Generation as Auxiliary Supervision Enhancing Visual Understanding at Zero Overhead via Decoupled Embedding Prediction

Anonymous ACL submission

Abstract

While Multimodal Large Language Models (MLLMs) have achieved remarkable progress, visual understanding and generation are typically treated as divergent objectives. Existing unified frameworks relying on discrete visual tokenization or diffusion often require training from scratch, as their generative targets are misaligned with the continuous representations required for visual understanding. Consequently, they struggle to directly enhance existing pretrained MLLMs. In this work, we present **GAS**, a generation-guided training framework that reinterprets visual generation as an auxiliary forcing function for representation learning. Concretely, GAS adapts **Next Embedding Prediction (NEP)** as a cross-modal generation paradigm within a decoupled Mixture-of-Transformers (MoT) architecture. By maintaining a shared foundational trunk and parallel upper layers, GAS allows generation losses to enrich the shared visual pathway with finer spatial precision and stronger visual retention, thereby improving complex multimodal reasoning. To maximize this synergy, we further construct highly correlated generation tasks that demand deep cognitive grounding rather than simple synthesis. Crucially, because the auxiliary generation branch is discarded after training, the **GAS** model family substantially improves multi-image perception, spatial comprehension, and temporal understanding with **zero inference overhead**. Extensive analyses further validate the underlying mechanism and demonstrate the feasibility of generation-guided training as a practical route to stronger multimodal understanding.

1 Introduction

Multimodal large language models (MLLMs) have achieved remarkable progress in visual understanding tasks, including image captioning, visual question answering, and visual grounding (Li et al., 2024a; Bai et al., 2025a; Zhu et al., 2025). The prevailing paradigm encodes visual inputs through vision encoder and projects the resulting representations into token space of large language model (LLM), which then performs reasoning and generates textual responses in an autoregressive manner (Liu et al., 2023; Radford et al., 2021). Although this architecture has proven remarkably effective for semantic-level comprehension, its training objective – next-token prediction (NTP) over text – provides only indirect supervision for visual perception. The model is trained to talk about images, yet never to learn from them because no training signal supervises the model to internalize the visual content itself at the pixel or region level.

This text-only supervision introduces limitations from two complementary perspectives. On one hand, natural language inherently suffers from an **expressiveness bottleneck**: it cannot precisely encode fine-grained visual details such as exact spatial relationships, pixel-level object boundaries, physical interaction dynamics, or subtle appearance variations (Shao et al., 2024; Shiri et al., 2024). When the sole training signal is text prediction, the model's visual representations are shaped only by what language can articulate, leaving the richer geometric, structural, and relational information latent and underutilized. On the other hand, there exist abundant **visually intensive tasks** like segmentation and image editing. Existing VLM architectures are structurally unable to consume such data for training, thereby forfeiting a vast reservoir of fine-grained visual supervision that could otherwise cultivate deeper perception and complex visual understanding capabilities. Furthermore, empirical evidence reveals that visual information progressively attenuates in deeper LLM layers during reasoning (Zhang et al., 2024; Ganz et al., 2024), precisely because the text-centric objective provides no mechanism to encourage persistent retention of visual features throughout the

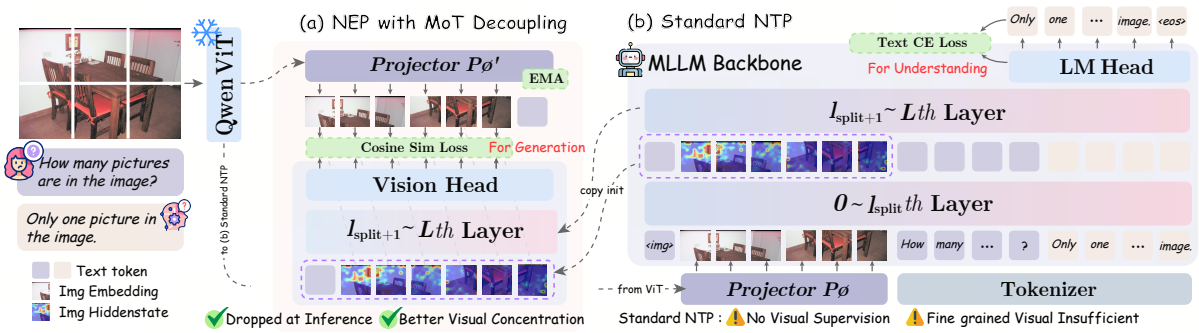


Figure 1: **Overview of our generation-guided training framework.** Given multimodal inputs, the understanding branch processes visual tokens from shared ViT encoder and predicts text via standard NTP paradigm. The generation branch, operating on independent MoT transformer layers, receives intermediate hidden states from the understanding backbone and predicts target image embeddings via Next Embedding Prediction (NEP). At inference time, the generation branch is entirely discarded.

inference chain.

A natural hypothesis emerges: if a model is additionally trained to predict visual content, like predicting pixel-level segmentation masks or synthesizing edited images, then its internal representations must necessarily encode richer and more precise visual information, which in turn should benefit understanding. This intuition has motivated the recent wave of unified multimodal models (UMMs) that jointly train understanding and generation (Team, 2024; Xie et al., 2025; Zhou et al., 2025; Deng et al., 2025). However, directly adopting the UMM paradigm to enhance understanding exposes two unresolved tensions:

First, current UMMs retain generation-side parameters at inference time (Deng et al., 2025; Shen et al., 2025), incurring latency and memory costs that hinder deployment. What's more, the generation objective in these systems is optimized for synthesis quality rather than for maximizing benefit to comprehension (Fu et al., 2026b), understanding performance often plateaus or even degrades relative to MLLMs of comparable scale, suggesting that simple joint training fails to transfer generative learning into understanding gains effectively.

Second, existing work typically reports aggregate benchmark numbers after introducing generation data, without dissecting the underlying mechanisms. Critical questions remain open: Do generation and understanding objectives inherently conflict? Which intermediate representations are most amenable to generative supervision for "feeding back" into understanding performance? Which types of generation tasks enhance which facets of comprehension? And through what pathways does generation training confer its advantages?

In this work, we propose a generation-guided training framework that resolves both tensions. Our core idea is to leverage generation tasks purely as training-time auxiliary supervision for understanding, rather than as an end goal. Architecturally, we adopt a Mixture-of-Transformers (MoT) design that introduces a dedicated generation branch with independent transformer parameters alongside the understanding branch, so that generation gradients refine shared visual representations without directly perturbing understanding-side parameters. The generation branch employs Next Embedding Prediction (NEP) paradigm, predicts continuous visual embeddings in the representation space purely autoregressively, enabling the generation objective to directly shape the representations consumed by understanding, without requiring decoders or de-tokenizers. At inference time, the generation branch is entirely discarded, yielding strictly zero additional cost.

On the data side, we systematically construct about 10M generation samples across five primary categories and 15 subtasks, covering a spectrum from pixel-level perception to high-level reasoning, and develop an automated synthesis pipeline produces high-quality training data without manual annotation. This framework allows us to conduct systematic analyses to elucidate how generation enhances complex visual understanding which leads to several intriguing findings on generation-understanding conflict, task-specific ablation and representation-level diagnostics.

We finally present **GAS** series, which validate our framework on 2B/4B-parameter and achieves consistent improvements across complex visual understanding evaluations. Our main contributions

are as follows:

- We propose a MoT-based decoupled training architecture with NEP generation supervision that effectively transfers generative learning to understanding at zero inference overhead.
- We construct diverse generation task categories with an automated data synthesis pipeline, and show that tasks with stronger latent correlations to understanding deliver the largest gains, while combining them yields further, complementary improvements on perception- and reasoning-oriented benchmarks.
- We provide systematic analyses- on parameter isolation, layer-wise supervision injection, per-task contribution, and representation-level diagnostics,offering actionable insights into why and how generation enhances understanding.

2 Methodology

Our framework is designed around a single guiding principle: generation tasks should serve as training-time auxiliary supervision that enriches the model's visual representations, without altering the inference-time architecture or cost. To realize this, we introduce three tightly integrated components: (1) Next Embedding Prediction (NEP) paradigm that formulates image generation as autoregressive embedding prediction within the same representation space used for understanding; (2) Mixture-of-Transformers (MoT) architecture that decouples generation and understanding parameters while preserving a shared visual representation pathway; and (3) systematic multi-type generation data construction strategy that maximizes the complementary benefits of diverse tasks.

