Architecture Decoupling Is Not All You Need For Unified Multimodal Model

Dian Zheng[1,*] Manyuan Zhang[2,†] Hongyu Li[2] Kai Zou[3] Hongbo Liu[4] Ziyu Guo[2]
Kaituo Feng[2] Yexin Liu[2] Ying Luo[2] Yan Feng[2] Peng Pei[2] Xunliang Cai[2] Hongsheng Li[1,✉]

[1]CUHK MMLab   [2]Meituan
[3]University of Science and Technology of China   [4]Tongji University

https://zhengdian1.github.io/AIA-project/

Abstract

Unified multimodal models for image generation and understanding represent a significant step toward AGI and have attracted widespread attention from researchers. The main challenge of this task lies in the difficulty in establishing an optimal training paradigm due to inherent conflicting targets in understanding and generation tasks. To alleviate these conflicts and pursue higher performance, many researchers adopt varying degrees of model decoupling (e.g., Double image encoders, MOE/MOT architecture, or frozen MLLM). However, excessive model decoupling can lead to the loss of interleave generation ability, undermining the original intent of unified models. In this work, we aim to explore how to mitigate task conflicts without resorting to model decoupling. Firstly, we analyze why decoupling alleviates conflicts by studying the cross-modal attention behavior of models. We observe that model decoupling essentially drives models toward task-specific multimodal interaction patterns, as seen in Qwen-VL and HunyuanImage, and that the more thorough the decoupling, the more consistent the behavior becomes. Motivated by this observation, we propose Attention Interaction Alignment (AIA) loss, which explicitly learns Task-Specific multimodal interaction patterns during training. To demonstrate the generalizability of our AIA loss, we apply it to Emu3 and Janus-Pro during SFT and post-training stage respectively. Without bells and whistles, AIA not only refines cross-modal attention patterns, but also boosts both generation and understanding performance.

1. Introduction

Unified Multimodal Model (UMM) is trained to perform two distinct tasks (*e.g.*, visual generation and understand-

---

ing) within a single network, aiming to enhance interpretability by visualizing intermediate processes (interleaved generation), while also improving single-task performance. This approach represents a significant step toward general artificial intelligence.

Although the original intention behind UMM is admirable, practical realities are harsh: visual understanding and generation tasks require distinct feature granularities and representations at different network layers. Early research [33, 39, 43] attempted fully unified architectures (*e.g.*, sharing the image encoder and base model), but the results lagged significantly behind single-task approaches. To address task conflicts and boost performance, some [5, 7, 20, 42] have begun to decouple model components to varying degrees, achieving promising results. Due to the inherent effectiveness of this strategy, more researchers are now pursuing decoupled architectures. However, this trend overlooks the core motivation of UMM [54]: *leveraging the unified model's capacity for cross-modal reasoning to enhance single-task performance.* Excessive decoupling risks losing this synergistic benefit, limiting the model's capability to transfer knowledge and generalize across tasks. Furthermore, architecture decoupling, such as using double image encoders, forces the cross-modal reasoning process to undergo additional decode-encode steps, which is inelegant, time-costing and can result in information loss.

To maintain the original intent of UMM while narrowing its performance gap with decoupled models, we first conducted a detailed analysis of the underlying causes of this gap. Since the core of interaction between generation and understanding tasks lies in cross-modal information exchange, we focused our investigation on cross-modal interaction patterns (see Sec 3.1 for details). As illustrated in Fig 1, we first observe that regardless of architecture decoupling degrees, *the two tasks show consistent negative correlation in cross-modal interaction patterns within each layer.* We further verified that this phenomenon is indepen-
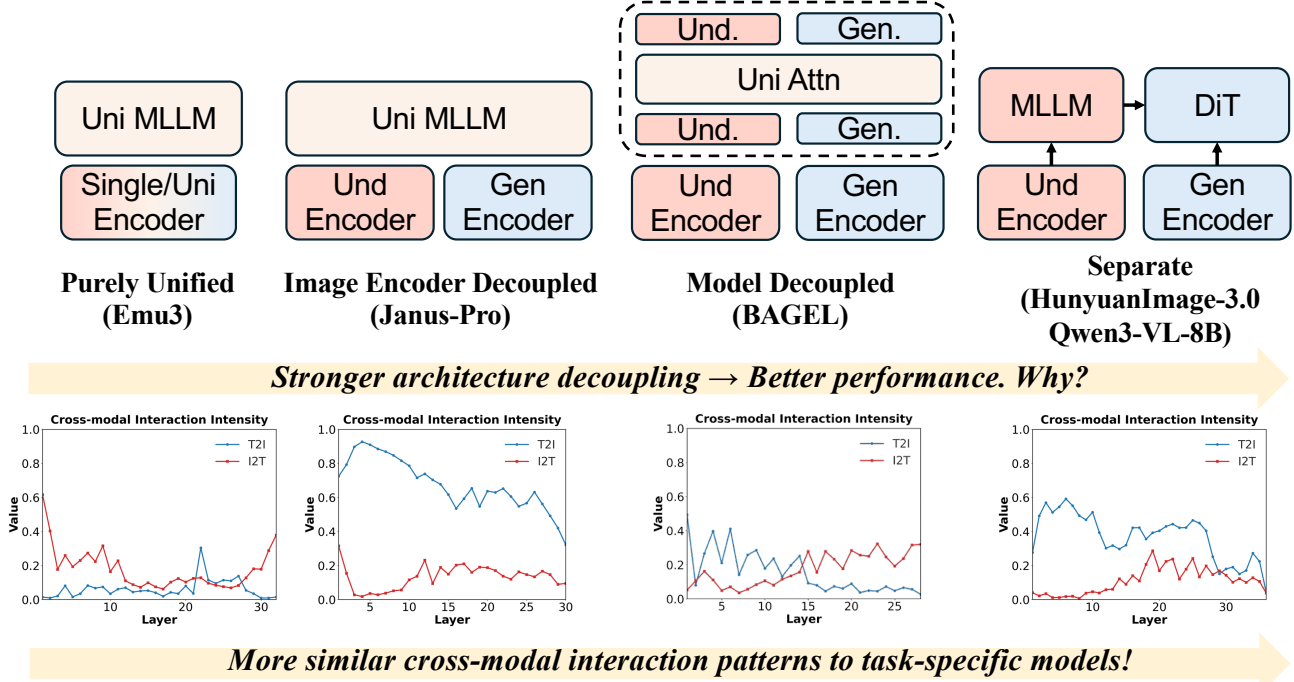
Figure 1. Various architectures of UMMs and its corresponding cross-modal interaction patterns. We arrange the models in order of increasing architecture decoupling. The row below illustrates the layer-wise cross-modal interaction intensity, with generation tasks shown in blue and understanding tasks in red; higher values indicate stronger interaction. The last column corresponds to HunyuanImage-3.0 and Qwen3-VL-8B, representing the interaction behavior of current SOTA task-specific generation and understanding methods. We observe that as decoupling increases, the negative correlation in cross-modal interaction patterns between understanding and generation tasks persists, but these patterns increasingly resemble those of task-specific models, leading to improved performance.

