Enhancing Alignment for Unified Multimodal Models via Semantically-Grounded Supervision

Jiyeong Kim[1], Yerim So[1], Hyesong Choi[2], Uiwon Hwang[1], and Dongbo Min[1†]

[1] Division of AI and Software, Ewha Womans University, South Korea
[2] Soongsil University, South Korea
{wldud8946, yrso, uiwon.hwang, dbmin}@ewha.ac.kr, hyesong@ssu.ac.kr

Abstract. Unified Multimodal Models (UMMs) have emerged as a promising paradigm that integrates multimodal understanding and generation within a unified modeling framework. However, current generative training paradigms suffer from inherent limitations. We present **Se**mantically-**Gro**unded **S**upervision (**SeGroS**), a fine-tuning framework designed to resolve the granularity mismatch and supervisory redundancy in UMMs. At its core, we propose a novel visual grounding map to construct two complementary supervision signals. First, we formulate semantic *Visual Hints* to compensate for the sparsity of text prompts. Second, we generate a semantically-grounded *Corrupted Input* to explicitly enhance the supervision of masking-based UMMs by restricting the Text-to-Image loss to core text-aligned regions. Extensive evaluations on GenEval, DPGBench, and CompBench demonstrate that SeGroS significantly improves generation fidelity and cross-modal alignment across various UMM architectures.
Project page: https://segros-project.github.io/

Keywords: Unified Multimodal Models · Text-to-Image Generation · Multimodal Alignment

1 Introduction

The evolution of Multimodal Large Language Models (MLLM) [2, 33] has shifted the prevailing paradigm from text-only processing [3, 12] to multimodal reasoning, enabling models to jointly interpret visual and linguistic signals. In extending to handle both understanding and image generation, early efforts often relied on combining independent models (*e.g.*, an LMM for reasoning and a dedicated image generator) [25, 40]. However, such decoupled designs suffer from weak integration between modalities [42] and incur engineering costs to maintain two separate models independently [43]. This led to the emergence of **U**nified **M**ultimodal **M**odels (**UMM**s) [15, 31, 32, 34, 37, 38], which integrate multimodal understanding and generation within a single sequence modeling framework by representing visual signals as tokenized inputs analogous to text tokens.

† Corresponding author.

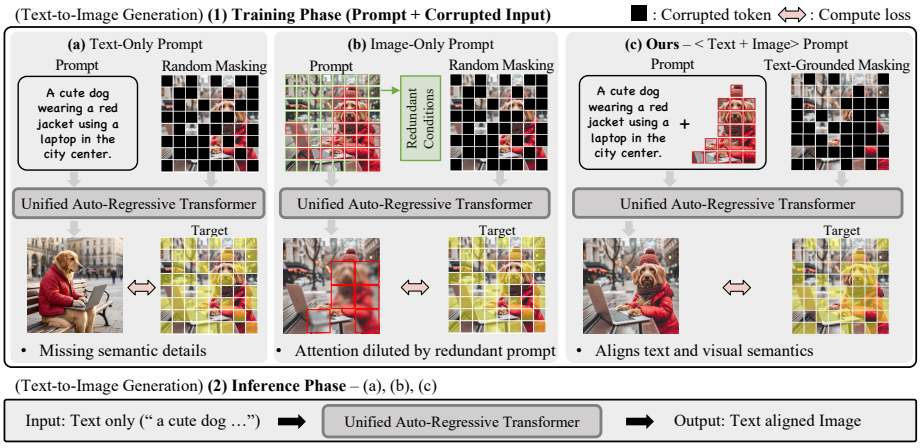

Fig. 1: Generative Training Paradigms in UMMs. (a) Text-conditioned training in UMMs [42, 47] uses the text prompt as the semantic specification and applies random masking to form the corrupted input for token-level reconstruction. (b) Image reconstruction-based approach, Reca [46], uses image prompts to reconstruct corrupted input, yet still relies on random masking. (c) Our approach, **Se**mantically-**Gro**unded **S**upervision (**SeGroS**), constructs its training signal by conditioning on text and explicitly text-aligned visual prompts. Furthermore, we selectively mask the core semantic regions to construct the corrupted input. This provides highly efficient supervision that concentrates model capacity on core structures rather than irrelevant regions.

Early UMMs [7, 14, 38] primarily adopt standard next-token prediction, generating visual tokens autoregressively in a unidirectional sequence. However, this sequential nature is computationally inefficient [4, 5, 26] and struggles to capture complex 2D spatial dependencies required for holistic 2D visual modeling [29, 39, 47]. To overcome these limitations, subsequent approaches shift toward mask prediction or denoising objectives, reconstructing corrupted visual tokens conditioned on both text and visible visual context. For instance, Show-o [47] introduces an **AR-diffusion** architecture that integrates autoregressive (AR) modeling with masked diffusion over visual tokens, enabling parallel decoding and faster generation. Harmon [42] incorporates masked autoregressive (MAR) modeling [29] within an **AR-MAR** framework to reconstruct corrupted visual tokens. Despite architectural differences, these paradigms share a common training strategy: the model processes a *corrupted input* in which part of the visual tokens are masked and the rest remain visible, and computes the reconstruction loss only on the masked subset (Fig. 1 (a)). Reconstructing these tokens conditioned on the *text prompt* provides a cross-modal learning signal.