2.1 Next Embedding Prediction (NEP)

Existing unified models typically adopt either discrete visual tokenization (Wang et al., 2026b; Wu et al., 2025) or diffusion-based generation (Yang et al., 2026; Xie et al., 2025). Both approaches introduce representation spaces fundamentally disjoint from the continuous embeddings consumed by the understanding branch, severely limiting how much generation training can benefit the shared visual representations (Team et al., 2026).

To bridge this gap, we propose **Next Embedding Prediction (NEP)** as a cross-modal generation paradigm. Recent literature (Xu et al., 2025a) has demonstrated that predicting continuous embeddings autoregressively serves as an exceptionally strong visual learner. Building upon this insight, our core motivation is to formulate image generation as autoregressive embedding prediction within the exact same continuous representation space used by the LLM.

This unified formulation provides a fundamental advantage: by anchoring the generative target directly in the continuous input space of the language model, it preserves a unified representation manifold for both perception and generation. Unlike discrete tokenization which maps images into a separate categorical codebook, or diffusion models that operate within specialized synthesis-oriented latent spaces, our continuous formulation avoids the need for cross-space translation. Consequently, the generative objective functions as a direct alignment mechanism. It compels the network to encode fine-grained spatial and structural visual details natively into the representation space of the LLM, ensuring that the acquired visual knowledge is readily transferable and directly exploitable for complex downstream reasoning.

Formulation and Supervision Target. Unlike single-image self-supervised pretraining (Xu et al., 2025a), NEP operates in a conditional, multi-task setting. Given a multimodal context containing an input image I and diverse linguistic instructions (e.g., for segmentation or grounding), let $\mathbf{x}_{\text{ctx}}$ denote the concatenated sequence of text tokens and source image embeddings. For a generative task with a target image I^{tgt}, the model autoregressively predicts the target embedding sequence $\hat{\mathbf{z}}^{\text{tgt}}$.

Crucially, rather than introducing an external reconstruction decoder to predict raw pixels, we extract ground-truth embeddings using the identical vision encoder followed by a visual-language projector. To ensure the target resides in the exact LLM input space, the target embeddings are defined as $\mathbf{z}^{\text{tgt}} = \text{Projector}(\text{ViT}(I^{\text{tgt}}))$. The model predicts each embedding conditioned on the context and previously generated embeddings:

$$\hat{z}_i^{\text{tgt}} = f_{\text{gen}}\big(\mathbf{x}_{\text{ctx}}, \hat{z}_{<i}^{\text{tgt}}; \Theta_{\text{gen}}\big), \tag{1}$$

where f_{gen} denotes the generative forward pass of the network parameterized by Θ_{gen}. To optimize this objective, both the predicted and target embeddings are first ℓ_2-normalized. The generation loss is then computed as the cosine distance in this

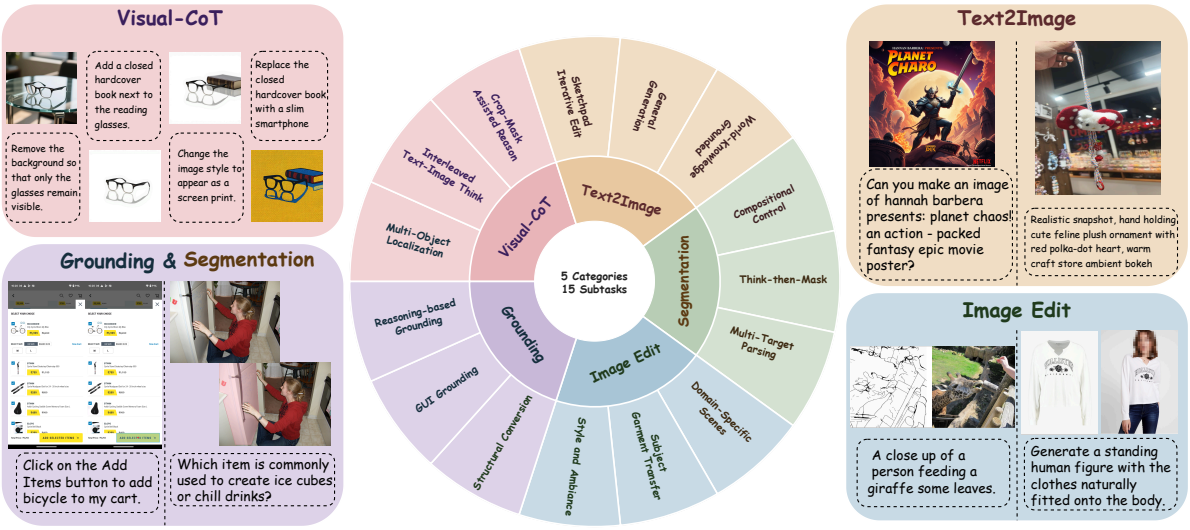


Figure 2: **Overview of generation training data.** Five primary task categories spanning 15 subtasks, jointly covering the fine-grained visual capabilities targeted by our generation-guided training.

continuous representation space:

$$\mathcal{L}_{\text{gen}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\hat{z}_i^{\text{tgt}} \cdot z_i^{\text{tgt}}}{\|\hat{z}_i^{\text{tgt}}\|_2 \|z_i^{\text{tgt}}\|_2}\right). \quad (2)$$

By optimizing directly within the understanding interface rather than pursuing photorealistic pixel synthesis, we explicitly prioritize the enrichment of deep structural representations over visual fidelity.

Target Stabilization via EMA. A critical challenge in predicting representations projected into the LLM space is that the active visual projector parameters are continuously updated during joint training (Chen et al., 2025a). Utilizing a dynamic projector to extract targets can lead to supervision drift and representation collapse. To prevent this, we maintain a *target projector* whose weights are updated via an Exponential Moving Average (EMA) of the active projector's weights. The stable target embeddings $\mathbf{z}^{\text{tgt}}$ are thus computed by passing the target image through the frozen ViT followed by this EMA target projector. Correspondingly, a dedicated **Vision Head** is applied to the output hidden states of the LLM's generation branch to predict these stable targets. As we will detail in Section 2.2, when coupled with a properly decoupled architecture, this robust NEP formulation safely injects precise spatial information into the shared representations without disrupting the LLM's text-reasoning capabilities.

2.2 MoT Decoupled Architecture

A straightforward approach to introducing generation supervision is to add the NEP loss directly onto the shared LLM backbone. However, our experiments (Section 4) and prior works reveal that this strategy degrades understanding capability (Deng et al., 2025; Li et al., 2025; Hao et al., 2025): the generative and understanding objectives impose conflicting optimization pressures, ultimately diluting the abstract semantic representations necessary for complex reasoning. To resolve this tension, we adopt the Mixture-of-Transformers (MoT) architecture (Liang et al., 2025), which decouples the upper task-specific parameters while preserving a shared lower trunk for indirect knowledge transfer.

Decoupled Architecture. To preserve visual structures that are typically abstracted in deeper layers focused on semantics, the model topology bifurcates at an intermediate layer l_{split}. The understanding branch continues to process the sequence through the remaining upper layers ($l > l_{\text{split}}$), optimized solely by the text cross-entropy loss ($\mathcal{L}_{\text{und}}$). Concurrently, we extend the backbone with a parallel generation branch comprising N_{gen} transformer layers. To guarantee a reasonable initialization and stabilize early training, the parameters of these generation layers (Θ_{gen}) are directly copied from the corresponding pre-trained layers of the understanding backbone. This branch extracts hidden states from layer l_{split} to autoregressively produce NEP predictions:

$$\begin{aligned} \mathbf{H}_{\text{gen}} &= \text{Transformer}_{\text{gen}}(\mathbf{H}^{(l_{\text{split}})}; \Theta_{\text{gen}}), \\ \hat{z}_i^{\text{tgt}} &= \text{VisionHead}(\mathbf{h}_{\text{gen},i}; \Theta_{\text{head}}), \end{aligned} \quad (3)$$

where $\mathbf{H}^{(l_{\text{split}})}$ denotes the sequence of hidden states

at layer l_{split}, and $\mathbf{h}_{\text{gen},i}$ is the i-th token representation in the generation output $\mathbf{H}_{\text{gen}}$. The generation branch is parameterized by Θ_{gen} and Θ_{head}, supervised entirely by the NEP loss ($\mathcal{L}_{\text{gen}}$).