dent of input type or length; rather, the model dynamically allocates cross-modal representational weights within layers based on task requirements. Moreover, as decoupling increases, the interaction patterns increasingly resemble those of single-task models. This suggests that *existing decoupled models do not eliminate the inherent conflict between tasks; instead, they make each task behave more like its single-task counterpart*, resulting in performance improvement.

Based on this observation, we propose Attention Interaction Alignment (AIA) loss, which will explicitly constrain the layer-wise cross-modal interaction intensity during training without architecture decoupling. Specifically, we first select models suitable as cross-modal interaction learning target. The chosen models must demonstrate both high performance and autoregressive capability to meet the requirements of UMM. For the understanding task, we use Qwen3-VL-8B. For generation, since SOTA methods rely on external diffusion heads and purely autoregressive approaches are less effective, we refer to HunyuanImage-3.0 [2], which combines an autoregressive backbone with a diffusion head. Although its generative ability primarily depends on the diffusion head, its attention patterns remain informative. For both understanding and generation tasks, we use 100 samples to extract the attention patterns at each

layer and compute their average. Recognizing that attention patterns are closely linked to model architecture and pretraining, we apply the Huber loss to relax layer-wise attention constraints, enabling more flexible allocation of attention weights.

To validate the effectiveness of our proposed AIA loss, we adopt it to two degrees of decoupling methods (*e.g.*, Emu3 for purely unified model, Janus-Pro for slight decoupling model). Experimental results demonstrate that our approach enhances both generation and understanding performance while narrowing the gap with more strongly decoupled methods. Although we can not surpass the performance of highly decoupled methods yet, we believe that it is a step to the real UMMs and our method will be fully open-sourced to help the development of the community. We highlight the main contributions of this paper below:

- We provide the first mechanistic analysis of unified models with different decoupling architectures through cross-modal attention interaction intensity, revealing that decoupling does not resolve task conflicts but merely shifts attention patterns toward single-task behaviors.
- We introduce AIA loss, a simple regularization to explicitly shift the cross-modal attention interaction patterns into the behavior of task-specific models without requir-
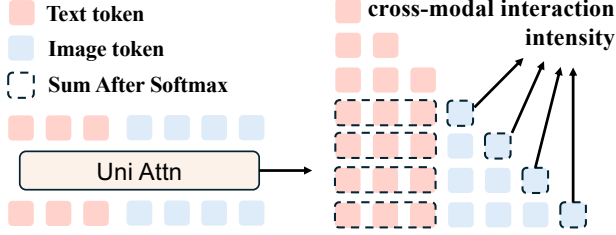
Figure 2. The pipeline of cross-modal interaction intensity calculation. We take text-to-image as an example, for each row, we compute the sum of all text token values, then average across all image tokens to obtain the intensity.

- ing architecture decoupling.
- Our method achieves significant improvements on widely-used generation and understanding benchmarks for both Emu3 and Janus-Pro.