Despite architectural progress, a fundamental bottleneck still persists across UMM paradigms: the granularity mismatch between textual conditions and visual targets. While visual tokens encode dense spatial structure and fine-grained details, text prompts provide only abstract semantic constraints [17, 32, 46]. As

a result, a single text description can correspond to multiple visually distinct images, rendering text-to-image supervision inherently ambiguous. As illustrated in Fig. 1 (a), the text prompt typically specifies primary objects (*e.g.*, a dog wearing a red jacket) but leaves specific attributes (*e.g.*, texture, illumination, pose, and spatial layout) underspecified. During training, however, the model is required to reconstruct one specific target image. Consequently, even semantically valid visual variations may be penalized if they do not exactly match the **text-underspecified factors** of the target, encouraging the model to fit incidental instance-level details rather than learning robust semantic alignment.

To compensate for missing fine-grained visual details in text-only conditioning, recent UMM training approach incorporates reconstruction objectives based on image-conditioned prompts (*i.e.*, *Visual Hints*) [46], as illustrated in Fig. 1 (b). By leveraging visual tokens from the image, this method provides fine-grained conditioning cues that are absent in the text prompts. However, in image-conditioned training, using all visual tokens as hints can be redundant, as uninformative patches may dilute attention and weaken semantic alignment. Furthermore, regardless of whether conditioning is provided by text [42, 47] or image prompts [46], existing UMMs typically define the corrupted input via random masking that is agnostic to semantic salience. Consequently, a substantial portion of reconstruction supervision is wasted on semantically irrelevant regions, encouraging the model to fit incidental visual details rather than strengthening text-aligned visual concepts.

In this work, we introduce **Se**mantically-**Gro**unded **S**upervision (**SeGroS**), a fine-tuning framework for UMMs that mitigates inefficient supervision arising from the granularity mismatch between coarse text and dense visual tokens. As illustrated in Fig. 1 (c), SeGroS departs from previously used conditioning training by providing semantically grounded visual guidance during training. At the core of SeGroS is a visual grounding score that explicitly quantifies the alignment between discriminative text tokens and image patches, allowing us to construct structured supervision. Based on this score, we derive two complementary components: *Visual Hints*, which provide semantically grounded visual embeddings to guide generation, and semantically grounded *Corrupted Input*, which allocates reconstruction supervision to text-aligned regions (Fig. 2).

The proposed framework consists of three key steps. We first perform **Discriminative Text Token Filtering** to identify linguistically salient tokens that have strong visual counterparts. Since uniformly comparing all tokens dilutes the alignment scores of core regions, we compute both intra-modal (text–text) and inter-modal (text–image) affinities to retain text tokens that dominate the visual semantics. Then, we construct a **Visual Grounding Map** by measuring the similarity between the *filtered* text tokens and each image patch. This map captures fine-grained text–image correspondences and identifies semantically coherent visual regions. Finally, guided by this map, patches with high grounding scores are explicitly extracted to serve as *Visual Hints*. Meanwhile, patches with the lowest scores are selected to form the unmasked context of the semantically-

grounded *Corrupted Input*. This naturally leaves the core semantic areas masked, forcing the model to actively reconstruct them rather than random patches.

Our main contributions are fourfold: **(1)** We propose SeGroS, a semantically grounded fine-tuning framework for UMMs that enhances cross-modal alignment by overcoming the text-image granularity mismatch. **(2)** Within SeGroS, we introduce a fine-grained grounding mechanism that first filters discriminative text tokens and then constructs a visual grounding map from these filtered tokens to extract text-aligned image regions. **(3)** Leveraging this grounding, we construct *Visual Hints* for conditioning and a semantically grounded *corrupted input*, concentrating supervision on core semantic regions. **(4)** We demonstrate that SeGroS improves generation fidelity and cross-modal alignment across diverse UMMs, achieving strong results on GenEval [16], DPGBench [19], and CompBench [20].

2 Preliminary

We formulate unified multimodal generation and understanding under masked/denoising UMM paradigms [42, 47]. To maintain an architecture-agnostic formulation, we abstract away model-specific instantiations (*e.g.*, masked diffusion denoising [47] or masked autoregressive decoding [42]) and instead focus on their shared training principle: reconstructing masked visual tokens from the corrupted input conditioned on the text prompt.

Unified Input Representations. We define all input modalities in a shared D-dimensional latent space. **Text.** Given a text prompt $\mathbf{T}$, we tokenize it into text tokens and map them to embeddings via a pre-trained embedding table, yielding $\mathbf{Z}_T \in \mathbb{R}^{L_T \times D}$, where L_T denotes the text sequence length. **Vision.** Given an image $\mathbf{I}$, a pretrained tokenizer/encoder yields either discrete tokens or continuous latents, which we represent in the shared space as $\mathbf{Z}_I \in \mathbb{R}^{N_I \times D}$, where N_I denotes the number of visual tokens.