Mechanism of Synergy. A critical question arises: how can generative training enhance understanding if their parameters diverge? The key lies in the asymmetric gradient flow across the shared first l_{split} layers. This partial decoupling induces two powerful, complementary effects:

(1) Direct Refinement of Shared Representations. During joint training, gradients from $\mathcal{L}_{\text{gen}}$ backpropagate through the generation branch into the visual projection and the shared trunk ($l \leq l_{\text{split}}$). This forces these lower layers to produce intermediate states $\mathbf{H}^{(l_{\text{split}})}$ that are enriched with precise, fine-grained visual details. Because $\mathbf{H}^{(l_{\text{split}})}$ acts as the direct input to the upper understanding branch, this visual enrichment is seamlessly inherited.

(2) Implicit Adaptation of Upper Layers. While the upper understanding layers ($l > l_{\text{split}}$) remain shielded from generative gradients, they are continuously fed the visually enriched representations from the shared trunk. Consequently, they are compelled to adapt under $\mathcal{L}_{\text{und}}$ to fully exploit these high-quality features. Empirically (Section 4), this adaptation manifests as sharpened attention heads that reliably anchor onto question-relevant visual regions, thereby enhancing evidence retrieval during multi-step reasoning.

Training and Inference. The overall objective minimizes the weighted sum $\mathcal{L} = \mathcal{L}_{\text{und}} + \lambda \cdot \mathcal{L}_{\text{gen}}$. At inference time, the entire generation branch (Θ_{gen} and Θ_{head}) is discarded. Only the understanding branch performs the forward pass, rendering the inference cost, latency, and memory footprint strictly identical to the standard baseline model.

2.3 Generation Task Construction

Prior unified multimodal works typically introduce generation data in the form of generic text-to-image synthesis or simple editing, treating generation merely as a self-contained skill. Under such configurations, understanding performance gains are often negligible or even negative. We argue that generic synthesis tasks, such as "generate an image of a cat" provide little training signal relevant to complex visual reasoning, as they merely demand a superficial alignment between basic nouns and visual concepts.

The core insight of our framework is that **generation can better enhances understanding when the generative act inherently depends on deep multimodal comprehension.** Rather than treating generation as a straightforward rendering process, we purposefully formulate it as the terminal output of a complex cognitive task. If a model must execute precise spatial localization, compositional scene analysis, or multi-step logical deduction just to figure out <u>what</u> or <u>where</u> to generate, the generation objective acts as a powerful forcing function. It compulsorily drives the network to develop the exact fine-grained perceptual and reasoning capabilities that pure text supervision fails to provide.

Guided by this principle, we select candidate generation tasks based on their <u>latent cognitive dependency</u> with the target understanding capabilities. We construct five primary categories of generation tasks including Grounding, Segmentation, Image Editing, Visual Chain-of-Thought and Text-to-Image, spanning 15 subtasks in total (illustrated in Figure 2), moving far beyond simple descriptive synthesis. Detailed definition of the tasks we construct can be found at Appendix A.1.

Crucially, our ablation studies (Section 4) confirm that these task categories provide <u>complementary supervision signals</u>. Empirically, on Counting & Spatial benchmarks the gain from combining all categories tends to exceed the best single-task gain, suggesting that diverse, comprehension-correlated generation tasks are not redundant; rather, they jointly cover different facets of fine-grained visual structure that no single task fully captures. Building on this finding, we develop a scalable data synthesis framework to produce high-quality generation data across diverse domains without manual annotation which detailed at Appenfix A.2. Combining above task selection principle with this automated construction and filtering pipeline, we ultimately scaled the total generation training corpus to approximately 10M samples spanning all five task categories.

3 Experiments

3.1 Experiments Setup

Model Configuration. We build our main models from scratch to avoid contamination from existing VLM weights. The understanding branch follows the Qwen3-VL architecture (Bai et al., 2025a) with a frozen Qwen3-VL-ViT and a Qwen3 (Yang et al., 2025) 2B/4B LLM backbone; the projector is

Model	#Params	MME test (Fu et al., 2026a)	MMMU val (Yue et al., 2024)	BLINK test (Fu et al., 2024)	RealWorldQA test	CharXiv DS (Wang et al., 2024b)	CharXiv RS (Wang et al., 2024b)	DynaMath test (Zou et al., 2025)	MathVision mini (Wang et al., 2024a)	MathVista mini (Lu et al., 2024)	LogicVista mini (Xiao et al., 2024)	VisuLogic test (Xu et al., 2025b)	CountBenchQA test (Paiss et al., 2023; Beyer et al., 2024)	CV-Bench 2D (Tong et al., 2024)	CV-Bench 3D (Tong et al., 2024)	Video-MME test (Fu et al., 2025)	MVBench test (Li et al., 2024c)
Understanding-only MLLMs																	
LLaVA-v1.5* (Liu et al., 2024)	7B	-	35.7	-	54.8	-	-	16.6	8.52	-	-	24.6	-	-	-	-	-
Qwen2.5-VL (Bai et al., 2025b)	3B	2204	49.2	48.6	64.1	-	-	-	14.8	-	-	-	-	-	-	54.1	61.3
Qwen3-VL (Bai et al., 2025a)	2B	2026	48.1	54.9	63.7	65.0	38.8	30.2	20.7	51.4	36.5	26.4	86.9	70.2	86.8	48.1	50.4
Unified Models																	
BAGEL* (Deng et al., 2025)	7B+7B	2381	55.3	-	72.8	-	-	-	-	73.2	44.3	41.7	82.5	-	-	-	-
Emu3* (Wang et al., 2026b)	8B	-	31.6	-	57.4	-	-	-	-	-	-	24.7	65.2	-	-	-	-
MetaMorph* (Tong et al., 2025)	8B	-	41.8	-	58.3	-	-	-	-	-	-	-	-	-	-	-	48.8
Janus-Pro (Chen et al., 2025c)	7B	1978*	41.0*	-	58.0*	-	-	22.2	15.8	45.3*	28.6	23.8	53.2	-	-	-	-
Show-o* (Xie et al., 2025)	1.3B	-	26.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Show-o2* (Xie et al., 2026)	1.5B	-	37.1	-	56.5	-	-	-	-	-	-	-	-	-	-	-	50.6
Cheers* (Zhang et al., 2026)	1.5B	-	36.0	-	60.9	-	-	-	-	50.5	-	-	-	-	-	-	-
Ours																	
GAS	2B	2003	46.6	48.0	64.8	64.4	34.3	47.9	20.4	56.4	32.7	28.6	90.1	73.2	73.8	46.2	47.8
-baseline		1986	46.9	46.7	64.1	64.0	34.6	46.2	19.1	54.4	36.0	26.8	87.7	69.9	73.2	45.2	49.0
GAS	4B	2146	57.0	56.9	70.3	78.4	40.5	55.2	27.0	68.1	37.1	23.6	90.8	75.4	83.6	54.4	52.4
-baseline		2092	54.9	57.8	70.2	77.9	40.3	57.6	25.3	65.6	37.1	26.1	90.1	75.1	82.1	55.0	54.2

Table 1: **Overall results** for a range of understanding-only MLLMs, Unified Models, baselines and our GAS family. The result of the best-performing model is in **bold**, and the second best is underlined. * denotes metrics taken directly from public reports rather than reproduced by us.

randomly initialized. The MoT generation branch is initialized from the corresponding backbone layers and branches from an intermediate split layer $l_s \approx L/2$.

Training Details. Training follows the two-stage procedure introduced in Section 2.2: Stage 1 aligns the generation branch while freezing the understanding side, and Stage 2 jointly optimizes understanding and NEP generation losses with the ViT kept frozen. The full optimization setup, hyper-parameters, EMA target update, and compute details are provided in Appendix C.

3.2 Evaluation Setup

We evaluate GAS on four capability groups: general perception, visual reasoning, counting & spatial understanding, and video understanding . This grouping mirrors the core capabilities targeted by our generation tasks and is used consistently throughout the result tables. Detailed benchmark descriptions and grouping rationale are deferred to Appendix C.