2. Related Works

2.1. Uunified Multimodal Models (UMMs)

In recent years, with the development of LLM and MLLM [1, 8, 19, 23, 24, 34, 48], autoregressive architectures have been thoroughly explored, demonstrating exceptional capabilities in large models. Researchers are now considering whether MLLMs can be integrated with image generation to form a unified model, enabling automatic interleaved reasoning at the latent level. Initial unified MLLMs, such as Liquid [43], Emu3 [39], and Chameleon [33], adhered strictly to this path but used VAE [9, 17, 36] as the image encoder, limiting their understanding capabilities. To address tokenizer representation conflicts, methods like UniLip [32] and TokenFlow [29] have constructed a unified tokenizer compatible with both understanding and generation (coarse-grained and fine-grained) feature requirements. The Janus series [5, 26, 40] and show-o [46, 47] series have attempted to decouple the image encoder and training objectives for generation and understanding, achieving further performance improvements. However, due to conflicts between understanding and generation in the backbone network, performance remains constrained. To alleviate task conflicts, approaches like BAGEL [7] and OneCat [20] have explored partially decoupled architectures (MOT, MOE), yielding promising results. Meanwhile, another group of researchers argues that generation hardly aids understanding, so Metaquery [27], OmniGen2 [42], UniWorld-V1 [22], and Blip3-o [4] have chosen to fix the MLLM and solely optimize the diffusion head, achieving current state-of-the-art results. However, as model architectures become increasingly decoupled, starting from the image encoder, models can no longer achieve automatic interleaved reasoning in the latent space but must undergo an inelegant decode-encode process. Furthermore, as the backbone network becomes de-

coupled, the model's ability to handle understanding and generation in a unified manner is further weakened, let alone completely fixing the MLLM, which has deviated from the original intent of UMM. This paper aims to explore why model decoupling can alleviate conflicts and enhance performance while attempting to achieve similar effects within a fully unified architecture.

2.2. Ultra Task-Specific Models

Through extensive technological iterations, understanding tasks have developed into a nearly fixed model architecture. Models like Qwen [1, 34, 48] and the Llava [8, 19] series employ a semantic encoder combined with an autoregressive architecture, achieving outstanding results. In contrast, generation tasks lack a fixed model architecture. Early models such as SDXL [28], SD3 [10], and the FLUX [18] series utilized CLIP [30] as the text encoder within a pure diffusion [15, 31] framework, achieving high aesthetic quality but showing some limitations in instruction compliance. SimpleAR [38] explored image generation using LLM as the base model within a purely autoregressive architecture, but the inherent information loss in discrete representations resulted in images with a pronounced blur. Qwen-Image [41], and HunyuanImage [2] combined the strengths of both approaches by replacing CLIP with MLLM and integrating a diffusion head after the MLLM, achieving excellent results in both instruction compliance and aesthetic quality. In this paper, we select the most performant models in understanding and generation tasks as references for cross-modal interaction patterns, thereby enhancing the reliability of the conclusions.

3. Attention Interaction Alignment

In this section, we first analyze that why model decoupling will alleviate the task conflicts and improve the performance based on cross-modal attention interaction pattern. We observe that regardless of architecture decoupling degrees, different tasks induce mutually exclusive cross-modal attention patterns across various layers, but push the pattern into task-specific model types. Based on this, we propose Attention Interaction Alignment loss to constrain the attention pattern based on the task-specific one during training.

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

Method. We take text-to-image generation as an example. As shown in Fig 2, after obtaining the attention map from

Table 1. Standard deviation across 100 samples for each model.

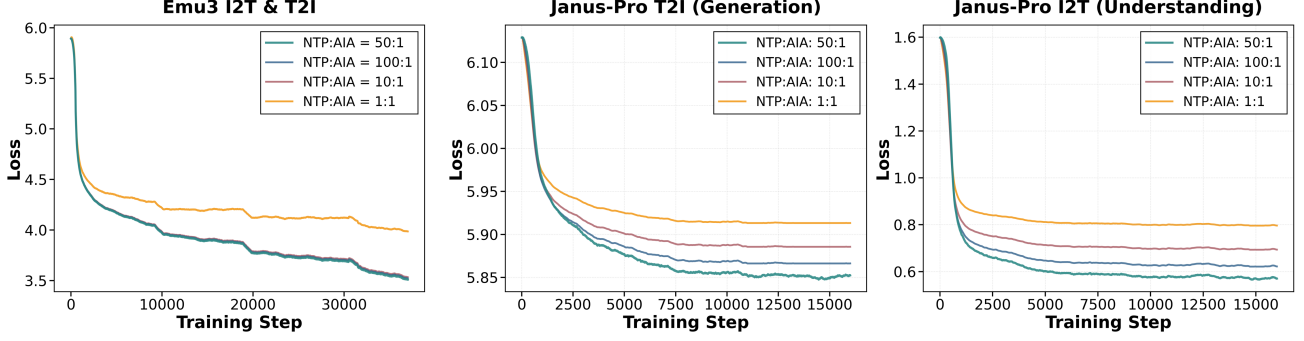| Method	Emu3	Janus-Pro	BAGEL	Task-Specific
Std	0.13	0.02	0.03	0.1

Figure 3. Training loss curve of Emu3 and Janus-Pro under various AIA coefficient. NTP and AIA means next-token-prediction and attention interaction alignment loss respectively. Note that we only show the NTP loss curve (excluding AIA loss) as it serves as the primary indicator of final performance and **the periodic drops in the Emu3 loss curve are due to learning rate schedule**.

the softmax operation, we calculate the sum of all text tokens in each row of the attention map. Then, we compute the average over all layers, all image tokens and all attention heads. This value represents the cross-modal interaction intensity between images and text for the given sample. The specific formula is as follows:

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

where I means the intensity, l is the layer index, N, H, Q, K represent token, head, query, key numbers respectively. The average results of 100 samples are shown in Fig 1. In Table 1, we further validate that the attention pattern of each model is independent of input length and type (*i.e.*, consisting of prompts like text rendering, short and dense caption and questions like caption, choices, *etc.*) by computing the standard deviation for each model.

