Semantic Generative Tuning for Unified Multimodal Models

Songsong Yu[1,2], Yuxin Chen[2], Ying Shan[2], and Yanwei Li[1⋆]

[1] School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China
[2] Tencent ARC Lab, Shenzhen, China

Abstract. Unified multimodal models (UMMs) strive to consolidate visual understanding and visual generation within a single architecture. However, prevailing training paradigms independently optimize understanding via sparse text signals and generation through dense pixel objectives. Such a decoupled strategy yields misaligned representation spaces, isolating visual understanding from generation and hindering their mutual reinforcement. This work presents the first systematic investigation into generative post-training, where we formulate hierarchical visual tasks as generative proxies to bridge the isolation in UMMs. Our empirical investigation reveals that high-level semantic tasks, particularly image segmentation, serve as optimal proxies. Unlike low-level tasks that distract models with texture details, segmentation provides structural semantics that significantly enhance both vision-centric perception and generative layout fidelity. Building upon these insights, we introduce **S**emantic **G**enerative **T**uning (**SGT**), a novel paradigm that leverages segmentation as a generative proxy to align and synergize multimodal capabilities. Mechanistic analyses further demonstrate that SGT fundamentally improves feature linear separability and optimizes visual-textual attention allocation pattern. Extensive evaluations show that SGT consistently improves both multimodal comprehension and generative fidelity across mainstream benchmarks. Our code is available on the Project Page.

Keywords: Unified Multimodal Models · Visual Understanding and Generation · Generative Tuning

1 Introduction

The rapid progress of multimodal models [3, 22, 27, 36, 58] has been fundamentally shaped by distinct research trajectories for understanding and generation. For understanding, models like LLaVA [36] formulate visual comprehension as a text-generation process, leveraging cross-modal alignment to map visual features into linguistic spaces for complex understanding and reasoning. As for generation, studies emphasize generative modeling [3, 13, 87], where diffusion-based architectures have established state-of-the-art performance in high-fidelity content

⋆ Corresponding author.

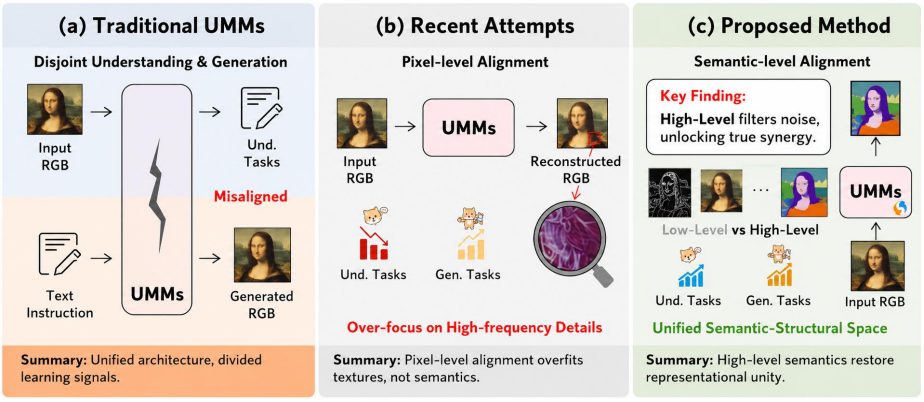


Fig. 1: Comparison of alignment strategies for UMMs. (a) Traditional UMMs optimize understanding and generation tasks separately, resulting in low synergy. (b) Recent pixel-level attempts [80] over-focus on high-frequency details, bringing suboptimal alignment. (c) Our proposed SGT achieves semantic-level alignment, filtering low-level noise and enabling true synergy between understanding and generation.

synthesis. While these specialized architectures exhibit significant proficiency within their respective domains, the emergent trend toward UMMs seeks to consolidate both visual comprehension and generation within a single streamlined framework [32, 49, 64, 72, 83]. This architectural convergence holds the potential to facilitate the transfer of bidirectional knowledge and foster mutual reinforcement between understanding and generation [8, 10, 24, 25, 46]. Consequently, this deep integration unlocks advanced capabilities, including interleaved image-text generation and in-context visual editing, establishing a robust foundation for general-purpose multimodal systems [48, 53].

Despite the structural unification, prevailing training paradigms optimize understanding and generation through divergent supervisory signals as shown in Fig. 1(a). Understanding tasks are predominantly driven by sparse text supervision (e.g., VQA datasets), while generative capabilities are optimized via low-level visual objectives (e.g., pixel or visual token reconstruction). This decoupled training strategy isolates two capabilities and hinders the model from capturing the inherent dependencies between visual understanding and generation. Consequently, UMMs often fail to achieve true mutual reinforcement, leaving the framework with a shared architecture but disjointed optimization processes.

As illustrated in Fig. 1(b), recent attempts [80] address this optimization divergence by employing visual reconstruction in the pixel space as a proxy task. Although this approach yields measurable improvements in generative capabilities, it remains questionable whether low-level visual reconstruction serves as the optimal proxy for synergizing understanding and generation. Since robust visual comprehension inherently relies on semantic information rather than the memorization of low-level textures [1], optimizing for pixel-perfect reconstruction compels the architecture to focus on irrelevant granular details. This distraction inherently limits the model's capacity to enhance visual understanding.

To resolve this critical inquiry, we conduct the first systematic investigation to evaluate the efficacy of various visual proxies in coupling understanding and generation as shown in Fig. 3a and Fig. 3b. Specifically, we establish a hierarchical taxonomy of visual objectives comprising low-level, mid-level, and high-level tasks. Each level encapsulates distinct degrees of spatial granularity and semantic information. This empirical investigation reveals that high-level semantic tasks, particularly image segmentation, serve as the optimal proxy. Unlike low-level tasks that over-emphasize textures, segmentation inherently aligns with the semantic demands of visual comprehension.

Guided by these findings, we introduce **S**emantic **G**enerative **T**uning (**SGT**) for UMMs, as illustrated in Fig. 1(c). This training paradigm leverages image segmentation as a generative proxy to tightly couple visual understanding and generation. To elucidate the underlying mechanisms, we investigate feature distributions and attention dynamics. Our analysis reveals that SGT fundamentally improves feature linear separability and optimizes visual-textual attention allocation. Consequently, this framework effectively enhances both vision-centric perception and generative layout fidelity across mainstream architectures and benchmarks.

The main contributions of this work are summarized as follows.