Corrupted Input. Following standard practices, at each training step we randomly choose a masking ratio $\gamma(t) \in [0.7, 1.0)$, where $\gamma(t)$ denotes the fraction of tokens treated as unknown (*e.g.*, $\gamma(t) = 0.7$ masks 70% of tokens). We then sample a mask index set $\mathcal{M} \subset \{1, \ldots, N_I\}$ with $|\mathcal{M}| = \lfloor \gamma(t)\, N_I \rfloor$. The masked and unmasked subsets are defined as $\mathbf{Z}_I^{\text{mask}} = \{z_i \mid i \in \mathcal{M}\}$ and $\mathbf{Z}_I^{\text{seen}} = \{z_i \mid i \notin \mathcal{M}\}$. The corrupted input $\widetilde{\mathbf{Z}}_I$ is constructed by replacing tokens in $\mathbf{Z}_I^{\text{mask}}$ with a learnable `[MASK]` embedding, while keeping $\mathbf{Z}_I^{\text{seen}}$ unchanged. This formulation reflects the common mask-and-reconstruct training paradigm in masked/denoising UMMs, with supervision restricted to the masked subset.

Training Objective. The unified model f_θ is optimized for Text-to-Image (T2I) generation by minimizing a generalized reconstruction loss:

$$\mathcal{L}_{\text{t2i}} = \ell\big(f_\theta(\mathbf{Z}_T, \widetilde{\mathbf{Z}}_I),\, \mathbf{Z}_I\big), \tag{1}$$

where the loss $\ell(\cdot)$ is computed only over the masked subset $\mathcal{M}$. The loss $\ell(\cdot, \cdot)$ is instantiated as negative log likelihood for discrete tokens (*e.g.*, Show-o [47]) or Mean Squared Error (MSE) for continuous latents (*e.g.*, Harmon [42]).

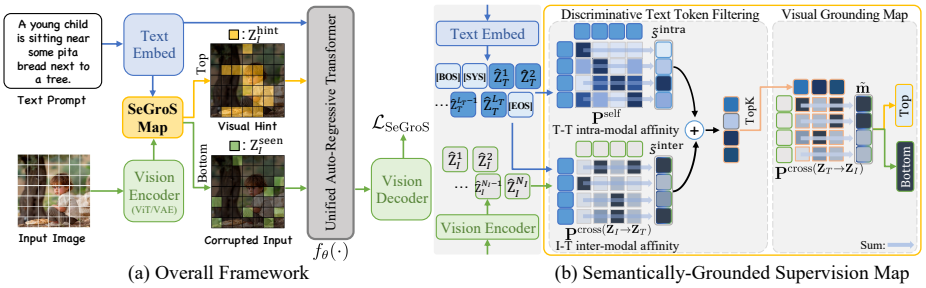

Fig. 2: Overview of SeGroS during training. (a) The SeGroS Map computes a grounding score for each visual token based on its semantic similarity to discriminative text tokens and constructs two complementary training signals. The *Visual Hints* preserve high-groundedness tokens as additional conditioning signals (yellow patches). The *Corrupted Input* retains low-groundedness tokens as unmasked context (green patches) while masking the remaining tokens, which are reconstructed by the unified autoregressive Transformer. (b) Discriminative Text Token Filtering identifies visually relevant text tokens via intra- and inter-modal affinities, which are then used to compute the Visual Grounding Map. At inference time, SeGroS Map generation is omitted, allowing the original UMMs to generate images without architectural modifications.

Furthermore, to enable multimodal understanding and text generation capabilities, we incorporate an autoregressive Image-to-Text (I2T) objective:

$$\mathcal{L}_{\text{i2t}} = \ell\big(f_\theta(\mathbf{Z}_I, \mathbf{Z}_{T,[\texttt{question}]}), \mathbf{Z}_{T,[\texttt{answer}]}\big), \tag{2}$$

where ℓ is the standard negative log-likelihood loss for autoregressively predicting the answer $\mathbf{Z}_{T,[\texttt{answer}]}$ conditioned on the image $\mathbf{Z}_I$ and question $\mathbf{Z}_{T,[\texttt{question}]}$.

3 Proposed Method

3.1 Motivation and Overview

Under text-only conditioning, Eq. (1) suffers from a *granularity mismatch*: sparse text lacks fine-grained constraints [17, 32, 46], resulting in ambiguous reconstruction supervision (Fig. 1 (a)). To address the sparsity issue of text prompts, Reca [46] introduces image-only conditioning, leveraging dense visual tokens as prompts, as described in Fig. 1 (b): $\mathcal{L}_{\text{reca}} = \ell\big(f_\theta(\mathbf{Z}_I, \widetilde{\mathbf{Z}}_I), \mathbf{Z}_I\big)$. The corrupted input $\widetilde{\mathbf{Z}}_I$ is reconstructed conditioned on the visual tokens $\mathbf{Z}_I$ (visual hint). However, conditioning on all visual tokens within the image prompt $\mathbf{Z}_I$ inherently introduces *supervisory redundancy*, which can dilute attention and weaken semantic alignment. Furthermore, semantic-agnostic random masking often assigns reconstruction targets to semantically irrelevant regions, wasting supervision on incidental details rather than strengthening text-aligned structure.