Rationality of task-specific attention pattern. Fig 1 shows that Qwen3-VL consistently exhibits low attention to image tokens, which aligns with the motivation behind token pruning methods [37, 51] in current understanding tasks. For generation, HunyuanImage maintains around 40% attention to text tokens in the first 80 layers, with a sharp decline in the final layers. This pattern matches the common consensus in generative models: shallow layers focus on building semantic representations and thus attend more to high-level text features, while the final layers shift toward pixel-level image features, resulting in reduced attention to text. The decrease in text attention in deeper layers further supports the claim in RecA [45] that textual information is insufficient at these stages. Except for Emu3 [39], all other architectures follow the single-task trend, indicating that unified models tend to learn in a task-specific manner once their architectures are decoupled.

Why model decoupling improves performance. From Table 1, we can first rule out the influence of input properties on cross-modal attention patterns. Fig 1 shows that the interaction curves for the two tasks have negative correlation,

indicating that, after unified training, the inherent conflict between tasks forces the model to allocate attention weights to different tasks at different layers. This mechanism helps the model self-mitigate cross-task interference, and the resulting allocation pattern becomes fixed. Furthermore, Fig 1 demonstrates that this negative correlation persists regardless of the degree of model decoupling. As the model becomes more decoupled, the attention interaction patterns increasingly resemble those of task-specific models, leading to improved performance. Specifically, when performing understanding tasks, Janus-Pro exhibits relatively high attention to image tokens in the initial layers, which limits its understanding capability. For Emu3, both generation and understanding tasks deviate significantly from the typical single-task interaction patterns, highlighting the inherent difficulty of fully unified learning within an autoregressive architecture.

3.2. Attention Interaction Alignment Loss

Based on the observation above, we propose attention interaction alignment loss, which will constrain the attention patterns explicitly during training. Specifically, we use the layer-wise intensity from the task-specific models in Fig 1 as learning targets, termed T_l. However, since the attention curves are fixed throughout network training, it is unsuitable to apply overly strict constraints for supervision. Therefore, we divide the values into several sub-stages according to their magnitudes and further relax the absolute constraint on individual values using the Huber loss. Taking Emu3 [39] as an example, the specific formula is as follows:

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

$$(2)$$

where I_l denotes the cross-modal intensity at layer l, and for the layer-wise target boundary T_l and Huber threshold δ_l, please see the Supplementary File. The AIA loss will

Table 2. System-level comparison on widely used image understanding and generation benchmarks. † means the result is re-implemented by ourselves. Types represent whether the model uses diffusion, autoregressive, or masked prediction for training. (Gray) means the result reported in original papers while nobody can re-implemented.

Method	Params	Types	Image Understanding						Image Generation	
			MMMU	MMBench	MMVP	MMVet	POPE	MME-P	GenEval	DPG
Gen. Only										
SDXL [28]	-	Diff				-			0.55	74.65
SD3-medium [10]	2B	Diff				-			0.74	84.08
Infinity [14]	2B	VAR				-			0.73	83.50
Infinity [14]	8B	VAR				-			0.79	86.60
FLUX.1-dev [18]	12B	Diff				-			0.82	84.00
Emu3-Gen [39]	8B	AR				-			0.66	80.60
Qwen-Image [41]	7B+20B	AR+Diff				-			0.87	88.32
Und. Only										
Emu3-Chat [39]	8B	AR	31.6	58.5	36.6	37.2	85.2	1244	-	-
Qwen2.5-VL [1]	3B	AR	53.1	79.1	-	61.8	-	-	-	-
Qwen2.5-VL [1]	7B	AR	58.6	83.5	-	66.6	-	1685	-	-
InternVL2.5 [6]	8B	AR	56.2	84.6	-	62.8	90.6	-	-	-
InternVL3 [53]	8B	AR	62.7	83.4	-	81.3	91.1	-	-	-
Qwen3-VL	8B	AR	69.6	85.0	-	-	-	-	-	-
Uni. Frozen MLLM										
MetaQuery-XL [27]	7B+1.6B	AR+Diff	58.6	83.5	-	66.6	-	1685	0.80	82.05
Blip3-o [4]	7B+1.4B	AR+Diff	58.6	83.5	-	66.6	-	1685	0.84	81.60
UniWorld-V1 [22]	7B+12B	AR+Diff	58.6	83.5	-	67.1	-	1685	0.84	81.38
OmniGen2 [42]	3B+4B	AR+Diff	53.1	79.1	-	61.8	-	-	0.86	83.57
Uni. MoE/MoT Arch										
BAGEL [7]	7B+7B	AR+Diff	55.3	85.0	69.3	67.2	-	1687	0.88	85.07
OneCat [20]	3B+3B+3B	AR+VAR	41.9	78.8	-	52.2	-	1630	0.90	84.53
Uni. Double Image Encoders / Training Objectives										
Show-o [46]	1.3B	AR+Mask	26.7	-	-	-	80.0	1097	0.69	67.27
Show-o2 [47]	7B	AR+Diff	**48.9**	**79.3**	-	-	-	<u>1621</u>	0.76	**86.14**
Janus-Pro [5]	7B	AR	41.0	65.54 (79.2)	47.3	**50.0**	87.4	1567	<u>0.80</u>	84.19
Janus-Pro + AIA (Ours)	7B	AR	<u>42.1</u>	<u>75.6</u>	**48.0**	<u>49.8</u>	**89.8**	**1656**	**0.81**	<u>84.49</u>
Uni. Purely										
Chameleon [33]	7B	AR	28.4	35.7	-	8.3	-	-	0.39	-
VILA-U [44]	7B	AR	<u>32.2</u>	**66.6**	**22.0**	**27.7**	**83.9**	**1336**	0.39	72.48
Emu3 [39] †	8B	AR	31.6	61.4	8.7	15.1	77.3	910	<u>0.60</u>	<u>79.24</u>
Emu3 + AIA (Ours)	8B	AR	**35.7**	<u>64.8</u>	<u>10.8</u>	<u>18.7</u>	<u>82.7</u>	1084	**0.67**	**81.20**