- We systematically explore generative tuning by formulating various visual tasks as generative proxies. Our analysis reveals that high-level semantic tasks, particularly image segmentation, significantly outperform low-level reconstruction in synergizing visual understanding and generation.
- Guided by these insights, we introduce SGT, a novel paradigm that leverages segmentation as a generative proxy to synergize multimodal capabilities. Mechanistic analyses further demonstrate that SGT fundamentally improves feature separability and optimizes visual-textual attention allocation.
- Extensive evaluations across mainstream UMM architectures validate the efficacy of SGT. By effectively mitigating representational misalignment, the proposed paradigm yields consistent improvements in both visual understanding and generation across diverse benchmarks. Specifically, the framework achieves a 6.02% performance increase over BAGEL [9] on the CV-Bench [60] evaluation and attains a 90.0% score on the GenEval [19].

2 Related Work

2.1 Unified Multimodal Models

Recent UMMs [40,44,75,78] focus on any-to-any processing within a single backbone through two primary trajectories. The first trajectory [18, 67] utilizes discrete visual tokenization and decoder-only autoregression to implement a unified next-token prediction framework. Models such as Emu3 [67], Janus-Pro [6], and VARGPT [93] support interleaved reasoning and mixed-modal generation through this paradigm. The second trajectory [9, 69, 73] employs hybrid architectures that combine causal language modeling with denoising objectives to

maintain synthesis quality while unifying reasoning, as demonstrated by Show-o [81,82] and Transfusion [92]. Research on representation and fusion, including TokenFlow [52] and Chameleon [57], further addresses the balance between semantic abstraction and structural integrity. These works collectively demonstrate that unified training and architectural convergence are essential for bridging the gap between semantic understanding and high-fidelity generation.

2.2 Representation Learning via Generative Objectives

Recent research has explored the utility of generative models, particularly diffusion [13, 15, 50, 71], for visual representation learning [11, 74, 85]. Initial approaches [43, 54, 59] utilize diffusion models as data augmenters to synthesize diverse training samples, thereby improving zero-shot classification and downstream recognition performance. Beyond data augmentation, several frameworks [7, 17, 20, 23, 70] reformulate generative processes as self-supervised objectives. For instance, SODA [23] optimizes semantic features through a diffusion-based bottleneck, while DDAE [70] interprets diffusion as a form of masked autoencoding for reconstruction-based learning. Recent evidence [30, 65, 68, 84, 90] further indicates that intermediate generative features capture rich semantic information that can complement contrastive representations or be directly transferred to recognition tasks. While existing efforts primarily focus on pixel-space reconstruction [26, 45, 63, 80, 88] to bolster visual representations for recognition or synthesis, our work introduces a systematic investigation into how classical visual tasks influence UMMs.

2.3 Reconstruction for Understanding and Alignment

Existing frameworks such as ReCA [80], DIVA [66], ROSS [63], and GenHancer [45] rely on exact pixel reconstruction to enhance model performance. We fundamentally diverge from this paradigm by abandoning raw pixel recovery to eliminate inherent representational redundancy. Crucially, we present the first systematic validation of how hierarchical visual proxy tasks impact the generative tuning of UMMs. By establishing this comprehensive taxonomy, we conclusively demonstrate that advanced visual tasks deliver the maximum performance improvements. Furthermore, while contemporary studies like UniMRG [55] explore isolated proxy tasks and Metamorph [61] observes the mutual influence between perception and synthesis, our work actively bridges the gap between discriminative and generative capabilities. This unified optimization explicitly establishes a shared semantic space to capture the structural abstraction essential for general purpose multimodal learning.

3 Semantic Generative Tuning

This section outlines the whole framework. It begins by formalizing the preliminaries of UMMs in Sec. 3.1. Then, Sec. 3.2 details the training strategies applied

to representative architectures such as BAGEL [9] and OmniGen2 [73]. For systematic evaluation over understanding and generative capabilities, Sec. 3.3 introduces a hierarchical suite of tasks within a generative tuning framework and assesses their influence on six core understanding metrics as well as generative performance.

3.1 Formulation

UMMs aim to integrate diverse modalities within a single architecture f_θ by mapping inputs from the textual space $\mathcal{T}$ and image space $\mathcal{I}$ into a shared representation space. Formally, given a text prompt $x \in \mathcal{T}$ and an optional reference image $v \in \mathcal{I}$, the model processes various tasks through different input combinations. For visual understanding tasks, UMMs typically process an input image using a semantic vision encoder and subsequently integrate the extracted features with language tokens for unified treatment within a language model. In the case of visual editing tasks, certain frameworks [6,9,49,73,76] supplement the semantic vision encoder with a variational autoencoder (VAE) to preserve fine-grained image details as well as to ensure identity consistency and high-quality generation.

Without loss of generality, we employ a dual encoder architecture as an illustrative example to introduce the general formulation of UMMs. Specifically, a ViT-based encoder $\Phi_{vit}(\cdot)$ extracts semantic tokens $z_{vit} \in \mathbb{R}^{L\times D}$ for multimodal reasoning, while a VAE-based encoder $\Phi_{vae}(\cdot)$ encodes the image into a latent space $z_{vae} \in \mathbb{R}^{H\times W\times C}$ to maintain structural and textural details. The mapping for these tasks is formulated as follows

$$y = \begin{cases} f_\theta(x, [z_{vit}]) & \text{Understanding: } y \in \mathcal{T} \\ f_\theta(x, [z_{noise}]) & \text{Generation: } y \in \mathcal{I} \\ f_\theta(x, [z_{vit}, z_{vae}, z_{noise}]) & \text{Editing: } y \in \mathcal{I} \end{cases} \tag{1}$$

where $[\cdot]$ denotes the set of optional inputs and z_{noise} represents the initial Gaussian noise utilized for generative processes. This formulation categorizes the operational scope of UMMs into three distinct functional paradigms. For visual understanding, the model leverages semantic features z_{vit} to generate textual responses $y \in \mathcal{T}$. In the context of visual generation, the model maps a text prompt x and the initial noise z_{noise} to a synthesized image $y \in \mathcal{I}$. For visual editing tasks, the framework integrates z_{vit}, z_{vae}, and the stochastic component z_{noise} to achieve high-fidelity image manipulation. Such a structure simultaneously yields representations across varying granularities to establish a robust foundation for UMMs.