3.3 Main Results

Table 1 reports main results of GAS against standard MLLMs and unified models across all benchmarks. All improvements are achieved with zero additional inference cost since the generation branch is discarded after training.

At the same 2B scale, GAS is highly competitive with baseline and other open-source models while showing markedly stronger reasoning and counting/spatial behavior: DynaMath increases from 46.2 to **47.9** (+1.5pp), MathVista from 54.4 to **56.4** (+2.0pp), CountBenchQA from 87.7 to **90.1**, CV-Bench-2D from 69.9 to **73.2**, and VisuLogic from 26.8 to **28.6**, all directly aligned with the spatial perception and visual reasoning capabilities targeted by our grounding, segmentation, and Visual-CoT generation tasks. Scaling to 4B, GAS establishes the strongest overall results in the unified family, leading on most of the 16 benchmarks we report (e.g., MMMU 57.0, CharXiv-DS 78.4 / RS 40.5, CountBenchQA 90.8) and surpassing significantly larger unified models such as BAGEL (7B+7B) and Emu3 (8B). The lone exception is VisuLogic at 4B (23.6), where the larger backbone shows a benchmark-specific regression that does not transfer to other reasoning suites, suggesting an idiosyncratic sensitivity of this puzzle-style benchmark to backbone scale. Together, these results show that our generation-guided framework consistently transfers fine-grained perception and reasoning gains to the understanding branch across both scales.

Additional qualitative cases are deferred to Appendix B. GAS consistently exhibits better spatial precision, finer visual discrimination, and more grounded reasoning chains, qualitatively matching the benchmark gains reported above. In the next section, we turn to controlled analyses that explain where these gains come from.

Stage	Configuration	Overall	Perception	Reasoning	Count&Spatial	Video	MME	MMMU	MathVista	MVBench
From Scratch	Baseline	47.25	56.55	31.38	76.92	46.60	1848	43.5	42.7	46.9
	+ MoT	**48.25**	**58.52**	**32.63**	75.15	**47.65**	**1895**	**45.8**	**47.1**	**49.5**
After Pretrain	Baseline	56.23	63.80	43.22	82.40	53.95	2169	53.2	72.3	55.1
	+ MoT	**56.71**	**63.88**	**44.12**	**82.95**	53.95	2170	51.1	71.2	54.6
After SFT	Baseline	52.65	61.32	38.23	79.82	51.40	2060	46.3	52.9	53.0
	+ MoT	**53.49**	**61.62**	**39.57**	**80.80**	**51.65**	**2096**	45.8	**53.6**	52.9

Table 2: **Generalization across training stages.** MoT generation-guided training yields improvements when applied to understanding backbones at different capability levels.

4 Analysis

We conduct a series of controlled experiments to dissect the mechanisms underlying our generation-guided training framework. All analyses are performed on the 2B-parameter model with 2.5B understanding combine with 2.5B generation data tokens, using the same evaluation suite described in Section 3, unless otherwise noted. To keep the main paper centered on the primary claims, we emphasize four questions here: (1) Does our generation-guided training providing consistent benefits across understanding backbones regardless of the pre-existing strength? (2) whether decoupling is necessary to avoid generation-understanding conflict, (3) whether transfer depends on the latent correlation of the generation tasks, and (4) what representational changes emerge after training. Additional studies on layer-wise injection, attention visualization, and text-only reasoning are deferred to the appendix.

4.1 Generalization Across Training Stages

Firstly, we verify that our generation-guided framework is not limited to a specific training stage or backbone quality by applying MoT training atop understanding backbones of varying capability levels. We evaluate three settings: (i) training from scratch (the raw LLM backbone), (ii) after pre-training with large-scale multimodal data, and (iii) after directly supervised fine-tuning (SFT) based on Qwen3-VL-2B (Bai et al., 2025a). Results are presented in Table 2.

MoT improves performance across all three stages, with the largest gain from scratch (+1.00pp overall), but still consistent gains after pretraining (+0.48pp) and during SFT (+0.84pp). This shows that generation-guided training acts as a stage-agnostic augmentation rather than a technique tied to a specific training regime.

Config.	Ovl.	Perc.	Reas.	C&S	Vid.
Und. only	47.63	57.22	**34.20**	70.78	45.60
Shared	46.00	54.95	32.25	70.20	45.15
MoT (Ours)	**48.25**	**58.52**	32.63	**75.15**	**47.65**

Table 3: **Generation-understanding conflict analysis.** Shared training causes clear degradation, while MoT decoupling turns generation supervision into consistent understanding gains.

4.2 Generation-Understanding Conflict and MoT Decoupling

A key question is whether generation and understanding objectives inherently conflict under joint training. We compare three settings: **Shared Backbone**, which applies NEP directly to the same LLM parameters used for understanding; **MoT Decoupled**, our design with an independent generation branch; and baseline Qwen3-VL architecture training from scratch.

Table 3 shows that shared-backbone setting degrades clearly relative to the baseline, confirming direct optimization conflict. By contrast, MoT not only avoids this drop but improves every aggregate metric, with the largest gains on Count&Spatial (+4.37pp) and Video (+2.05pp). This makes decoupling a necessary condition for converting generation supervision into useful transfer.

4.3 Per-Task Generation Contribution

We next investigate which generation tasks contribute most effectively to understanding enhancement by individually adding 200k samples of each five task categories to understanding-only baseline.

All five task categories improve at least one aggregate dimension, but with clearly different profiles, supporting our latent task correlation hypothesis. While segmentation and grounding are most effective for Count&Spatial (+2.00pp and +2.37pp), consistent with their direct supervision on fine-

Config.	Ovl.	Perc.	Reas.	C&S	Vid.
Und. only	47.63	57.22	34.20	70.78	45.60
+ T2I	48.12	57.70	34.66	71.65	45.85
+ Edit	47.77	56.70	34.12	72.95	45.70
+ Seg	47.43	55.32	34.35	72.78	45.55
+ Grd	47.92	57.28	33.92	73.15	45.95
+ V-CoT	47.86	57.80	34.26	71.45	45.20
+ All	**48.30**	57.08	34.28	**75.72**	45.40

Table 4: **Per-task ablation.** Each row adds a single generation task category to the understanding-only baseline; "All" combines all five tasks totally 1000k samples.

Config.	Ovl.	Perc.	Reas.	C&S	Vid.
L8	47.06	56.80	31.52	**75.88**	45.40
L14	47.85	**58.58**	31.64	75.28	47.60
L14 + warmup	**48.25**	58.52	**32.63**	75.15	**47.65**
L14 + unf. ViT	46.20	55.80	31.31	72.60	45.25
L20	47.25	58.20	31.43	74.60	45.45

Table 5: **Layer-wise injection ablation.** A mid-layer split ($L/2$) with warmup gives the best overall trade-off.

grained spatial structure. T2I provides moderate gains across dimensions, and Visual-CoT is most helpful for reasoning-oriented transfer.

Most notably, the combined setting gives the best Count&Spatial result (75.72), suggesting that these tasks are complementary rather than redundant. The practical takeaway is that composing generation tasks with different latent correlations to understanding is a productive design lever, especially for perception- and counting-oriented capabilities.

Effect of Task Correlation on Understanding Gains. We further verify this point with a controlled T2I study in Appendix C.2, where correlation-aware prompt rewriting substantially outperforms generic short-caption T2I under the same data volume.

4.4 Layer-wise Supervision Injection

Our MoT architecture extracts features from an intermediate layer l_s of the understanding backbone to feed into the generation branch. We therefore compare split layers 8, 14 (approximately $L/2$ for our 2B, 28-layer model), and 20, together with progressive warmup and ViT unfreezing.

The results support two conclusions. First, supervision injected near the middle of the backbone provides the best overall trade-off, suggesting that mid-level representations best balance residual spatial structure with the semantic abstraction needed for broad transfer, while the slightly stronger C&S result at layer 8 indicates that earlier layers remain more specialized for low-level spatial cues. Second, progressive warmup further improves the layer-14 setting (+0.40pp overall), whereas unfreezing the ViT causes a clear drop (-1.65pp), reinforcing our design choice to keep the visual encoder fixed and route generation benefits through the shared projector and lower LLM trunk.