To empirically validate this hypothesis, we conducted an ablation on the image-only prompting setting of Reca [46] to test whether conditioning on all

visual tokens is necessary. We ranked image patches by a text–image affinity score (used only for patch selection) and kept only the top 30% as visual hints. Despite using 3.3× fewer hint tokens, this prompt pruning improves GenEval [16] from 78.7→79.2 and DPGBench [19] from 84.7→85.2 on the Harmon-0.5B model [42]. (More details are provided in Appendix Tab. 3.) These preliminary results explicitly validate our claim: conditioning on the entirety of dense image prompts inherently introduces redundancy, which stems from *text-underspecified* regions, proving that semantically selective conditioning on *text-grounded* regions is essential.

Driven by this finding, we propose **Se**mantically-**Gro**unded **S**upervision (**SeGroS**), a fine-tuning framework that restructures dense visual supervision according to text–image semantic correspondence (Fig. 2). We explicitly identify which image regions are semantically grounded in the text and allocate supervision accordingly. SeGroS operates in three conceptual steps. First, we perform **Discriminative Text Token Filtering** to extract linguistically salient tokens that are likely to have visual counterparts, retaining key entities and interactions that dominate the scene semantics. Second, we construct a **Visual Grounding Map** that measures patch-level similarity between the filtered text tokens and visual features, quantifying how strongly each image patch aligns with the textual semantics. Finally, based on the grounding map, we construct two complementary supervision signals: (1) *Semantic Visual Hints*, consisting of highly aligned regions provided as explicit conditioning signals, and (2) *Semantically grounded Corrupted Input*, which is input for reconstruction. Specifically, low-groundedness tokens are retained as unmasked context, while high-groundedness tokens are masked and reconstructed by the unified autoregressive Transformer. By concentrating the reconstruction objective on semantically coherent regions, SeGroS provides structured supervision that strengthens cross-modal alignment while reducing redundant visual noise.

3.2 Discriminative Text Token Filtering

We first refine the text prompt $\mathbf{Z}_T$ by identifying tokens that are most relevant for visual grounding, as treating all text tokens equally can dilute alignment scores for core visual regions. A text token is considered discriminative only if it is both linguistically salient and visually grounded. To this end, we evaluate each token from two complementary perspectives. We compute the **Text Intra-modal Affinity** to measure its global linguistic importance within the text prompt, and the **Text–Image Inter-modal Affinity** to assess its correspondence with visual tokens. Tokens that score highly on both criteria are retained as discriminative text tokens for subsequent groundedness estimation.

Text Intra-modal Affinity. The first step is to identify the core concepts within the text prompt itself. Given the ℓ_2-normalized text embeddings $\hat{\mathbf{Z}}_T \in \mathbb{R}^{L_T \times D}$ extracted by the tokenizer prior to processing by the unified multimodal model f_θ, we compute the self-affinity matrix $\mathbf{S} \in \mathbb{R}^{L_T \times L_T}$ as $\mathbf{S} = \hat{\mathbf{Z}}_T \hat{\mathbf{Z}}_T^\top$ to capture

contextual relationships between tokens. This effectively captures phrase-level coupling to better preserve actions (*e.g.*, boys-lying). To focus on core semantic concepts, we exclude special tokens (*e.g.*, `[BOS]`, `[EOS]`) from the calculation, which are known to act as attention sinks [24, 44] that bias attention weights away from discriminative content. We then derive the attention probability map $\mathbf{P}^{\text{self}}$ by applying a row-wise softmax with a temperature parameter τ:

$$\mathbf{P}_{kj}^{\text{self}} = \frac{\exp(\mathbf{S}_{kj}/\tau)}{\sum_{m=1}^{L_T} \exp(\mathbf{S}_{km}/\tau)}. \tag{3}$$

Finally, the intra-modal affinity score s_j^{intra} for the j-th text token is obtained by aggregating the attention it receives from all other tokens in the sequence:

$$s_j^{\text{intra}} = \sum_{k=1}^{L_T} \mathbf{P}_{kj}^{\text{self}}. \tag{4}$$

Image-Text Inter-modal Affinity. While intra-modal affinity identifies semantically dominant tokens, it does not guarantee their visual congruence with the actual image content. To address this, we evaluate the visual relevance of each text token by measuring its direct influence on the image features. We compute the image-text inter-modal affinity matrix $\mathbf{A} \in \mathbb{R}^{N_I \times L_T}$ as $\mathbf{A} = \hat{\mathbf{Z}}_I \hat{\mathbf{Z}}_T^\top$, using the ℓ_2-normalized visual token embeddings $\hat{\mathbf{Z}}_I$ and text token embeddings $\hat{\mathbf{Z}}_T$, where both are derived from their respective tokenizers prior to processing by the unified multimodal model f_θ. Similar to the text intra-modal affinity, we calculate the inter-modal attention probability $\mathbf{P}^{\text{cross}(\mathbf{Z}_I \to \mathbf{Z}_T)}$ that represents the likelihood of visual token i attending to text token j:

$$\mathbf{P}_{ij}^{\text{cross}(\mathbf{Z}_I \to \mathbf{Z}_T)} = \frac{\exp(\mathbf{A}_{ij}/\tau)}{\sum_{k=1}^{L_T} \exp(\mathbf{A}_{ik}/\tau)}. \tag{5}$$