3.2 Motivation and Hierarchical Visual Task Taxonomy

Recent advances [4,33,45,63,66] indicate that reconstructing visual inputs from learned embeddings significantly enhances the representation quality of visual embeddings. However, pixel-space reconstruction fundamentally optimizes image

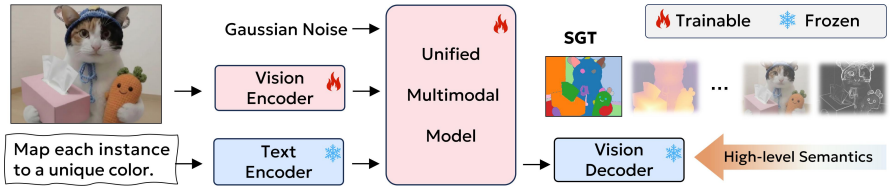


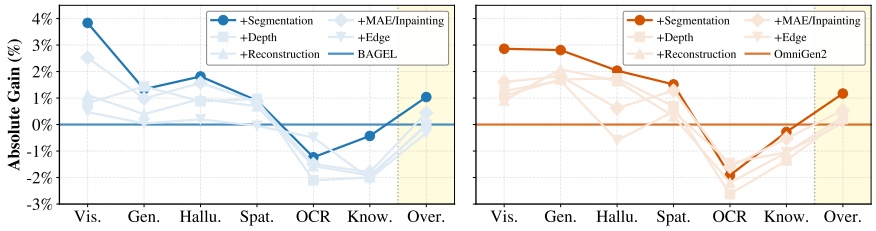
Fig. 2: Overview of the generative tuning paradigm. An RGB image and a concise textual instruction are processed by respective vision and text encoders to extract independent embeddings. UMMs then integrate these embeddings and map the representations to the designated task. Because empirical evaluations demonstrate that visual generation targets at an advanced semantic level yield the most significant performance gains, SGT explicitly adopts image segmentation as its generative objective.

fidelity rather than cross-modal semantic alignment, and its objective is not invariably the most relevant for visual understanding and reasoning. Driven by this insight, we pose the question of whether pixel-space reconstruction is truly the optimal choice for UMMs.

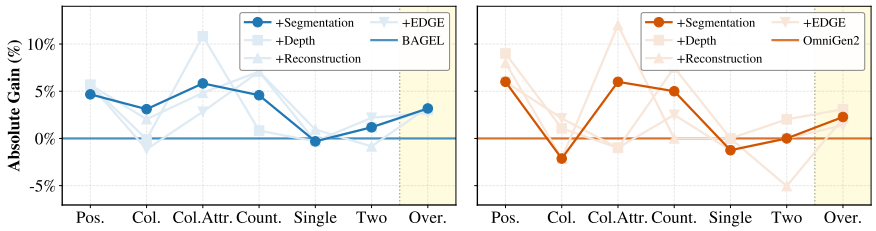
In response to this question, we establish a hierarchical taxonomy to investigate the impact of different levels of visual tasks on UMMs within the generative tuning framework. Formally, we model the generative tuning as a conditional generation process $y = f_\theta(x, [z_{\text{vit}}, z_{\text{noise}}])$, where the output y resides in the visual space. We define the training objective as $\mathcal{L} = L(f_\theta(x, [z_{vit}, z_{noise}]), \hat{y})$, where x denotes a concise natural language instruction tailored to the specific task, and $\hat{y}$ represents the target visual representation as depicted in Fig. 2. Here, $\hat{y}$ denotes the ground truth for diverse visual tasks. Crucially, to isolate the impact of task granularity, we exclusively utilize visual data for generative tuning during this investigative phase, strictly excluding other data types such as visual question answering, text-to-image generation, or standard image editing data. To ensure a rigorous comparison, all tasks are evaluated using the same set of input RGB images and an identical volume of training data. Specifically, our evaluation covers high-level tasks (segmentation, object detection), mid-level tasks (depth estimation, inpainting), and low-level tasks (edge detection).

3.3 From Empirical Observations to the SGT Paradigm

We begin by evaluating visual proxy tasks across different levels based on empirical model performance variations. To establish a comprehensive and systematic evaluation protocol, we draw inspiration from the taxonomy proposed in Cambrian-1 [60]. Specifically, we augment the original categories of general VQA [5, 89], vision-centric perception [60, 62], chart/OCR [39, 47], and mathematical reasoning [41, 42] with spatial reasoning [35, 86] and hallucination resistance [21, 31] to enable a more holistic assessment. Each capability score is derived from the unweighted average of two representative benchmarks. Generative capabilities are evaluated via GenEval [19]. We validate our findings across both BAGEL [9] and OmniGen2 [73] to ensure architectural generalizability,



(a) Understanding capability gains



(b) Generation capability gains

Fig. 3: Empirical evaluation of the hierarchical task ladder across diverse understanding and generation dimensions. (a) High-level proxy tasks yield greater performance gains than low-level tasks in multimodal understanding. (b) Various generative objectives consistently improve performance in the position dimension, yielding comparable overall gains. (From left to right): Position, Colors, Color Attributes, Counting, Single Object, Two Objects, and Overall. The results represent the average performance computed across twelve random seeds.

with specific model details provided in Sec. 4.1. Our empirical analysis yields three crucial observations, as visualized in Fig. 3a and Fig. 3b.

Observation 1: High-level semantic tasks outperform low-level cues. Our analysis indicates that high-level tasks yield substantially greater benefits for multimodal understanding than their mid- or low-level counterparts. As evidenced in Fig. 3a, high-level objectives such as image segmentation consistently outperform mid-level tasks (e.g., depth estimation) and low-level tasks (e.g., edge detection). We attribute this to the strong alignment between high-level semantics and the reasoning requirements of understanding models. High-level supervision encourages the extraction of semantic and structural essence, whereas low-level tasks may compel the model to overfit to intricate textural details that are often redundant for complex reasoning. This observation aligns with findings in GenHancer [45] and the design philosophy of I-JEPA [1].

Observation 2: Visual supervision enhances perception, not reasoning. The generative tuning paradigm predominantly fortifies fundamental visual perception rather than linguistic priors or abstract logical reasoning. While we observe significant performance gains in vision-centric tasks, spatial reasoning, and hallucination resistance, capabilities in chart recognition and mathematical knowledge remain static or exhibit marginal decline, as shown in Fig. 3a. This

Table 1: Statistics of the training data.