4.5 Representation-Level Diagnostics

To understand why generation guidance improves understanding, we briefly examine the learned representations. As detailed in Appendix Figure 4, the MoT-trained model preserves substantially more visual information in middle and deep layers than the baseline, and its hidden states yield stronger region-level linear probes while remaining comparable on global semantic probing. Appendix Figure 5 further shows that this benefit is reflected at inference time as sharper attention over question-relevant visual regions, rather than the diffuse visual focus observed in the baseline. Appendix Table 7 finally verifies that these representational gains do not come at the expense of pure text reasoning.

5 Conclusion and Future Work

We present a generation-guided training framework that improves multimodal understanding by using visual generation purely as training-time auxiliary supervision. Concretely, NEP provides a generation target aligned with the LLM input space, while the MoT decoupled design avoids direct optimization conflict and preserves zero additional inference cost. Across diverse understanding benchmarks, the framework delivers consistent gains, and our analyses show that the improvement arises from better visual retention, more focused visual attention, and task-dependent transfer shaped by generation-task correlation.

These results suggest that visual generation is most effective not as an inference-time capability to preserve, but as a carefully structured supervisory signal for building stronger understanding models. Future work may extend this paradigm to longer-context multi-image / hour-long video settings.

6 Limitations

This work focuses on generation-guided training for visual understanding in image- and video-centric multimodal settings. While the proposed framework is shown to be effective across diverse perception, reasoning, counting, spatial, and temporal benchmarks, we do not exhaustively study other modalities such as audio, 3D, or embodied sensory streams. In addition, our analysis is centered on the current GAS design and training scale; further investigation of longer-context settings, and broader task mixtures may provide a more complete picture of how generation-guided supervision behaves in other regimes. These are natural directions for extending the framework rather than constraints on its practical use in the settings studied here.

7 Ethics Statement

Our method uses automatically constructed generation data and large-scale web-derived sources, which may inherit biases, cultural imbalances, or incomplete coverage present in upstream datasets and pretrained components. Although the generation branch is used only during training and discarded at inference time, models trained with our framework may still reflect such biases in downstream predictions. In addition, stronger multimodal understanding systems can be deployed in both beneficial and potentially sensitive applications, so practitioners should evaluate them carefully before use in high-stakes settings. We therefore recommend standard safeguards, including dataset filtering, bias-aware evaluation, and human oversight when deploying models trained with generation-guided supervision.

References

2024. Introducing Gemini 2.0: our new AI model for the agentic era.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL Technical Report. arXiv preprint. ArXiv:2511.21631 [cs].

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, and 8 others. 2025b. Qwen2.5-VL Technical Report. Preprint, arXiv:2502.13923.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. PaliGemma: A versatile 3B VLM for transfer. arXiv preprint arXiv:2407.07726.

Fengjiao Chen, Minhao Jing, Weitao Lu, Yan Feng, Xiaoyu Li, and Xuezhi Cao. 2025a. Unihetero: Could generation enhance understanding for vision-language-model at large data scale? arXiv preprint arXiv:2512.23512.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, and 1 others. 2025b. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36:49250–49267.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2026a. Mme: A comprehensive evaluation benchmark for multimodal large language models. Advances in Neural Information Processing Systems, 38.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 24108–24118.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. 2026b. Lance: Unified multimodal modeling by multi-task synergy. arXiv preprint arXiv:2605.18678.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision, pages 148–166. Springer.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 13861–13871.

Google. 2025. Gemini 3 flash: frontier intelligence built for speed.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. 2025. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. arXiv preprint arXiv:2509.24365.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 787–798.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, and 1 others. 2025. Explain before you answer: A survey on compositional visual reasoning. arXiv preprint arXiv:2508.17298.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22195–22206.

Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. 2025. Unifork: Exploring modality alignment for unified multimodal understanding and generation. arXiv preprint arXiv:2506.17202.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2026. Videochat-flash: Hierarchical compression for long-context video modeling. In The Fourteenth International Conference on Learning Representations.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. 2025. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. Transactions on Machine Learning Research.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv preprint. ArXiv:2304.08485 [cs].

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. 2025. Video-xl-pro: Reconstructive token compression for extremely long video understanding. Preprint, arXiv:2503.18478.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In International Conference on Learning Representations, volume 2024, pages 23439–23554.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11–20.

OpenAI. 2025. gpt-5-system-card.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching CLIP to Count to Ten. arXiv preprint arXiv:2302.12066.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In SC20: international conference for high performance computing, networking, storage and analysis, pages 1–16. IEEE.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37:8612–8642.

Tao Shen, Xin Wan, Taicai Chen, Rui Zhang, Junwen Pan, Dawei Lu, Fanding Lei, Zhilin Lu, Yunfei Yang, Chen Cheng, and 1 others. 2025. Mammothmoda2: A unified ar-diffusion framework for multimodal understanding and generation. arXiv preprint arXiv:2511.18262.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and LILI YU. 2026. LMFusion: Adapting pretrained language models for multimodal generation. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21440–21455.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 26160–26169.

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

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, and 1 others. 2026. Longcat-next: Lexicalizing modalities as discrete tokens. arXiv preprint arXiv:2603.27538.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems, 37:87310–87356.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2025. Metamorph: Multimodal understanding and generation via instruction tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17001–17012.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786.

Jiacong Wang, Zijian Kang, Haochen Wang, LiangXiao, Ya Wang, Jiawen Li, Bohong Wu, Ran Jiao, Haiyong Jiang, ChaoFeng, and Jun Xiao. 2026a. VGR: Visual grounded reasoning. In The Fourteenth International Conference on Learning Representations.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems, 37:95095–95169.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, and 7 others. 2026b. Multimodal learning with next-token prediction for large multimodal models. Nature, 650(8101):327–333.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems, 37:113569–113697.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2025. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 12966–12977.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. arXiv preprint arXiv:2407.04973.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o: One single transformer to unify multimodal understanding and generation. In The Thirteenth International Conference on Learning Representations.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 2026. Show-o2: Improved native unified multimodal models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X Yu. 2025a. Next-embedding prediction makes strong vision learners. arXiv preprint arXiv:2512.16922.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. 2026. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In The Fourteenth International Conference on Learning Representations.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025b. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. arXiv preprint arXiv:2504.15279.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2026. Mmada: Multimodal large diffusion language models. Advances in Neural Information Processing Systems, 38:138867–138907.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9556–9567.

Yichen Zhang, Da Peng, Zonghao Guo, Zijian Zhang, Xuesong Yang, Tong Sun, Shichu Sun, Yidan Zhang, Yanghao Li, Haiyan Zhao, and 1 others. 2026. Cheers: Decoupling patch details from semantic representations enables unified multimodal comprehension and generation. arXiv preprint arXiv:2603.12793.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are visually-grounded language models bad at image classification? Advances in Neural Information Processing Systems, 37:51727–51753.

Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2025. Transfusion: Predict the next token and diffuse images with one multi-modal model. In The Thirteenth International Conference on Learning Representations.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In International Conference on Learning Representations, volume 2025, pages 48337–48383.

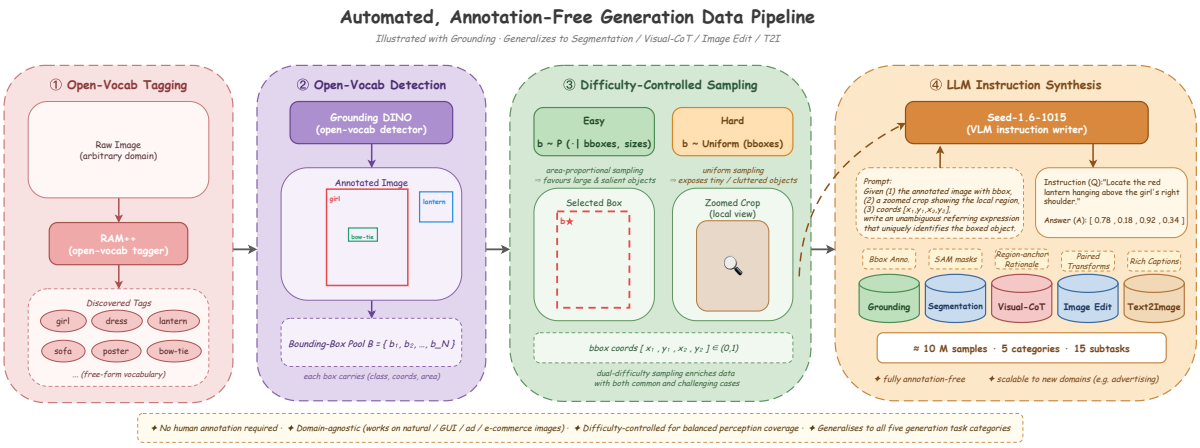


Figure 3: **Generation task construction pipeline.** Illustration of our automated, annotation-free data construction pipeline, taking grounding as a representative example.