combine with next-token-prediction loss, term $\mathcal{L}_{\text{NTP}}$ as:

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

where λ is set to 40 in the experiment by default.

3.3. Training Details Under Different Architectures

We incorporate the AIA loss into Emu3 during the supervised fine-tuning (SFT) stage and into Janus-Pro during the post-training stage to demonstrate the effectiveness of our method across different scenarios and explore the challenges of integrating our loss at various training phases.

Emu3. We load the pretrained (PT) weights of Emu3 and perform SFT training using our own data. Note that Emu3 only employs unified training at this stage; the results reported in the original paper were obtained through separate training. Since Emu3's performance at the pretraining stage is relatively poor, as shown in Fig 3, we incorporate the AIA loss with varying weights during the SFT stage and observe that the NTP loss convergence trends remain nearly identical. This indicates that, when the learning target is set appropriately, incorporating the AIA loss during SFT does not significantly affect the model's pretrained knowledge.

Janus-Pro. Since Janus-Pro only provides the final SFT weights, we perform post-training on this basis. When data quality does not differ significantly, this greatly increases the difficulty of tuning, as the model's distribution is already highly fixed at this stage (*i.e.*, a point further supported by the variance comparison between Emu3 and Janus-Pro in Table 1). However, this also aligns more closely with realistic settings: *how to perform fine-grained tuning when only the final weights are accessible and attention distribution adjustment is required*. As shown in Fig 3, the model is

Table 3. Quantitative comparison for the ablation study about the data quality, AIA loss, atention pattern, λ, and data sampling ratio.

Method	Image Understanding						Image Generation	
	MMMU	MMBench	MMVP	MMVet	POPE	MME-P	GenEval	DPG
Data Quality and The effectiveness of AIA Loss								
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**	**49.8**	**89.8**	**1656.4**	**0.81**	**84.49**
w/o stage-level intensity	40.2	67.4	43.3	44.0	87.6	1543.2	0.79	83.79
w/o Huber	41.2	73.1	47.6	47.7	88.2	1613.9	0.80	84.15
w/o AIA (baseline)	40.7	71.5	47.1	49.2	88.1	1593.1	0.80	84.15
Attention Pattern Selection (Janus-Pro)								
Qwen3-VL+FLUX	40.2	72.1	45.7	46.0	87.5	1555.1	0.76	82.89
Qwen3-VL+SimpleAR	41.5	74.3	47.9	48.6	89.5	1623.4	0.80	84.12
Qwen3-VL+HunyuanImage-3.0	**42.1**	**75.6**	**48.0**	**49.8**	**89.8**	**1656.4**	**0.81**	**84.49**
Qwen3-VL+Qwen-Image	40.3	72.5	44.3	46.6	88.0	1546.7	0.75	83.01
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	46.8	48.2	87.2	1545.6	0.80	83.26
NTP:AIA: 50:1	**42.1**	**75.6**	**48.0**	**49.8**	**89.8**	**1656.4**	**0.81**	**84.49**
NTP:AIA: 100:1	41.3	71.3	47.5	49.7	87.9	1595.2	0.80	84.11
Data Sampling Ratio (Janus-Pro)								
Gen:Und: 1:1	**42.1**	**75.6**	**48.0**	**49.8**	**89.8**	**1656.4**	**0.81**	**84.49**
Gen:Und: 2:1	41.8	74.2	46.9	49.3	89.1	1643.6	0.81	84.23
Gen:Und: 4:1	41.2	72.8	47.5	48.5	88.3	1621.4	0.80	83.87
Gen:Und: 1:2	42.0	74.8	47.7	49.3	89.3	1645.9	0.79	83.12

highly sensitive to the weight of the AIA loss at this stage, yet it can still achieve the desired effect when λ is set appropriately. We further validate this in Section 4.3.