Data Source	SGT	General	Doc/Chart/Screen	Math/Reasoning	General OCR	Language
Number	190k	180k	103k	101k	45k	72k

divergence indicates that while visually-derived supervision enhances representation quality to boost perceptual capabilities, it does not impart additional knowledge or logical reasoning skills.

Observation 3: Various proxy tasks consistently improve spatial fidelity. Diverging from the trends associated with varying granularities observed in understanding benchmarks, the generative tuning paradigm consistently enhances overall generation quality. Moreover, as illustrated in Fig. 3b, the model demonstrates consistent performance gains on position-aware tasks. This suggests that visual proxy tasks inherently provide explicit spatial constraints, regardless of their semantic granularity. Empirically, the process of reconstructing these visual structures forces the model to maintain accurate spatial layouts, thereby naturally enhancing its alignment with positional prompts. This observation aligns with insights reported in RecA [80].

Synthesizing these three observations, we conclude that within the generative tuning framework, employing high-level semantic proxy tasks for generative tuning yields optimal enhancements for UMMs. Consequently, we advocate for a novel training paradigm termed Semantic Generative Tuning (SGT). This approach strategically leverages high-level visual proxies, especially image segmentation, to refine the internal representations of UMMs, thereby harmonizing visual understanding and generation within a unified framework. Additional experiments show that semantic instance and panoptic segmentation, as well as class-agnostic segmentation, consistently yield comparable improvements.

4 Experiments

We first detail the experimental configurations and the selection of models in Sec. 4.1. Sec. 4.2 presents a unified study that (i) benchmarks our approach against state-of-the-art UMMs on diverse understanding and generation tasks and (ii) evaluates alternative visual proxy tasks. Furthermore, we investigate the optimal data recipe and the scaling properties in Sec. 4.3. In Sec. 4.4, we analyze how the SGT paradigm alters the feature space and attention allocation of UMMs, in order to uncover deeper underlying causes.

4.1 Experimental Setup

Datasets. Although Sec. 3.3 confirms that semantic generative tuning is highly effective in isolation, we further construct a holistic post-training to fully unleash the potential of SGT. By synergizing SGT with 500k supervised fine-tuning samples from LLaVA-OneVision [29], we demonstrate its robustness and scalability.

To strictly preclude data overlap between the training and evaluation phases, we source all images for SGT exclusively from the SAM [28] dataset. Specifically, we curate 190k samples for the SGT dataset, with the detailed source distribution outlined in Table 1. Regarding the VQA data, we align data mixture with the official recipe provided by LLaVA-OneVision [29].

Model selection. We conduct our experiments on two mainstream UMM architectures, BAGEL [9] and OmniGen2 [73], to evaluate our method across distinct design philosophies. Beyond an approximate twofold difference in parameter scale, these models differ fundamentally in their feature interaction mechanisms and training paradigms. Specifically, BAGEL adopts a Mixture of Transformers framework to facilitate layer-wise feature sharing throughout the network. Conversely, OmniGen2 utilizes hidden states from the understanding module as semantic guidance to steer the generative process. Their training strategies also diverge considerably, as BAGEL employs a native interleaved training process, whereas OmniGen2 pairs a frozen pre-trained vision language model [2] with a diffusion module trained from scratch. To further validate the universality of the SGT paradigm beyond these UMMs, we extend our preliminary evaluation to single visual encoder architectures. This architectural diversity ensures the broad applicability of SGT.

Evaluation benchmarks. To comprehensively assess multimodal understanding, we utilize the VLMEvalKit [12] to evaluate model performance across a diverse suite of benchmarks. This carefully curated selection encompasses spatial reasoning, robustness against hallucinations, general visual question answering, knowledge reasoning, and vision-centric perception to ensure a holistic evaluation. Specifically, we conduct these assessments on CV-Bench [60], MMVP [62], VSR [35], SIBench-mini [86], POPE [31], Hallusion [21], MMBench-TEST-EN 1.1 [38], MMMU-val [89], RWQA [79], MathVista [41], BLINK [16], MME [14], and MMStar [5]. Furthermore, we employ GenEval [19] and GEdit-Bench-En [37] to measure text-to-image generation and image editing capabilities respectively.

4.2 Main Results

Comparison with state-of-the-art UMMs. We present a comprehensive comparison between our proposed models and existing leading UMMs in Table 2. SGT-BAGEL and SGT-Gen2 represent the enhanced variants of BAGEL and OmniGen2. We train these variants using segmentation data from the SAM dataset [28] alongside visual understanding instruction tuning. Quantitative evaluations indicate that both SGT-BAGEL and SGT-Gen2 consistently outperform their original baseline architectures and surpass a broad range of competitive models across multiple benchmarks. This widespread superiority demonstrates the efficacy of integrating high-level semantic generative objective into the fine-tuning of UMMs. Furthermore, our framework achieves favorable performance in generative tasks. As Fig. 4 illustrates, SGT demonstrates superior adherence to complex textual prompts including spatial and color instructions when compared to the baseline model. Such qualitative improvements confirm that SGT exerts a synergistic benefit on the overall capabilities of UMMs.

Table 2: Comparison with state-of-the-art UMMs. Best results are in **bold**, second best are underlined. "–" indicates not reported. ✗ indicates the model does not support image editing. † refers to methods using LLM rewriter. *Results are taken from previous works. ‡ indicates that the first term denotes understanding parameters, and the second denotes generation parameters.

Model	Params	Visual Understanding						Visual Generation	
		MMVP	VSR	Hallu.	MMStar	RWQA	MathV.	GenEval	GEdit-Bench-En
Small-scale Models (≤4B)									
*Show-o$_{512}$ [81]	1.3B	50.00	54.26	46.06	–	38.17	–	68.0	✗
Harmon [77]	1.5B	60.00	60.88	46.69	38.00	48.00	33.70	73.0	✗
ReCA-Harmon [80]	1.5B	47.00	–	36.70	25.53	43.53	24.50	**90.0**	✗
*UniLIP [56]	2B	73.00	65.55	60.57	–	64.18	–	**90.0**	–
*UniMRG [55]	3.6B	**74.67**	73.90	**64.56**	–	**66.01**	–	55.8	✗
*OpenUni [76]	2B	71.67	66.69	60.88	–	65.23	–	51.0	-
OmniGen2 [73]	3B+4B‡	65.00	77.52	62.35	55.07	64.41	63.50	76.6	6.63
SGT-Gen2	3B+4B‡	68.33	**78.85**	64.25	**57.07**	65.10	**64.00**	78.9	**6.83**
Large-scale Models (≥7B)									
Chameleon [57]	7B	50.00	–	31.13	28.93	39.00	21.90	39.0	✗
Janus-Pro [6]	7B	63.00	71.03	60.15	46.80	41.83	42.60	80.0	✗
*Emu3 [67]	8B	–	–	–	–	57.40	–	66.0†	✗
UniWorld-v1 [34]	7B+12B‡	77.67	**83.34**	68.35	63.90	67.58	68.20	84.0†	4.85
BAGEL [9]	7B+7B‡	83.00	80.45	68.34	67.46	71.26	73.10	88.0†	6.64
SGT-BAGEL	7B+7B‡	**83.33**	81.54	**70.24**	**68.33**	**72.42**	**73.90**	**90.0**†	**6.94**

