.

Generality across Model and Data Scales. Our main experiments use the 7B Janus-Pro model with 3M training samples. Appendix Table A1 extends the comparison to 1B and 7B models and to 600K and 3M data scales. AIA improves over the corresponding baseline in all four settings, showing that the effect is not limited to one model or data scale.

The Appendix further discusses reference-pattern selection's implications for unified-model design and the right way of UMMs in the future.

5 Conclusion and Limitations

We revisit architectural decoupling in UMMs through layer-wise cross-modal attention interaction and reveal that stronger decoupling is accompanied by increasingly task-specific, reference-like interaction patterns. Motivated by this finding, we propose **AIA**, which combines stage-level interaction alignment with a Huber loss to transfer coarse task-specific interaction structure while preserving layer-wise flexibility. AIA consistently improves both understanding and generation on Emu3 and Janus-Pro, together with gains in interleaved cross-task evaluation. These results suggest that strong unified multimodal performance may rely less on forcing heterogeneous tasks toward a shared interaction regime, and more on developing an appropriate interaction behavior.

Limitations. Our controlled architectural study is limited to a relatively small-scale setting. Extending these findings to larger and more diverse UMMs (pixel-based) remains future work.

AI USE STATEMENT

We used generative AI tools only to refine the writing and improve the clarity and readability of the manuscript. All content was reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

Section 3 specifies the cross-modal interaction intensity, AIA objective, and training procedure. Section 4 reports the datasets, model and training settings, and evaluation benchmarks. The appendix analyzes model and data scales, reference-pattern selection, and unified-model design.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

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

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

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.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

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 *ICML*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025.

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

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, et al. Shotbench: Expert-level cinematic understanding in vision-language models. In *NeurIPS*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*, 2025.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. In *NeurIPS*, 2025.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, 2025.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*, 2025.

Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.

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

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. Sparsemm: Head sparsity emerges from visual concept responses in mllms. *arXiv preprint arXiv:2506.05344*, 2025a.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025b.

Junke Wang, Xiao Wang, Jiacheng Pan, Xuefeng Hu, Feng Li, Jingxiang Sun, Chaorui Deng, Zilong Chen, Yunpeng Chen, Kaibin Tian, et al. Arm: An autoregressive large multimodal model with unified discrete representations. *arXiv preprint arXiv:2606.11188*, 2026.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024a.

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*, 2025b.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *IJCV*, 2024b.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024c.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024d.

Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025a.

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 *ICLR*, 2025b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *NeurIPS*, 2025c.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. *arXiv preprint arXiv:2512.13687*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

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 *CVPR*, 2024.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

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. *arXiv preprint arXiv:2408.11039*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Table A1: Generality analysis. We compare the performance of the baseline and our method across different model sizes (1B, 7B) and training data scales (600k, 3M).

Data Size	Method	Image Understanding						Image Generation	
		MMMU	MMBench	MMVP	MMVet	POPE	MME-P	GenEval	DPG
Data-600k									
1B Model	Janus-Pro	36.1	61.2	43.1	37.4	86.5	1448.0	0.72	82.93
	w AIA	**37.2**	**65.6**	**43.8**	**38.5**	**87.1**	**1467.4**	**0.73**	**83.21**
7B Model	Janus-Pro	40.1	70.4	47.3	48.4	87.5	1587.4	0.80	84.06
	w AIA	**41.1**	**73.7**	**48.7**	**49.0**	**88.9**	**1637.8**	**0.80**	**84.78**
Data-3M									
1B Model	Janus-Pro	36.6	62.1	42.4	38.9	86.7	1424.0	0.73	83.01
	w AIA	**37.7**	**67.4**	**43.8**	**39.4**	**87.6**	**1487**	**0.75**	**83.67**
7B Model	Janus-Pro	40.7	71.5	47.3	49.2	88.1	1593.1	0.80	84.19
	w AIA	**42.1**	**75.6**	**48.0**	**49.8**	**89.8**	**1656.4**	**0.81**	**84.49**

Table A2: Composition of the samples used for cross-modal interaction analysis. We use 1,000 generation samples and 1,000 understanding samples, with generation prompts grouped by prompt length and understanding samples grouped by task type.

Task	Category	Samples	Total
Generation	1–20 tokens	330	1000
	20–50 tokens	268	
	50–100 tokens	193	
	>100 tokens	209	
Understanding	VQA	400	1000
	Captioning	300	
	Mathematical Reasoning	300	

A Discussions

Interaction Pattern Differences Across Reference Models. As shown in Fig 4, understanding models from different families and scales exhibit similar cross-modal interaction patterns, consistent with their broadly shared autoregressive architecture. Generation models are more diverse because they span different modeling paradigms. Models with bidirectional attention, including FLUX, Qwen-Image, and HunyuanImage-3.0, maintain relatively stable cross-modal interaction over many layers, whereas the causal SimpleAR model progressively reduces attention to text as image tokens are generated. More importantly, understanding and generation remain different even when they use similar attention mechanisms. This supports our main premise that task-dependent interaction is not merely an artifact of one particular UMM architecture.

What is the Right Path toward Unified Models? Across all architectures we study, understanding and generation retain distinct cross-modal interaction patterns. The same difference is also visible when the two tasks are trained independently under similar autoregressive formulations, suggesting that these task-dependent interaction differences reflect their distinct representational requirements rather than being an artifact of one particular unified architecture. Architectural decoupling is effective in practice, but our analysis shows that these differences remain as decoupling becomes stronger. Instead, stronger decoupling is associated with interaction patterns that are closer to strong task-dependent references. This suggests that handling the understanding–generation conflict may not require forcing the two tasks toward a shared interaction regime; a more direct approach is to let the shared model accommodate distinct but appropriate interaction behaviors for different tasks. AIA is a concrete instance of this idea, combining coarse interaction alignment with flexible layer-wise adaptation while keeping the architecture unchanged.

A complementary direction is to reduce the amount of fine-grained reconstruction handled by the shared backbone. Understanding mainly requires high-level semantic information, whereas generation must recover much finer visual detail (Yu et al., 2025). Even with a unified tokenizer (Tang et al., 2025; Qu et al., 2025; Yao et al., 2025; Ma et al., 2025; Lu et al., 2025), this difference remains. One possible design is therefore to let the unified model predict a more abstract visual representation and use a diffusion head to recover fine details, as also motivated by RecA (Xie et al., 2025a). Such a design could reduce the reconstruction burden on the shared backbone while preserving unified high-level processing.