4. Experiment

4.1. Experimental Setup

Datasets. We primarily use open-source image generation and understanding datasets for training, including ShareGPT-4V [3], BLIP3-o [4], and OpenSora [52] for generation, and LLaVA-OneVision-1.0 [19], Mammoth-VL [13] for understanding, resulting in 1.5M samples for each task. Since the quality of these datasets is not as high as that of Janus-Pro's data, we incorporate 200K internal samples to align the data quality.

Implementation Details. For Emu3, we proportionally resize images to approximately 720×720 resolution for both understanding and generation tasks. The understanding and generation data are balanced at a 1:1 ratio. Under the Deep-Speed ZeRO-3 framework, the entire training process for the 8B model took approximately 10 days on a cluster of 8 nodes, each equipped with 8 NVIDIA H800 (80GB) GPUs. For Janus-Pro, we proportionally resize images to approximately 384×384 resolution for both understanding and gen-

eration tasks. The understanding and generation data are balanced at a 1:1 ratio. Under the FSDP framework, the entire training process for the 7B model took approximately 1 day on a cluster of 8 nodes, each equipped with 8 NVIDIA H800 (80GB) GPUs.

4.2. Evaluation on the benchmarks

Multimodal Understanding. We evaluate our model on six widely recognized benchmarks—MME [11], MMBench (1.0-EN) [25], MMVet [49], MMMU [50], POPE [21], and MMVP [35]—which together form a compact yet thorough assessment framework encompassing perception, cognition, and multimodal reasoning, with robust capability to distinguish performance differences among leading models. The result is shown in Table 2, within each decoupling degree, we achieve state-of-the-art results under the same training configuration, and narrow the performance gap with models employing higher decoupling degrees or different training paradigms.

Text-to-Image Generation. We follow Janus-Pro [5] and report results on widely used image generation benchmarks Geneval [12] and DPG-Bench [16]. As shown in Table 2, Equipping with the AIA loss regulation, we improve the performance of Janus-Pro and Emu3 and narrow the gap
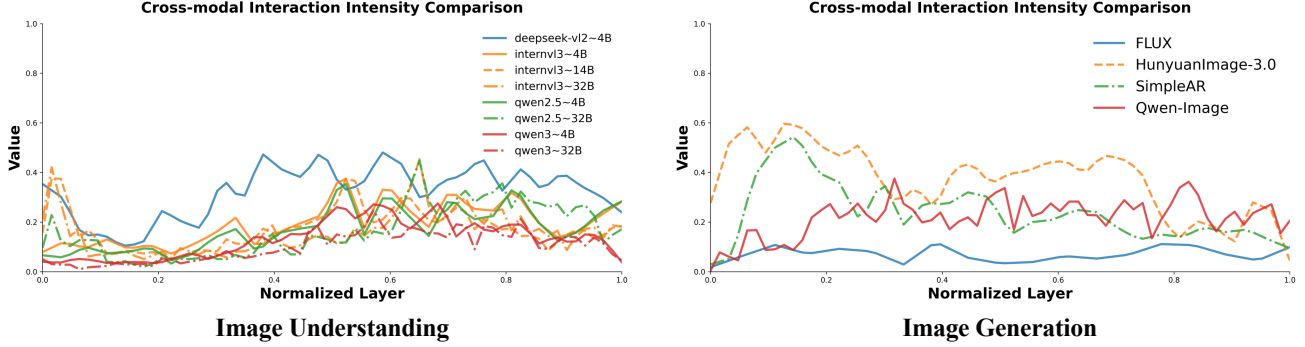
Figure 4. Cross-Modal Attention Patterns Visualization of Different Single-Task Models.

with models with higher decoupling degrees.

4.3. Ablation Study

Due to the slow training speed of Emu3, we primarily conduct ablation studies on Janus-Pro. As mentioned in Sec 3.3, this represents a more challenging setting, thereby better demonstrating the effectiveness of our method.

Data Quality Analysis. Since the training data of Janus-Pro is not publicly available, we first assess the quality of our own dataset to validate our method's effectiveness. As shown in Table 3 (*i.e.*, w/o AIA), further fine-tuning Janus-Pro with our data yields comparable performance to the original model in both understanding and generation, indicating that our data quality itself does not introduce performance gains. Note that the Emu3 results (w/o AIA) cannot be directly compared with those reported in the original Emu3 paper, as their performance was obtained through single-task SFT. The relatively low unified performance, particularly in understanding, highlights the inherent limitations of using a VAE as the image encoder and the challenges of purely unified training.

The Effectiveness of AIA loss. The result in Table 3 shows that incorporating our AIA loss leads to improved performance in both understanding and generation, with earlier integration (*e.g.*, during the Emu3 SFT stage) yielding greater improvements. Fig 5 further illustrates the changes in cross-modal attention patterns after applying the AIA loss. With AIA regularization, the attention patterns of both Emu3 and Janus-Pro shift closer to those of single-task models, confirming that aligning attention patterns toward task-specific behaviors indeed enhances model performance, and that the AIA loss effectively improves results while reshaping attention patterns. However, it could also be observed that when adding AIA loss to purly unified model architecture Emu3, the cross-modal attention pattern is more harder to change (*i.e.*, the generation attention pattern in Emu3 merely captures the correct directional trend across layers, but the actual values are still incorrect).