A Concrete Description of Generation Tasks and Construction Pipeline

A.1 Detailed Definition of Tasks

- **Grounding**: Region localization tasks that compel the model to perfectly resolve complex linguistic referring expressions before predicting absolute spatial coordinates, directly cultivating fine-grained spatial reasoning.

- **Segmentation**: Pixel-level mask prediction tasks that force the model to comprehend semantic boundaries and physical extents based on intricate textual instructions, reinforcing instance-level discrimination.

- **Image Editing**: Conditional image transformation tasks (e.g., style transfer, object removal) that require a deep, compositional understanding of scene layouts and object interdependencies before executing the edit.

- **Visual Chain-of-Thought**: Multi-step reasoning tasks that demand the model to actively generate intermediate visual evidence (e.g., highlighting key regions) to anchor multi-step logical deduction, simulating a "thinking with images" process.

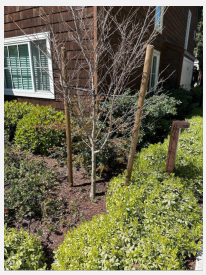
- **Text-to-Image (T2I)**: Carefully curated image synthesis tasks grounded in rich world knowledge and dense structural prompts, establishing foundational cross-modal alignment beyond naive concept mapping.

A.2 Detailed Description of Automated Construction and Filtering Pipeline

Taking grounding as a representative example, our automated pipeline operates in four stages: (1) open-vocabulary tagging via RAM++ to discover all salient objects; (2) open-vocabulary detection via Grounding DINO to localize each tag; (3) difficulty-controlled sampling (area-proportional for easy, uniform for hard) to select target boxes; and (4) LLM-based instruction synthesis using the annotated image and a zoomed crop to generate unambiguous referring expressions. A visual example of this construction pipeline is provided in Figure 3. This pipeline generalizes to broad image domains and enables scaling of generation data to new scenarios (e.g., advertising) with minimal engineering effort.

B More Qualitative Cases of GAS

Qualitative Case 1: RealWorldQA



Question: How many wooden poles are in this photo?

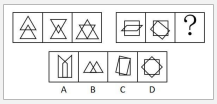
Ours 4B:

```
<think>
To determine the number of wooden poles in the photo, let's
analyze the image step by step:
1. **Foreground Analysis**: The image shows a small tree
with a thin trunk. There are two visible wooden poles
supporting the tree. These poles are positioned on either
```

side of the tree, providing stability.
2. **Background and Surroundings**: The background includes a building with windows and some greenery. There are no additional wooden poles visible in the background or elsewhere in the image.
3. **Counting the Poles**: Based on the visible elements, there are exactly two wooden poles supporting the tree.
Given this analysis, the correct answer is:
C. There are 2 wooden poles.
</think>

The final answer is $\boxed{C}$.

Qualitative Case 2: VisuLogic (Xu et al., 2025b)



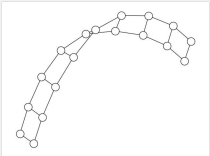
Question: Choose the most appropriate one from the four given options to fill in the question mark so that a certain regularity is presented: A, B, C, D.

Ours 4B:

<think>
To solve this problem, we first analyze the shapes and number of lines in each figure, then identify the underlying pattern.
1. **Analyze shape and line count:**
Observing the figures in the first and second rows, each figure is composed of **triangles** or **quadrilaterals** (rhombus-like shapes), with a regular line count.
- Row 1, Fig. 1: two triangles, lines $= 2 \times 3 = 6$;
- Row 1, Fig. 2: two quadrilaterals, lines $= 2 \times 4 = 8$;
- Row 1, Fig. 3: two triangles, lines $= 2 \times 3 = 6$;
- Row 2, Fig. 1: two quadrilaterals, lines $= 2 \times 4 = 8$;
- Row 2, Fig. 2: two triangles, lines $= 2 \times 3 = 6$;
- Row 2, Fig. 3 (the question mark): should follow the same shape/line-count pattern and stacking style.
2. **Identify the pattern:**
The figures in both rows alternate between **triangle pairs** and **quadrilateral pairs**, with line counts of 6 and 8 respectively. All figures share the same two-shape overlapping composition with consistent contour and inner-line connectivity.
3. **Match the options:**
Option D consists of two overlapping quadrilaterals (rhombus-like), whose line count and composition match the pattern (Row 2 starts with a quadrilateral pair, so the question mark should also be a quadrilateral pair). Option A is a triangle pair with a mismatched shape; Option B is a triangle pair with inconsistent line count; Option C is a quadrilateral but with a different stacking style.
Therefore, the correct answer is $\boxed{D}$.
</think>

The correct answer is $\boxed{D}$.

Qualitative Case 3: DynaMath (Zou et al., 2025)



Question: Find the chromatic number of the following graph.

Ours 4B:

<think> To determine the chromatic number of the given graph, we need to find the minimum number of colors required to color the vertices such that no two adjacent vertices share the same color. Let's analyze the graph step by step.
1. **Identify the Structure of the Graph:** - The graph appears to be a bipartite graph. A bipartite graph is a graph whose vertices can be divided into two disjoint sets such that every edge connects a vertex in one set to a vertex in the other set. - In this graph, we can observe that the vertices can be divided into two sets: one set containing the vertices on the left side of the graph and the other set containing the vertices on the right side of the graph.
2. **Verify Bipartiteness:** - Check if there are any odd-length cycles in the graph. If there are no odd-length cycles, the graph is bipartite. - Upon inspection, the graph does not contain any odd-length cycles. All cycles in the graph are even-length cycles.
3. **Determine the Chromatic Number:** - For a bipartite graph, the chromatic number is always 2. This is because we can color one set of vertices with one color and the other set with a different color, ensuring that no two adjacent vertices share the same color.
4. **Conclusion:** - Since the graph is bipartite, the chromatic number is 2.
</think>

The final answer is $\boxed{2}$.

Qualitative Case 4: CV-Bench (Tong et al., 2024)

Question: Estimate the real-world distances between objects in this image. Which object is closer to the chair (highlighted by a red box), the television (highlighted by a blue box) or the lamp (highlighted by a green box)? A. television, B. lamp

Ours 4B:

<think>
To determine which object is closer to the chair (highlighted by the red box), we need to analyze the spatial relationships in the image:
1. **Chair (red box)**: This is located in the foreground, near the dining table.
2. **Television (blue box)**: This is positioned on a stand against the wall, further back in the room.
3. **Lamp (green box)**: This is placed on the floor near the sofa, closer to the center of the room but still farther back than the chair.
Analysis: - The **chair** is in the foreground, very close to the viewer.
- The **television** is on a stand against the wall, which is significantly farther away from the chair.
- The **lamp** is on the floor near the sofa, which is also farther away from the chair compared to the chair's position.
Conclusion: - The **chair** is the closest object to itself, and the **lamp** is closer to the chair than the **television**.
Thus, the correct answer is:
B. lamp
</think>

The final answer is $\boxed{B}$.

C Detailed Experimental Setup

Model Configuration. To ensure a controlled environment devoid of data contamination from existing VLM weights, we build our main models from scratch. Adopting the Qwen3-VL architecture (Bai et al., 2025a) for the understanding branch, we utilize Qwen3-VL-ViT as a visual encoder initialized with pretrained weights, which remains frozen throughout all training stages. For the LLM backbone, we employ the 2B/4B variants of Qwen3 (Yang et al., 2025). We use the raw LLM weights without prior multimodal alignment, and the ViT-LLM projector is randomly initialized, ensuring that all visual-language capability is acquired solely through our training pipeline. The MoT generation branch consists of L_{mot} transformer layers initialized from the corresponding pre-trained layers of the understanding backbone, with hidden states extracted from the intermediate layer $l_s \approx L/2$ of the understanding backbone.