The inter-modal affinity score s_j^{inter} is defined by aggregating the attention the j-th text token receives from the entire visual context. This metric filters out text concepts that have little visual grounding in the image:

$$s_j^{\text{inter}(\mathbf{Z}_I \to \mathbf{Z}_T)} = \sum_{i=1}^{N_I} \mathbf{P}_{ij}^{\text{cross}}. \tag{6}$$

After computing the intra-modal centrality (s^{intra}) and inter-modal relevance ($s^{\text{inter}(\mathbf{Z}_I \to \mathbf{Z}_T)}$), we integrate these metrics to determine the final semantic importance of each text token.

Discriminative Token Selection. Finally, to identify the most discriminative tokens, we integrate the linguistic and visual insights derived above. Since the two affinity scores operate on different scales, we normalize them via Min-Max scaling to ensure a balanced contribution. For each score vector $\mathbf{s} \in \{\mathbf{s}^{\text{intra}}, \mathbf{s}^{\text{inter}(\mathbf{Z}_I \to \mathbf{Z}_T)}\}$,

we compute $\tilde{s}_j = \frac{s_j - \min(\mathbf{s})}{\max(\mathbf{s}) - \min(\mathbf{s})}$. The unified importance score Ω_j is then defined as the sum of the normalized metrics:

$$\Omega_j = \tilde{s}_j^{\text{intra}} + \tilde{s}_j^{\text{inter}}. \tag{7}$$

We select the indices $\mathcal{K}$ of the top K_T tokens, where $K_T = \lfloor \rho L_T \rfloor$ is determined by the sequence length L_T and a text preservation ratio ρ:

$$\mathcal{K} = \text{TopK}(\Omega, K_T) = \{j \mid \text{rank}(\Omega_j) \leq K_T\}. \tag{8}$$

Accordingly, we define the binary text-filtering mask $\mathbf{w} \in \{0,1\}^{L_T}$:

$$w_j = \mathbb{1}(j \in \mathcal{K}) = \begin{cases} 1 & \text{if } j \in \mathcal{K} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

3.3 Visual Grounding Map

With the binary text-filtering mask $\mathbf{w}$ identified, we evaluate how strongly each visual token aligns with the filtered text tokens, yielding the Visual Grounding Map. This map serves as the decisive criterion for constructing two complementary training signals: *Visual Hints* and *Corrupted Input*. Similar to the image-to-text affinity, we calculate the text-to-image attention probability $\mathbf{P}^{\text{cross}(\mathbf{Z}_T \to \mathbf{Z}_I)}$ that represents the likelihood of text token j attending to visual token i:

$$\mathbf{P}_{ji}^{\text{cross}(\mathbf{Z}_T \to \mathbf{Z}_I)} = \frac{\exp(\mathbf{A}_{ji}^{\top}/\tau)}{\sum_{k=1}^{N_I} \exp(\mathbf{A}_{jk}^{\top}/\tau)}. \tag{10}$$

Using this probability, we compute the visual grounding score m_i for each image patch i by aggregating the text-to-image attention probabilities weighted by the text filtering mask $\mathbf{w}$:

$$m_i = \sum_{j=1}^{L_T} w_j \cdot \mathbf{P}_{ji}^{\text{cross}(\mathbf{Z}_T \to \mathbf{Z}_I)}, \quad \text{for } i \in \{1, \ldots, N_I\}. \tag{11}$$

These scores are collected into a vector $\mathbf{m} = [m_1, \ldots, m_{N_I}] \in \mathbb{R}^{N_I}$, which is then normalized to the $[0,1]$ range via Min–Max scaling, defined as $\bar{m}_i = \frac{m_i - \min(\mathbf{m})}{\max(\mathbf{m}) - \min(\mathbf{m})}$. Directly applying deterministic selection to this score map enforces a fixed construction of *Visual Hints* and *Corrupted Input*, repeatedly isolating the exact same regions across epochs. While text strictly requires such determinism to preserve core semantics, images are fundamentally spatially redundant, often containing multiple patches with similar grounding scores. Thus, exclusive reliance on deterministic masking leads to a performance decline [10]. To mitigate this, we inject a slight uniform noise $\boldsymbol{\xi} \sim \mathcal{U}([0, 0.5]^{N_I})$ into the normalized visual scores $\bar{\mathbf{m}}$ before the Top/Bottom selection, formulated as:

$$\tilde{\mathbf{m}} = \bar{\mathbf{m}} + \boldsymbol{\xi}. \tag{12}$$

3.4 Visual Hints and Corrupted Input Construction

Based on the perturbed visual grounding map $\tilde{\mathbf{m}}$, we map the visual tokens to their respective roles using two distinct criteria, as shown in Fig. 2. First, to resolve the linguistic ambiguity of the sparse text prompt, we construct *Visual Hints* ($\mathbf{Z}_I^{\text{hint}}$) to serve as an explicit visual prompt. Specifically, given a predefined selection ratio η, we define these hints by selecting the top-$\lfloor \eta N_I \rfloor$ tokens with the highest scores, representing the visual parts most highly aligned with the text. Our empirical analysis (Tab. 3) identifies the range $\eta \in [0.3, 0.5]$ as the optimal balance for providing effective semantic cues; thus, we adopt $\eta \in$[0.3, 0.4].