Then we validate the effectiveness of huber loss and

stage-level intensity. Both components are designed to relax overly strict attention constraints. As shown in the table 3, removing either leads to performance below the baseline, confirming that excessively rigid attention constraints hinder training. The Huber loss and stage-level intensity provide coarse optimization targets while allowing the model flexibility for self-adjustment, effectively addressing this issue and validating the effectiveness of AIA loss.

Task-specific Attention Patterns Selection. To identify the most suitable task-specific attention patterns as training targets for unified models, we compare Deepseek-VL2 [23], the InternVL [6, 53] series, and the Qwen [1] series for understanding tasks, and FLUX [18], SimpleAR [38], Qwen-Image [41] and HunyuanImage-3.0 [2] for generation tasks. Fig 4 illustrates the corresponding attention patterns of these models. We observe that attention patterns for understanding tasks remain highly consistent across different models, whereas those for generation tasks exhibit significant variations. Consequently, we select Qwen3-VL-8B as the target for understanding and combine it with attention patterns from four generation models. Table 3 shows that the attention pattern from HunyuanImage-3.0 achieves the best performance as a learning target, while that from Qwen-Image yields mediocre results. We attribute this to differences in attention mechanisms: Qwen-Image employs a fixed MLLM module, limiting us to extracting full attention patterns from its diffusion head, which are not well-suited for autoregressive models. However, we also argue that HunyuanImage-3.0 may not represent the optimal choice, as its training incorporates understanding tasks and undergoes reinforcement learning, making its attention patterns potentially suboptimal for the PT or SFT stages of unified model training.

λ **Selection.** Here, we analyze the coefficient of the AIA loss during fine-tuning on Janus-Pro. We show the results in Table 3, where NTP:AIA denotes the loss weight ratio between the two objectives. By modifying attention patterns in a model with well-established pretrained knowledge is highly sensitive to the loss coefficient (*i.e.*, also validated in
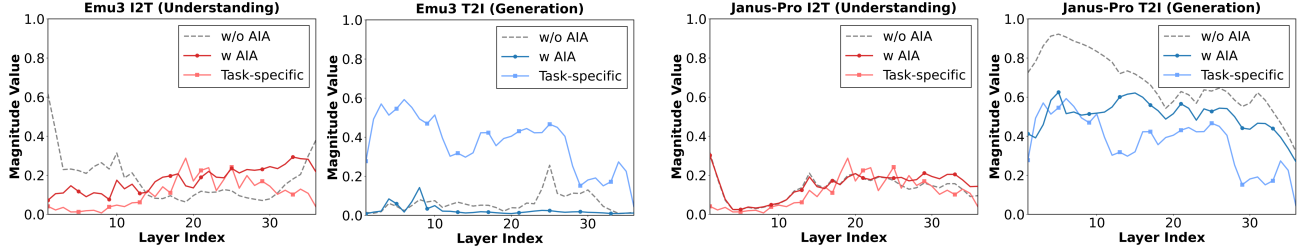
7

Figure 5. Visualization of cross-modal attention patterns modification after AIA training. Task-specific models are Qwen3-VL-8B for understanding and HunyuanImage-3.0 for generation, with understanding tasks shown in blue and generation tasks in red.

Fig 3). The coefficient must be carefully balanced—strong enough to influence model optimization, yet not so dominant as to disrupt the existing knowledge.

Is Data Sampling Ratios still Matter with AIA loss? Previous methods, beyond architecture decoupling, primarily mitigate task conflicts by adjusting the ratio of understanding and generation data. As concluded in BAGEL, understanding tasks converge faster than generation tasks, leading to a data distribution heavily skewed toward generation in later training stages. In this work, we alleviate task conflicts through the AIA loss and further investigate how this affects the model's sensitivity to data sampling ratios. As shown in Table 3, we find that a balanced 1:1 ratio achieves the best performance, contrary to the conventional high-generation, low-understanding distribution. This suggests that with the AIA loss, **understanding and generation tasks not only experience reduced conflict but also exhibit synergistic effects** (*i.e.*, training with both tasks outperforms using generation data alone). We believe this represents a significant step forward in unified model training.

4.4. Discussions

The Task-Specific Attention Patterns Difference in Various Models. As shown in Fig 4, for understanding tasks, models across different series and sizes exhibit nearly identical cross-modal attention patterns due to their shared autoregressive architecture, reflecting the maturity of current understanding architectures. In contrast, generation tasks present diverse patterns due to two distinct architectural paradigms: autoregressive with diffusion head and pure diffusion. Notably, models within the same architectural type display consistent attention patterns. We further observe that generation attention patterns closely follow their underlying attention mechanisms: models with bidirectional attention (FLUX, Qwen-Image) maintain consistent cross-modal interaction intensity across all layers, while those with causal attention (HunyuanImage-3.0, SimpleAR) progressively reduce attention to text as image tokens are generated—completely different from understanding tasks. Interestingly, even with the same attention mechanism, understanding and generation tasks exhibit distinct cross-modal patterns. This difference may hold the key to resolving task

conflicts in unified training.