Implementation Details. Training proceeds in two stages. In **Stage 1 (MoT Align)**, we freeze all understanding-side parameters and train only the MoT generation branch using T2I generation data so as to activate the generation pathway and align it with the existing visual representation space without perturbing the pretrained understanding capability. In **Stage 2 (Joint Training)**, understanding and generation data are used simultaneously. The understanding branch is optimized with text cross-entropy loss, while the generation branch is optimized with NEP loss under a progressive weighting schedule that linearly ramps λ from 0.015 to 1.0 over 4k steps. This setting allows generation gradients to refine the shared trunk while leaving the ViT frozen, thereby enabling the indirect transfer mechanism described in Section 2.2. Our GAS training paradigm incurs an approximately 11% overhead in GPU hours.

As for hyper-parameters, we use the AdamW (Kingma and Ba, 2014) optimizer with a cosine learning rate schedule, a peak learning rate of 5×10^{-5} for both the understanding branch and the MoT generation branch. We maintain a target projector EMA with decay rate 0.999 for training stability, following $\theta_{\text{EMA}}^{(t)} = 0.999 \cdot \theta_{\text{EMA}}^{(t-1)} + 0.001 \cdot \theta^{(t)}$. Training is conducted on 32 GPUs with DeepSpeed ZeRO-2 (Rajbhandari et al., 2020), using a global batch size of 128 and a maximum sequence length of 8192 tokens.

Evaluation Setup. To comprehensively assess the impact of generation-guided training on complex visual understanding, we organize evaluation into four capability dimensions:

- **General Perception (Multi-Image):** MME (Fu et al., 2026a), MMMU (Yue et al., 2024), BLINK (Fu et al., 2024), and RealWorldQA, covering broad visual perception, world knowledge integration, and multi-image comprehension.
- **Visual Reasoning:** CharXiv (Wang et al., 2024b), DynaMath (Zou et al., 2025), MathVision (Wang et al., 2024a), MathVista (Lu et al., 2024), LogicVista (Xiao et al., 2024), and VisuLogic (Xu et al., 2025b), targeting chart, diagram, and geometric reasoning with multi-step inference.
- **Counting & Spatial Understanding:** CountBench (Paiss et al., 2023; Beyer et al., 2024) and CV-Bench (Tong et al., 2024), which probe object counting accuracy and spatial relationship understanding.
- **Video Understanding:** Video-MME (Fu et al., 2025) and MVBench (Li et al., 2024c), which test temporal reasoning and sustained visual attention across frames.

C.1 Representation-Level Diagnostics

To understand how generation training improves understanding at the representation level, we conduct three complementary analyses: (i) visual information retention across layers, (ii) attention pattern visualization, and (iii) linear probing diagnostics.

Visual Information Retention. We measure the cosine similarity between the visual token representations at each LLM layer and the original visual features at the LLM input, tracking how much visual information is preserved as processing deepens. Figure 4 (left) shows that in the baseline model, visual information progressively attenuates beyond layer $L/3$, with cosine similarity dropping sharply in the latter half of the network. The MoT-trained model, in contrast, maintains significantly higher visual similarity throughout the middle and deep layers (beyond layer 9), demonstrating that generation supervision encourages the network to retain visual features rather than allowing them to be "squeezed out" by language-dominant processing. The generation branch, by requiring accurate

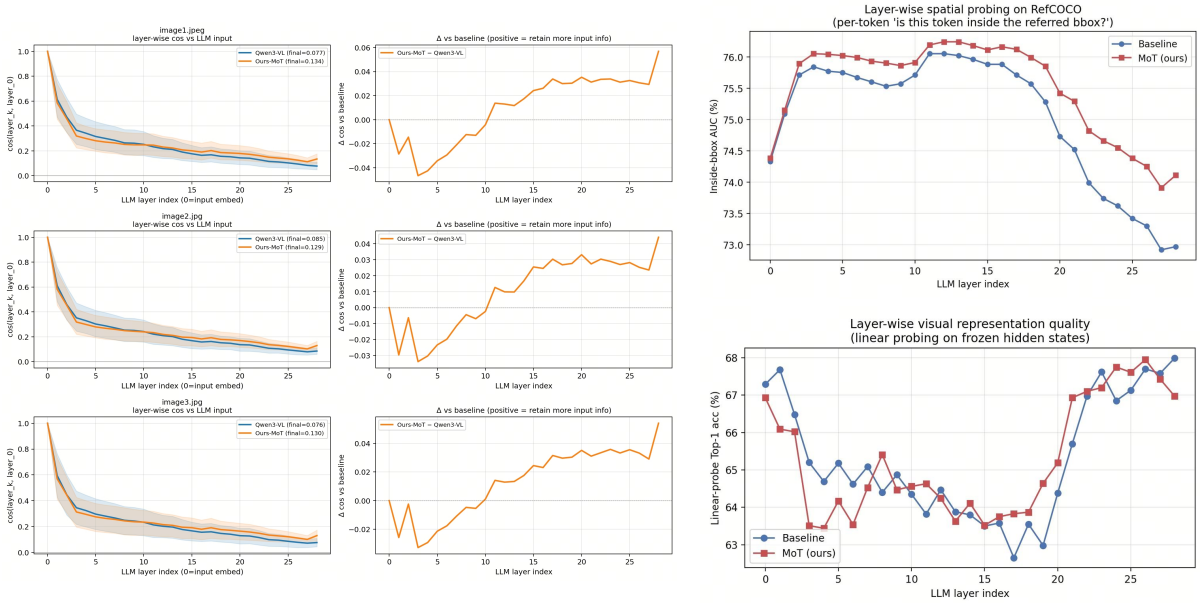


Figure 4: **Visual information retention and layer-wise linear probing.** Left: cosine similarity between per-layer visual tokens and the input visual features, and suggests that generation training preserves visual information in deep layers where the baseline rapidly decays. Right top / bottom: linear probing accuracy on RefCOCO (region-level) and ImageNet (global). The two models perform comparably on ImageNet, but the MoT-trained model gains a widening advantage on RefCOCO at deeper layers, indicating that generation supervision specifically enriches fine-grained, region-level representations.

embedding prediction from intermediate representations, creates an implicit incentive for the backbone to preserve visual information at those layers, which then propagates benefit to the understanding branch operating on the same representations.

Layer-wise Linear Probing. To quantify the discriminative quality of visual representations at each layer, we extract hidden states corresponding to visual tokens and train linear classifiers on ImageNet (Deng et al., 2009) (global classification) and RefCOCO (Mao et al., 2016; Kazemzadeh et al., 2014) (region-level referring expression comprehension). As shown in Figure 4 (right), the two models perform comparably on ImageNet across all layers. However, on RefCOCO, the MoT-trained model exhibits a growing advantage that widens consistently from layer 8 onward, ultimately achieving substantially higher probing accuracy at the final layer. This pattern reveals that generation training specifically enriches fine-grained, region-level visual representations rather than generic semantic features, precisely aligning with the observation that our framework's strongest gains appear on spatial and grounding benchmarks. The widening gap at deeper layers further corroborates the retention analysis: because the MoT model preserves more spatial visual information in deep layers, its representations remain useful for fine-grained tasks even after extensive language-driven processing.

C.2 Effect of Task Correlation on Understanding Gains

We hypothesize that the conflict issue is not the modality itself, but the weak correlation between generic short-caption T2I and the visual capabilities probed by complex understanding benchmarks. To verify this, we collect 200k additional T2I samples from COYO-700M and compare two variants: (i) **T2I (easy)**, standard short-caption synthesis, and (ii) **T2I (rewrite)**, where prompts are enriched with world knowledge, multi-object compositions, and explicit attribute/relation cues to better match the scene structures required by our reasoning and perception benchmarks.