Ablation study. We conduct comprehensive ablation studies to systematically evaluate the isolated impact of SFT data alongside its joint training dynamics with visual tasks across varying semantic levels. As shown in Table 3, SFT+SGT utilizes the segmentation task from the SAM dataset as the generative target, whereas SFT+Reconstruction and SFT+Edge employ image reconstruction and edge detection as their respective proxy tasks. All three tasks yield performance gains across the majority of perception-centric understanding benchmarks. We observe notable improvements in vision-focused evaluations such as MMVP and CV-Bench, spatial reasoning assessments including VSR and SIBench-mini, and various hallucination robustness tests. Crucially, SGT yields the most substantial performance gains among the evaluated proxy tasks. This outcome directly corroborates the findings detailed in Sec. 3.3 and validates this semantic approach as the optimal target for generative tuning. In generative evaluations, all three proxy tasks achieve comparable gains in text-to-image synthesis, while gains in image editing are positive but smaller in magnitude. This discrepancy suggests that while generative tuning successfully aligns representational spaces, driving further substantial gains in complex generative editing may require the integration of explicit image editing data. Finally, consistent performance improvements observed across both the BAGEL and OmniGen2 architectures underscore the generalizability and robustness of SGT.

4.3 More Explorations

Optimal data recipe. While our analysis in Sec. 3.3 indicates that SGT independently enhances both understanding and generation, we posit that a

Table 3: Unified performance comparison on various benchmarks. The best results in each group are highlighted in **bold**. The results reported for the GenEval benchmark represent the average performance computed across twelve random seeds.

	Vision-Centric		Spatial Reasoning		Hallucination		General			Generation	
Method	CV-Bench	MMVP	VSR	SIBench	POPE	Hallusion	MMBench	MMMU	MMStar	GenEval	GEdit-Bench-En
Base Model: OmniGen2											
OmniGen2 (Base)	65.94	65.00	77.52	43.29	85.97	62.35	77.04	42.11	55.07	76.58	6.63
SFT	65.99	66.00	77.61	44.37	86.25	64.35	77.04	43.22	55.73	74.54	6.32
SFT+Edge	66.67	65.33	77.99	**45.51**	86.10	63.72	76.88	42.78	55.53	77.45	6.79
SFT+Reconstruction	66.71	66.33	78.18	45.41	85.92	**65.19**	77.00	44.44	55.73	77.53	6.81
SFT+SGT	**66.91**	**68.33**	**78.85**	45.37	**87.29**	64.25	**77.09**	**45.89**	**57.07**	**78.86**	**6.83**
Base Model: BAGEL											
BAGEL (Base)	73.21	83.00	80.45	48.95	85.69	68.34	81.25	46.77	67.46	78.21	6.52
SFT	74.61	82.67	80.69	49.34	86.77	67.92	81.86	47.33	66.93	77.18	6.49
SFT+Edge	74.56	**83.67**	80.83	49.51	86.48	68.66	81.89	47.56	67.20	79.96	6.72
SFT+Reconstruction	75.23	83.33	80.83	**50.59**	87.98	68.03	82.18	46.56	67.40	80.82	6.75
SFT+SGT	**79.23**	83.33	**81.54**	50.18	**88.32**	**70.24**	**83.84**	**48.56**	**68.33**	**80.95**	**6.94**

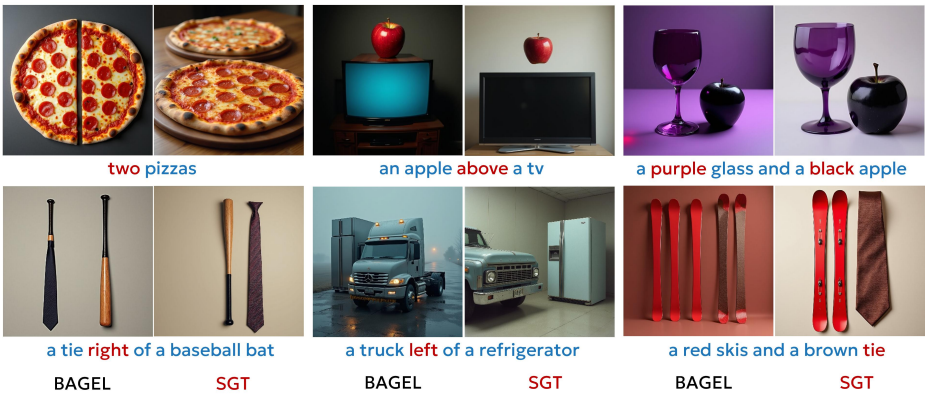
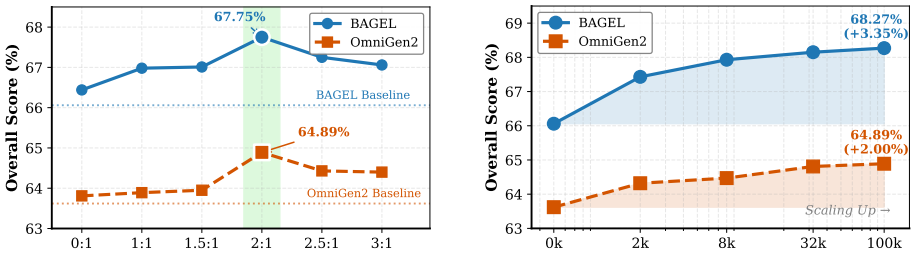


Fig. 4: Qualitative comparison on compositional text-to-image generation.

comprehensive post-training regime must synergize SGT objectives with SFT data to maximize performance. Therefore, we conduct an ablation study to determine the optimal data sampling recipe between VQA instructions and segmentation-based visual targets within each training batch. To ensure a robust assessment, we aggregate performance across eight diverse understanding benchmarks [5, 14, 16, 21, 31, 60, 62, 89] and report the average normalized score. As illustrated in Fig. 5a, a 1:2 intra-batch ratio of VQA to segmentation data yields the most significant improvements in this aggregate metric for both the BAGEL and OmniGen2 architectures. Regarding generative tasks, we observe that performance scales positively with the proportion of generative samples within the training batch. Balancing these multi-faceted requirements, we adopt the 1:2 ratio for our final configuration. We reserve the exploration of more complex tripartite mixing strategies involving understanding, generation, and SGT for future research.