What is the Right Path toward Unified Models? Through our investigation, we find that regardless of architecture decoupling strategies, generation and understanding tasks consistently exhibit mutually exclusive cross-modal interactions within the same network layers. This aligns with our earlier observation that even under the same autoregressive architecture, understanding and generation display distinct cross-modal patterns when trained independently, suggesting that the task conflicts transcend s any architecture design. However, this raises an intriguing question: *if models invariably learn task-exclusive interaction patterns during unified training—regardless of architecture choices—could this actually represent the correct behavior for unified models?* Despite the negative correlation between tasks, models can identify the current task through input sequences (image-text or text-image) and special tokens (*e.g.*, <img start>), and automatically adjust cross-modal interaction accordingly. With appropriate explicit guidance methods like our AIA, task conflicts may not be an issue to avoid, but rather a natural characteristic to manage.

An alternative path toward unification would be to remove all task-distinguishing cues (*i.e.*, adopt a unified tokenizer, eliminate task-related special tokens, and use interleaved data to obscure input formats), forcing the model to learn a truly unified space from the input alone. While this approach may resolve negative correlations between tasks, it would significantly increase training difficulty. We leave this exploration to future work.

5. Conclusion

In this work, we investigated the fundamental challenge of unified multimodal models training: the inherent conflict between image generation and understanding tasks. While existing approaches rely on model decoupling—such as dual image encoders, MOE/MOT architectures, or fixed MLLMs—to mitigate these conflicts, such strategies often sacrifice the interleaved generation capability that defines true unified models. Through systematic analysis of cross-modal attention behaviors, we revealed that model decoupling essentially guides models toward task-specific mul-

8

timodal interaction patterns, with more aggressive decoupling yielding stronger alignment to single-task behaviors. Building on this insight, we proposed Attention Interaction Alignment (AIA) loss, which explicitly learns task-specific interaction patterns without architecture modifications. We validated AIA on Emu3 during SFT and Janus-Pro during post-training, demonstrating that it not only refines cross-modal attention patterns but also consistently improves both generation and understanding performance across different training stages and architectures. This work represents a significant step toward achieving high-performance unified models while preserving their core capability of seamless cross-modal reasoning.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 5, 7

[2] 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. 2, 3, 7

[3] 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*, 2025. 6

[4] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 3, 5, 6

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Januspro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 1, 3, 5, 6

[6] 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. 5, 7

[7] 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. 1, 3, 5

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, 2024. 3

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[10] 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. 3, 5

[11] 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. 6

[12] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 6

[13] 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. 6

[14] 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. 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[16] 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. 6

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[18] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3, 5, 7

[19] 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. 3, 6

[20] 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. 1, 3, 5

[21] 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. 6

[22] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 3, 5

[23] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 3, 7

[24] 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. 3

[25] 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. 6

[26] 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. 3

[27] 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. 3, 5

[28] 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. 3, 5

[29] 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. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3

[32] 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. 3

[33] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 3, 5

[34] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 3

[35] 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. 6

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 3

[37] Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. Sparsemm: Head sparsity emerges from visual concept responses in mllms. *arXiv preprint arXiv:2506.05344*, 2025. 4

[38] 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*, 2025. 3, 7

[39] 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. 1, 3, 4, 5

[40] 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*, 2024. 3

[41] 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*, 2025. 3, 5, 7

[42] 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. 1, 3, 5

[43] 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*, 2024. 1, 3

[44] 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*, 2024. 5

[45] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025. 4

[46] 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*, 2025. 3, 5

[47] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *NeurIPS*, 2025. 3, 5

[48] 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. 3

[49] 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. 6

[50] 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. 6

[51] 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. 4

[52] 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. 6

[53] 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. 5, 7

[54] Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025. 1

Architecture Decoupling Is Not All You Need For Unified Multimodal Model

Supplementary Material

In this supplementary file, we provide comprehensive details of the layer-wise target boundary T_l and Huber threshold δ_l in Sec F.

F. Details on layer-wise target boundary T_l and Huber threshold δ_l

Here we show the detailed hyper-parameters of Huber loss in Emu3 and Janus-Pro in Table S4 and Table S5 respectively. We use nearly identical parameters for each task across both models to enhance reliability. Note that the slight differences in the layer ranges of l are due to the different number of layers in Emu3 and Janus-Pro.

Table S4. The hyper-parameter of Huber loss in Emu3.

Layer Range	Generation (δ_l, T_l)	Understanding (δ_l, T_l)
$0 \leq l < 10$	$(0.2, 0.4)$	$(0.05, 0.1)$
$10 \leq l < 20$	$(0.1, 0.4)$	$(0.05, 0.15)$
$20 \leq l < 25$	$(0.1, 0.4)$	$(0.05, 0.3)$
$25 \leq l \leq 30$	$(0.05, 0.2)$	$(0.05, 0.3)$
$l > 30$	$(0.05, 0.2)$	$(0.05, 0.2)$

Table S5. The hyper-parameter of Huber loss in Janus-Pro.

Layer Range	Generation (δ_l, T_l)	Understanding (δ_l, T_l)
$0 \leq l < 10$	$(0.2, 0.4)$	$(0.05, 0.1)$
$10 \leq l < 20$	$(0.1, 0.4)$	$(0.05, 0.15)$
$20 \leq l < 25$	$(0.1, 0.4)$	$(0.05, 0.3)$
$25 \leq l \leq 29$	$(0.05, 0.2)$	$(0.05, 0.3)$
$l > 29$	$(0.05, 0.2)$	$(0.05, 0.2)$