Second, we define a semantically grounded *Corrupted Input* ($\widetilde{\mathbf{Z}}_I^{\text{SeGroS}}$) to focus reconstruction on text-aligned regions. Instead of randomly sampling a mask set [42, 46, 47], we deliberately retain semantically low-salience regions as visible context and mask the core semantic regions as reconstruction targets. Given a masking ratio $\gamma(t) \in [0.7, 1.0)$, we set $K_{\text{mask}} = \lfloor \gamma(t) N_I \rfloor$ and $K_{\text{seen}} = N_I - K_{\text{mask}}$. We select the bottom-$K_{\text{seen}}$ indices from the grounding map $\tilde{\mathbf{m}}$ to remain visible: $\mathcal{S} = \text{BottomK}(\tilde{\mathbf{m}}, K_{\text{seen}})$, and define $\mathbf{Z}_I^{\text{seen}} = \{z_i \mid i \in \mathcal{S}\}$. The mask index set is then given by $\mathcal{M} = \{1, \ldots, N_I\} \setminus \mathcal{S}$, and the corrupted input is formed by replacing $\{z_i \mid i \in \mathcal{M}\}$ with a learnable `[MASK]` embedding.

Final Objective. We optimize a joint objective where *Visual Hints* provide additional conditioning alongside the text prompt, and the semantically grounded *Corrupted Input* specifies the masked reconstruction targets. Conditioned on $\mathbf{Z}_T$ and $\mathbf{Z}_I^{\text{hint}}$, the model f_θ processes the partially masked sequence $\widetilde{\mathbf{Z}}_I^{\text{SeGroS}}$ and reconstructs the original visual tokens at the masked positions. To preserve the model's image-to-text (I2T) capability, we additionally include an autoregressive I2T objective. The total loss is:

$$\mathcal{L}_{\text{total}} = \underbrace{\ell\big(f_\theta(\mathbf{Z}_T, \mathbf{Z}_I^{\text{hint}}, \widetilde{\mathbf{Z}}_I^{\text{SeGroS}}), \mathbf{Z}_I\big)}_{\mathcal{L}_{\text{SeGroS}}} + \lambda \mathcal{L}_{\text{i2t}}, \tag{13}$$

where $\ell(\cdot, \cdot)$ is evaluated only on the masked subset of $\widetilde{\mathbf{Z}}_I^{\text{SeGroS}}$. λ is a fixed hyperparameter that balances the two loss terms.

4 Experiments

4.1 Settings

Unified Multimodal Models (UMMs). To demonstrate the robustness of SeGroS, we conducted extensive experiments across three families of UMMs, encompassing diverse model scales (from 0.5B to 3.6B parameters) and generation resolutions (256×256 and 512×512). All models were initialized with their official pre-trained checkpoints before fine-tuning. The details are as follows:
(I) Show-o [47] unifies a masking-based diffusion model within a single autoregressive framework (**AR+Diffusion**). We adopted both Show-o-256 and

Table 1: Comparison on text-to-image generation across UMMs. We compared the proposed method with the Supervised Fine-Tuning (SFT) and Reca [46]. '*' indicates results reproduced using the official implementations.