(a) Segmentation-to-VQA Ratio **(b)** Scalability

Fig. 5: Ablation studies on segmentation data integration. (a) **The optimal data mixture.** We analyze the Segmentation-to-VQA ratio within each training batch, observing that both models achieve optimal performance at a 2:1 ratio. (b) **Data scalability.** Performance improves consistently as the segmentation dataset expands from 2k to 100k samples (BAGEL: +3.3%, OmniGen2: +2.0%), confirming that our visual proxy task yields scalable benefits for multimodal understanding.

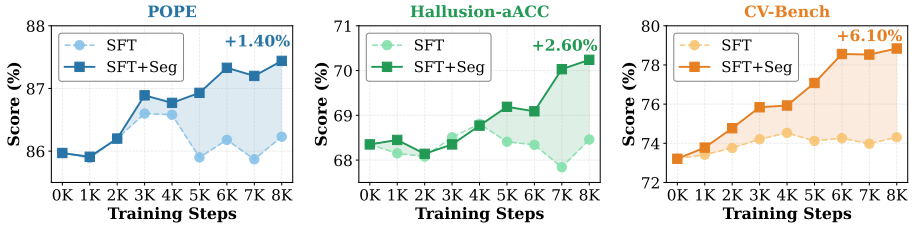


Fig. 6: Training dynamics with different SFT:Seg ratios. We compare the training curves of BAGEL under 1:0 (baseline, no segmentation) and 1:2 (with segmentation data) ratios across three benchmarks.

Scaling properties of SGT. To verify the scalability of SGT, we fix the VQA SFT data and systematically scale the segmentation training data. We report the aggregate performance across the eight representative benchmarks described previously. As illustrated in Fig. 5b, the average normalized score exhibits a monotonic increase commensurate with the volume of segmentation data. Furthermore, an analysis of the training dynamics in Fig. 6 reveals that the integration of segmentation objectives significantly accelerates convergence on challenging benchmarks such as CV-Bench and Hallusion. Compared to the baseline trained exclusively on VQA SFT data, our strategy consistently achieves superior performance during optimization. This demonstrates that SGT serves as a scalable approach to continuously enhance multimodal capabilities.

4.4 Mechanistic Insights: Why Semantic Proxies Unlock Synergy?

To further elucidate the impact of the SGT, we employ BAGEL as a representative architecture to investigate specific representational shifts at both the feature and attention levels. Our analysis examines the model's internal dynamics across

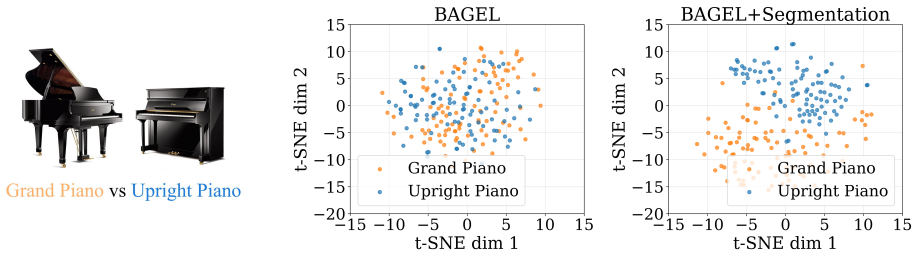
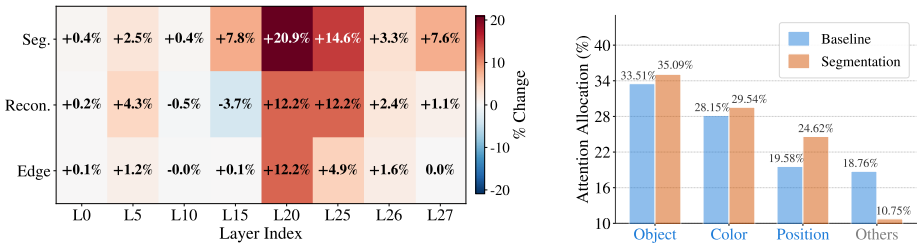


Fig. 7: Feature space analysis on fine-grained classes. Left panels display the visually confusable categories Grand Piano and Upright Piano. The corresponding tSNE visualizations on the right reveal that while the baseline BAGEL yields entangled feature spaces, our proposed BAGEL+Segmentation learns highly discriminative embeddings and achieves clear class separation.



(a) Vision-language attention allocation. **(b)** Key words attention allocation.

Fig. 8: Analysis of attention patterns. (a) Layer-wise changes in attention to visual features for three proxy tasks relative to the BAGEL baseline, demonstrating a consistent increase in visual focus in deeper layers. (b) Attention distribution over text tokens. The segmentation objective effectively enhances the focus on critical tokens (Object, Color, Relation).

three dimensions encompassing the feature space structure of the visual encoder, the cross-modal attention patterns within the understanding module, and the attention distribution during generation.

Finding 1: SGT promotes feature linear separability. We first visualize the visual embeddings z_{vit} using t-SNE as shown in Fig. 7. The projections reveal that training with segmentation data enhances the linear separability of categories that are semantically similar yet structurally distinct, such as upright and grand pianos. In contrast to the baseline model which often yields diffuse clusters, the incorporation of segmentation supervision significantly improves both the intra-class compactness and inter-class separability of the visual representations.

Finding 2: Mitigating linguistic over-reliance. We examine the cross-modal attention dynamics within the understanding module, as shown in Fig. 8a. Specifically, we observe a higher concentration of attention on visual tokens within the deeper transformer layers compared to the baseline. This distribution

indicates that the model anchors its reasoning process more firmly in visual evidence, effectively counteracting the over-reliance on linguistic priors that often leads to hallucination [51, 91]. Crucially, high-level segmentation tasks induce a more pronounced attention shift than low-level objectives.