The two variants share the same task category and data volume, yet diverge sharply: generic T2I yields only marginal overall gains (+0.24pp) and even hurts Perception (-0.75pp), whereas the rewrite variant improves every dimension and lifts Overall by +0.71pp. This result closes the loop with our latent task correlation principle: transfer scales with task-level correlation, not simply with

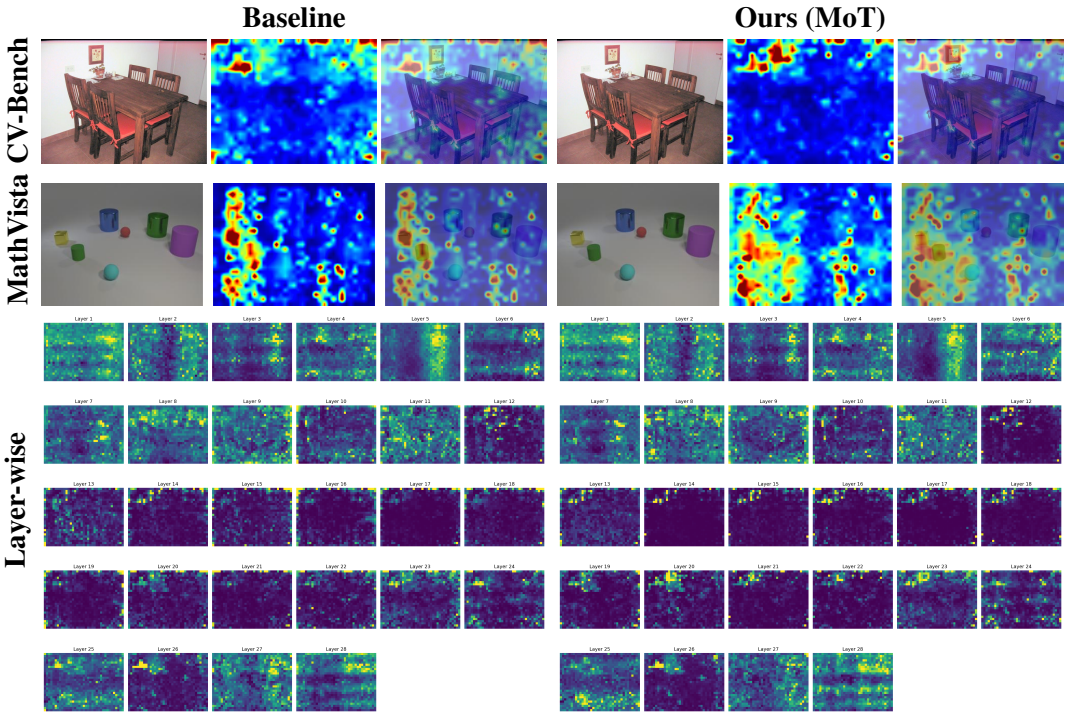


Figure 5: **Layer-wise attention visualization on representative understanding samples.** The first two rows compare attention over visual tokens between the baseline and our MoT-trained model on CV-Bench-2D and MathVista. The bottom row shows the full layer-by-layer attention evolution (layers 1–28): the baseline's attention rapidly diffuses and collapses across mid-to-deep layers, whereas the MoT-trained model preserves sharp, task-relevant focus throughout the network depth.

Configuration	Overall	Perception	Reasoning	Count&Spatial	Video
Und. only baseline	47.63	57.22	34.20	70.78	45.60
+ T2I (easy)	47.87	56.47	34.37	72.47	46.55
+ T2I (rewrite)	**48.34**	**57.53**	**34.58**	**73.62**	45.95

Table 6: **Effect of T2I prompt correlation with understanding tasks.** Aligning T2I prompts with the compositional, knowledge-grounded scene structures required by understanding benchmarks substantially amplifies the enhancement effect.

Model	ZebraLogic			MMLU-Redux
	Puzzle Acc	Cell Acc	Small Acc	Acc
Qwen3-VL baseline	18.0	12.74	52.19	52.66
GAS	**20.8**	**14.79**	**56.56**	**53.89**

Table 7: **Text-only reasoning evaluation on ZebraLogic and MMLU-Redux.**

sample count.

C.3 Attention Visualization and Text-only Reasoning

Qualitative Case 5: CV-Bench-2D (Attention Analysis)

Question: How many pictures are in the image? A. 1 B. 3 C. 2 D. 0
Ground Truth: A
Baseline: <think> The image shows a room with a wooden table and chairs; no separate pictures appear, so the answer is D (0). </think> Answer: D ✗
Ours (MoT): <think> The image is itself a single coherent scene (a dining setup); among the choices, A (1) corresponds to the single picture present. </think> Answer: A ✓

Qualitative Case 6: MathVista-MINI (Attention Analysis)

Question: Subtract all red things. Subtract all tiny matte balls. How many objects are left?
Ground Truth: 5
Baseline: <think> Identifies the red matte ball as both "red" and "tiny matte ball", but then double-counts the remaining items, mistakenly listing 6 leftover objects. </think> Answer: 6 ✗
Ours (MoT): <think> Total 7 objects; the red sphere is the only object that is both red and a tiny matte ball, so it is removed once. Remaining objects: yellow cube, green small cylinder, cyan sphere, blue / green / purple large cylinders, giving $7 - 1 - 1 = 5$. </think> Answer: 5 ✓

D Related Works Discussion

Multimodal large language models (MLLMs) have emerged as the prevailing paradigm for image and video understanding, typically by aligning pretrained visual encoders with powerful language

backbones. Representative early systems include Flamingo (Alayrac et al., 2022) and InstructBLIP (Dai et al., 2023), whereas subsequent open-source families such as the LLaVA series (Li et al., 2024a; Liu et al., 2023; Li et al., 2024b) and Qwen-VL series (Bai et al., 2025a,b) have further advanced image perception and understanding. Most studies in this line largely follow the LLaVA paradigm (Liu et al., 2023), in which visual inputs are first encoded by a vision encoder (Radford et al., 2021; Tschannen et al., 2025) and subsequently concatenated with text tokens for joint modeling by a language model decoder. Proprietary systems such as GPT (OpenAI, 2025) and Gemini (Int, 2024; Google, 2025) likewise exhibit strong multimodal understanding capabilities. Recent advances have further extended these models to multimodal visual reasoning (Xu et al., 2026; Wang et al., 2026a; Ke et al., 2025) and video understanding (Liu et al., 2025; Shu et al., 2025; Li et al., 2026). Despite their notable strengths in semantic abstraction and cross-modal alignment, these models are primarily optimized for text generation rather than learning from native visual information. The training signal flows exclusively from vision to language, leaving fine-grained visual evidence under-supervised.

Recent unified multimodal models (UMMs) attempt to close this gap by jointly training understanding and generation within a single framework, thereby introducing visual synthesis objectives that force the model to engage more deeply with image structure. One line follows a fully autoregressive formulation: Chameleon (Team, 2024) and Emu3 (Wang et al., 2026b) cast both understanding and generation into next-token prediction over a shared discrete visual token space. Another line adopts AR-diffusion hybrid formulations, combining language modeling for text with diffusion or flow-based objectives for visual generation, as in Transfusion (Zhou et al., 2025), Show-o (Xie et al., 2025, 2026), and BLIP3-o (Chen et al., 2025b). Several works further explore architectural decoupling to mitigate the conflict between generation and understanding objectives: Janus (Chen et al., 2025c) decouples visual encoding pathways while sharing a single LLM backbone; MetaMorph (Tong et al., 2025) introduces Visual-Predictive Instruction Tuning and observes that improving visual understanding implicitly benefits generation; LMFusion (Shi et al., 2026) adds modality-specific transformer layers to preserve pretrained language capabilities during generation training; and BAGEL (Deng et al., 2025) scales on trillions of interleaved tokens, demonstrating emergent compositional reasoning abilities in generation tasks.

However, despite these architectural innovations, existing UMMs predominantly pursue generation quality or emergent capabilities from scale, while treating the understanding branch as a secondary beneficiary rather than the primary optimization target. Moreover, the introduction of generation objectives frequently degrades understanding performance due to conflicting gradient signals, a tension that current methods address by either accepting the tradeoff or increasing model capacity (Team et al., 2026). We position GAS as a systematic attempt to use generation supervision explicitly for understanding improvement, with architectural decoupling and task selection both designed around that goal.