UMMs	Method	GenEval ↑							DPGBench ↑	CompBench ↑
		Single	Two	Count	Color	Position	Attr.	Overall	Overall	Overall
Fine-Tuning Dataset: MidjourneyV6 Dataset										
Show-o-256 [47]	w/o SFT	97.4	63.3	52.1	82.3	14.2	30.3	56.6	70.65	73.58
	SFT*	97.8	63.4	53.4	80.3	16.5	33.5	57.5	74.6	75.37
	Reca	97.4	**73.6**	**56.0**	83.8	20.3	40.2	61.9	75.7	77.5*
	Ours	**98.1**	72.5	54.4	**86.4**	**21.0**	**41.0**	**62.22**	**76.52**	**78.30**
Show-o-512 [47]	w/o SFT	97.2	80.3	61.9	78.2	27.3	52.3	66.2	82.21	80.53
	SFT*	97.2	84.1	**66.6**	80.9	26.8	46.8	67.03	81.8	80.02
	Reca	**98.1**	**93.4**	64.7	79.8	**38.0**	**55.8**	**71.63***	84.94	**84.47**
	Ours	97.8	91.4	65.0	**81.9**	35.0	53.0	70.69	**85.18**	84.35
OpenUni-1.6B [41]	w/o SFT	96.8	63.3	46.4	80.1	18.5	30.8	56.0	76.29	75.47
	SFT*	97.8	76.8	55.9	81.9	25.5	36.0	62.32	79.7	81.25
	Reca	96.6	**85.4**	52.5	**84.3**	**46.5**	50.8	69.33*	80.45	83.1*
	Ours	**98.8**	84.6	**56.9**	84.0	38.8	**54.0**	**69.5**	**81.33**	**83.83**
OpenUni-3.6B [41]	w/o SFT	99.1	71.8	51.9	83.9	23.3	41.6	61.9	79.02	78.84
	SFT*	98.8	81.3	**55.9**	85.9	25.0	48.8	65.94	80.45	82.6
	Reca	99.1	92.7	52.3	**87.1**	43.8	70.3	74.1	82.75	86.0*
	Ours	**99.7**	**95.2**	52.5	85.1	**46.8**	**73.0**	**75.37**	**83.37**	**86.14**
Harmon-0.5B [42]	w/o SFT	99.7	80.5	55.8	86.7	32.2	49.7	67.6	80.12	80.34
	SFT*	100	86.4	64.4	87.5	37.8	56.5	72.1	82.5	83.5
	Reca	99.9	**92.3**	59.4	**91.7**	58.5	70.7	78.7	84.67	85.7*
	Ours	**100**	90.2	**65.3**	91.5	**66.0**	**75.8**	**81.5**	**85.4**	**86.9**
Fine-Tuning Dataset: BLIP3o-60k Dataset										
Harmon-1.5B [42]	w/o SFT	99.4	87.3	68.7	86.4	44.9	51.1	72.9	80.93	81.36
	SFT*	99.1	94.7	77.8	86.4	78.3	74.0	85.0	85.6	87.32
	Reca	–	–	–	–	–	–	85.2	86.5	87.2*
	Ours	**100**	**97.7**	**79.1**	**90.4**	**83.5**	**81.3**	**88.66**	**86.58**	**88.08**

Show-o-512 variants to evaluate performance across different image resolutions. **(II) Harmon** [42] integrates a Masked Autoregressive (MAR)-based generative model into a unified autoregressive architecture (**AR+MAR**). We utilized Harmon-0.5B and Harmon-1.5B to assess the scalability of our method across varying model capacities. **(III) OpenUni** [41] integrates an autoregressive understanding model with a diffusion-based generation model (**AR+Diffusion**). To validate our approach, we adopt the base OpenUni models (1.6B and 3.6B) operating at a 512×512 resolution, excluding the GPT-4o-Image distillation.

Datasets. For fine-tuning, we combined high-quality generation data with an auxiliary instruction dataset, aligning with the experimental setup of Reca [46] for a fair comparison. We used MidjourneyV6 [11] and BLIP3o-60k [6] to improve aesthetic quality and semantic alignment, while incorporating LLaVA-Instruct-150K [33] to preserve the model's understanding capability.

Hyperparameter. We set the softmax temperature $\tau = 1.0$ for all softmax operations, text preservation ratio $\rho = 0.4$ in (8), and joint objective weight $\lambda = 1.0$ in (13). The visual hint selection ratio is $\eta = 0.3$ (or 0.4 for Show-o

Table 2: Visual understanding performance.

Method	MME	$\text{POPE}_{(\text{Acc})}$	$\text{POPE}_{(\text{F1})}$	GQA	MMMU	SEED
w/o SFT	1195	83.8	83.9	**58.8**	34.7	65.2
SFT	1210	84.0	84.1	58.5	35.3	65.1
Reca [46]	1190	83.9	83.4	58.5	34.9	65.1
Ours	**1217**	**84.5**	**84.2**	58.7	**36.0**	**65.5**

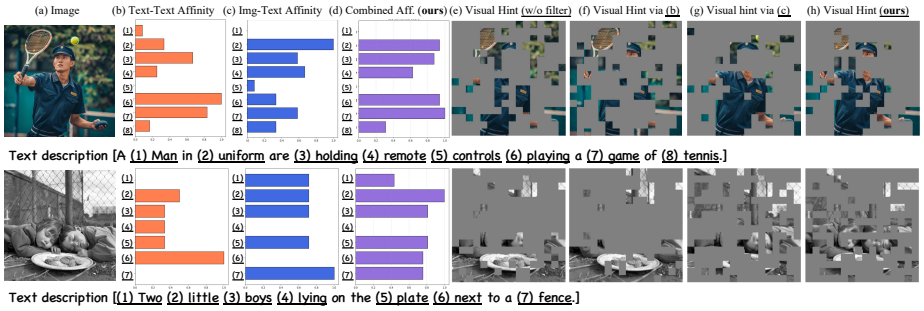

Fig. 3: Visualization of text token affinity and derived visual hints using the MidjourneyV6 [11] text-image dataset. (a) The original input image with text description. (e–h) The extracted visual hints are represented by the image prompt. For (f–h), we project the text scores from (b–d) to compute the visual grounding map, retaining only the top 30% of visual tokens with the highest groundedness scores. (e) shows the image prompt generated using all text words without discriminative filtering. Text scores exclude non-semantic functional stopwords.