Finding 3: Amplifying critical tokens and suppressing irrelevant cues. We investigate the generative capability using prompts containing position and attribute constraints sampled from GenEval [19]. We quantify the cross-attention weights allocated to critical tokens, specifically position, color, and object identity. As illustrated in Fig. 8b, the integration of segmentation data amplifies the model's focus on these attribute-specifying tokens. This further demonstrates that SGT effectively narrows the representational gap within UMMs and compels the models to prioritize intrinsically meaningful features.

5 Limitations

While SGT effectively aligns understanding and generation for natural scenes, relying exclusively on segmentation data constrains performance on symbolically dense and knowledge-intensive tasks, as shown in Fig. 3a. This observation indicates that SGT functions best as a foundational alignment strategy rather than a standalone training solution. The paradigm successfully retains its symbolic proficiency when SGT is augmented with VQA data, as shown in Table 2. Future research will explore a comprehensive post-training pipeline integrating the SGT alignment strategy with understanding data, generative targets, and reinforcement learning frameworks to achieve optimal cross-modal performance.

6 Conclusion

This work proposes a fine-tuning paradigm for UMMs to mitigate the optimization divergence between visual understanding and generation. Previous attempts leverage pixel space reconstruction to improve multimodal alignment but inadvertently introduce granular visual noise that ultimately leads to suboptimal performance. To overcome this limitation, we introduce Semantic Generative Tuning as a novel paradigm that shifts the alignment proxy from the pixel space to the semantic space. Mechanistic analyses reveal that this semantic integration fundamentally improves feature linear separability and optimizes attention allocation to directly mitigate representational misalignment. Extensive empirical evaluations across mainstream architectures demonstrate that SGT consistently yields significant improvements in both visual understanding accuracy and generative layout fidelity. The principles established by this paradigm highlight that aligning multimodal capabilities at the semantic level serves as a crucial foundation for developing cohesive and versatile UMMs.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. arXiv:2301.08243 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv:2502.13923 (2025)
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators, accessed 25 June 2026
4. Chen, F., Jing, M., Lu, W., Feng, Y., Li, X., Cao, X.: Unihetero: Could generation enhance understanding for vision-language-model at large data scale? arXiv:2512.23512 (2025)
5. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? NeurIPS **37**, 27056–27087 (2024)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv:2501.17811 (2025)
7. Chen, X., Liu, Z., Xie, S., He, K.: Deconstructing denoising diffusion models for self-supervised learning. arXiv:2401.14404 (2024)
8. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. NeurIPS **36**, 49250–49267 (2023)
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv:2505.14683 (2025)
10. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: DreamLLM: Synergistic multimodal comprehension and creation. In: ICLR (2024), https://openreview.net/forum?id=y01KGvd9Bw, accessed 25 June 2026
11. Du, S., Guo, J., Li, B., Cui, S., Xu, Z., Luo, Y., Wei, Y., Gai, K., Wang, X., Wu, K., et al.: Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. arXiv:2511.23386 (2025)
12. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: ACMMM. pp. 11198–11201 (2024)
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
14. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394 (2023)
15. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: ECCV. pp. 241–258. Springer (2024)

16. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. arXiv:2404.12390 (2024)
17. Fuest, M., Ma, P., Gui, M., Schusterbauer, J., Hu, V.T., Ommer, B.: Diffusion models and representation learning: A survey. arXiv:2407.00783 (2024)
18. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv:2404.14396 (2024)
19. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. NeurIPS **36**, 52132–52152 (2023)
20. Graikos, A., Yellapragada, S., Le, M.Q., Kapse, S., Prasanna, P., Saltz, J., Samaras, D.: Learned representation-guided diffusion models for large-image generation. In: CVPR. pp. 8532–8542 (2024)
21. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR. pp. 14375–14385 (2024)
22. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: CVPR. pp. 15733–15744 (2025)
23. Hudson, D.A., Zoran, D., Malinowski, M., Lampinen, A.K., Jaegle, A., McClelland, J.L., Matthey, L., Hill, F., Lerchner, A.: Soda: Bottleneck diffusion models for representation learning. In: CVPR. pp. 23115–23127 (2024)
24. Jin, Y., Sun, Z., Xu, K., Chen, L., Jiang, H., Huang, Q., Song, C., Liu, Y., Zhang, D., Song, Y., et al.: Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. arXiv:2402.03161 (2024)
25. Jin, Y., Xu, K., Chen, L., Liao, C., Tan, J., Huang, Q., Chen, B., Lei, C., Liu, A., Song, C., et al.: Unified language-vision pretraining in llm with dynamic discrete visual tokenization. arXiv:2309.04669 (2023)
26. Ju, H., Huang, S., Li, H., Ding, Z., Liu, S., Wang, M., Zheng, Z.: From instruction to event: Sound-triggered mobile manipulation. arXiv:2601.21667 (2026)
27. Ju, H., Huang, S., Liu, S., Zheng, Z.: Video2bev: Transforming drone videos to bevs for video-based geo-localization. In: ICCV. pp. 27073–27083 (2025)
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
29. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv:2408.03326 (2024)
30. Li, J., Yu, S., Wang, Y., Wang, L., Lu, H.: Selm: Selective mechanism based audio-visual segmentation. In: ACM MM. pp. 3926–3935 (2024)
31. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP. pp. 292–305 (2023)
32. Li, Z., Liu, Z., Zhang, Q., Lin, B., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv:2510.16888 (2025)
33. Liao, X., He, Q., Xu, K., Qu, X., Li, Y., Wei, W., Yao, A.: Va-π: Variational policy alignment for pixel-aware autoregressive generation. arXiv:2512.19680 (2025)
34. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv:2506.03147 (2025)

35. Liu, F., Emerson, G.E.T., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics (2023)
36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
37. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1x-edit: A practical framework for general image editing. arXiv:2504.17761 (2025)
38. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: ECCV. pp. 216–233. Springer (2024)
39. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
40. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv:2312.17172 (2023)
41. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: ICLR (2024)
42. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
43. Luo, R., Li, Y., Chen, L., He, W., Lin, T.E., Liu, Z., Zhang, L., Song, Z., Xia, X., Liu, T., et al.: Deem: Diffusion models serve as the eyes of large language models for image perception. arXiv:2405.15232 (2024)
44. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv:2502.20321 (2025)
45. Ma, S., Ge, Y., Wang, T., Guo, Y., Ge, Y., Shan, Y.: Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. In: ICCV. pp. 24402–24412 (2025)
46. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: CVPR. pp. 7739–7751 (2025)
47. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV. pp. 2200–2209 (2021)
48. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Ning, K., Feng, C., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv:2503.07265 (2025)
49. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., Hou, J., Xie, S.: Transfer between modalities with metaqueries. arXiv:2504.06256 (2025)
50. Parihar, R., Sachidanand, V., Mani, S., Karmali, T., Venkatesh Babu, R.: Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In: ECCV. pp. 469–487. Springer (2024)
51. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: CVPR (2022)
52. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: CVPR. pp. 2545–2555 (2025)

53. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Lmfusion: Adapting pretrained language models for multimodal generation. arXiv:2412.15188 (2024)
54. Shipard, J., Wiliem, A., Thanh, K.N., Xiang, W., Fookes, C.: Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In: CVPR. pp. 769–778 (2023)
55. Su, Z., Wei, H., Cen, K., Wang, Y., Chen, G., Yuan, C., Chu, X.: Generation enhances understanding in unified multimodal models via multi-representation generation. arXiv:2601.21406 (2026)
56. Tang, H., Xie, C., Bao, X., Weng, T., Li, P., Zheng, Y., Wang, L.: Unilip: Adapting clip for unified multimodal understanding, generation and editing. arXiv:2507.23278 (2025)
57. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv:2405.09818 (2024)
58. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. NeurIPS **37**, 84839–84865 (2024)
59. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. NeurIPS **36**, 48382–48402 (2023)
60. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. NeurIPs **37**, 87310–87356 (2024)
61. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv:2412.14164 (2024)
62. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: CVPR. pp. 9568–9578 (2024)
63. Wang, H., Zheng, A., Zhao, Y., Wang, T., Ge, Z., Zhang, X., Zhang, Z.: Reconstructive visual instruction tuning. arXiv:2410.09575 (2024)
64. Wang, P., Peng, Y., Gan, Y., Hu, L., Xie, T., Wang, X., Wei, Y., Tang, C., Zhu, B., Li, C., et al.: Skywork unipic: Unified autoregressive modeling for visual understanding and generation. arXiv:2508.03320 (2025)
65. Wang, S., Pei, M., Sun, L., Deng, C., Li, Y., Shao, K., Tian, Z., Zhang, H., Wang, J.: Spatialviz-bench: A cognitively-grounded benchmark for diagnosing spatial visualization in mllms. In: ICLR (2025)
66. Wang, W., Sun, Q., Zhang, F., Tang, Y., Liu, J., Wang, X.: Diffusion feedback helps clip see better. arXiv:2407.20171 (2024)
67. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv:2409.18869 (2024)
68. Wang, Y., Schiff, Y., Gokaslan, A., Pan, W., Wang, F., De Sa, C., Kuleshov, V.: Infodiffusion: Representation learning using information maximizing diffusion models. In: ICML. pp. 36336–36354. PMLR (2023)
69. Wang, Z., Chen, Z., Gou, C., Li, F., Deng, C., Zhu, D., Li, K., Yu, W., Tu, H., Fan, H., et al.: Lightfusion: A light-weighted, double fusion framework for unified multimodal understanding and generation. arXiv:2510.22946 (2025)
70. Wei, C., Mangalam, K., Huang, P.Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A., Feichtenhofer, C.: Diffusion models as masked autoencoders. In: ICCV. pp. 16284–16294 (2023)

71. Weng, N., Pegios, P., Petersen, E., Feragen, A., Bigdeli, S.: Fast diffusion-based counterfactuals for shortcut removal and generation. In: ECCV. pp. 338–357. Springer (2024)
72. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv:2410.13848 (2024)
73. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv:2506.18871 (2025)
74. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. arXiv:2507.01467 (2025)
75. Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., Bai, X.: Liquid: Language models are scalable and unified multi-modal generators. IJCV (2025)
76. Wu, S., Wu, Z., Gong, Z., Tao, Q., Jin, S., Li, Q., Li, W., Loy, C.C.: Openuni: A simple baseline for unified multimodal understanding and generation. arXiv:2505.23661 (2025)
77. Wu, S., Zhang, W., Xu, L., Jin, S., Wu, Z., Tao, Q., Liu, W., Li, W., Loy, C.C.: Harmonizing visual representations for unified multimodal understanding and generation. In: ICCV. pp. 17739–17750 (2025)
78. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. arXiv:2409.04429 (2024)
79. xAI: Grok-1.5 vision preview. https://x.ai/news/grok-1.5v (2024), Accessed 25 June 2026
80. Xie, J., Darrell, T., Zettlemoyer, L., Wang, X.: Reconstruction alignment improves unified multimodal models. arXiv:2509.07295 (2025)
81. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv:2408.12528 (2024)
82. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv:2506.15564 (2025)
83. Yang, J., Yin, D., Zhou, Y., Rao, F., Zhai, W., Cao, Y., Zha, Z.J.: Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling. In: CVPR. pp. 7974–7985 (2025)
84. Yang, X., Wang, X.: Diffusion model as representation learner. In: ICCV. pp. 18938–18949 (2023)
85. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv:2410.06940 (2024)
86. Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al.: How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv:2509.18905 (2025)
87. Yu, S., Chen, Y., Qi, Z., Xie, Z., Wang, Y., Wang, L., Shan, Y., Lu, H.: Mono2stereo: A benchmark and empirical study for stereo conversion. In: CVPR. pp. 21847–21856 (2025)
88. Yu, S., Wang, Y., Zhuge, Y., Wang, L., Lu, H.: Dme: Unveiling the bias for better generalized monocular depth estimation. In: AAAI. vol. 38, pp. 6817–6825 (2024)
89. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal un-

derstanding and reasoning benchmark for expert agi. In: CVPR. pp. 9556–9567 (2024)
90. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: ICCV. pp. 5729–5739 (2023)
91. Zheng, X., Liao, C., Fu, Y., Lei, K., Lyu, Y., Jiang, L., Ren, B., Chen, J., Wang, J., Li, C., et al.: Mllms are deeply affected by modality bias. arXiv:2505.18657 (2025)
92. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv:2408.11039 (2024)
93. Zhuang, X., Xie, Y., Deng, Y., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv:2501.12327 (2025)