512×512). Following prior UMMs [42, 47], we adopt their masking schedules for sampling $\gamma(t)$.

4.2 Effectiveness of SeGroS in Text-to-Image Generation

As shown in Tab. 1, SeGroS consistently improves upon standard supervised fine-tuning (SFT) across all evaluated UMM backbones and datasets, yielding higher overall scores on GenEval [16], DPGBench [19], and CompBench [20] in most cases. Compared to Reca, SeGroS achieves higher DPGBench scores universally and superior GenEval overall scores in five of six settings (excepting a marginal drop on Show-o-512, where our DPGBench still leads). Notably, the gains are most pronounced in composition-sensitive GenEval categories (e.g., *Position* and *Attr.*), suggesting that our grounding-aware supervision effectively allocates learning capacity toward prompt-relevant structures. For instance, on OpenUni-3.6B, SeGroS improves the GenEval overall score from 65.94% (SFT) and 74.1% (Reca) to 75.37%, while substantially boosting *Position* accuracy (from 25.0% to 46.8%). We observe a similar trend when fine-tuning on BLIP3o-60k, where SeGroS pushes Harmon-1.5B to an 88.66% GenEval overall score, indicating robustness across both backbone designs and data distributions. These quantitative trends are further supported qualitatively in Fig. 4, where SeGroS better satisfies compositional constraints such as object counting and spatial relations. Fig. 5 further showcases its high-fidelity generation.

4.3 Image-to-Text Understanding Results

We evaluate image-to-text (I2T) understanding on the Harmon-1.5B [42] fine-tuned with the BLIP3o-60k dataset (Tab. 2). Building upon Harmon's observation

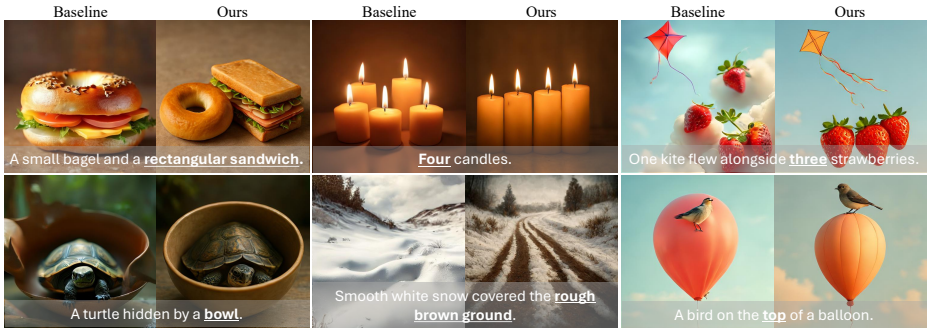

Fig. 4: Qualitative comparison of T2I results using Harmon-1.5B [42] fine-tuned on the BLIP3o-60k [6] dataset. Baseline indicates the model before SFT. More visualizations for other backbones are provided in the supplementary material.

Table 3: Effect of Visual Hint Ratio.

Ratio (η)	GenEval	DPGBench
0%	72.1	82.5
10%	80.4	84.5
30%	**81.5**	**85.4**
50%	81.2	85.1
80%	79.8	84.9
100%	79.2	84.5

Table 4: Ablation on text filtering affinity.

Metric		Performance	
$\tilde{s}^{\text{intra}}$	$\tilde{s}^{\text{inter}}$	GenEval	DPGBench
–	–	79.2	84.0
✓	–	79.8	84.8
–	✓	81.5	84.8
✓	✓	**81.5**	**85.4**

Table 5: Ablation on visual hint and corrupted input.

Visual Hint		$\mathbf{Z}_I^{\text{seen}}$		Performance	
Top	Bot.	Top	Bot.	GenEval	DPGBench
–	✓	✓	–	79.5	84.6
–	✓	–	✓	80.4	84.5
✓	–	✓	–	79.9	84.8
✓	–	–	✓	**81.5**	**85.4**

that improved generative modeling benefits visual understanding, we find that SeGroS effectively leverages this synergy. Relative to w/o SFT, SeGroS yields a +22 gain on MME [13] and a +1.3 gain on MMMU [50], with consistent improvements on POPE [30] and SEED [27]. GQA [21] performance remains stable (58.7 vs. 58.8), verifying the preservation of visual understanding capabilities. Moreover, SeGroS consistently outperforms both standard SFT and the reconstruction-based Reca [46] variant across these benchmarks, indicating that grounding-aware supervision strengthens semantic representations.

4.4 Ablation study

In this section, we conduct ablation studies to verify the effectiveness of each component in SeGroS. All ablation models are trained on the MidjourneyV6 dataset [11] using the Harmon-0.5B [42] backbone. We maintain the same hyperparameter settings as described in Tab. 1.

Redundancy Analysis on Visual Hints. We hypothesize that utilizing the entire image as a visual hint introduces spatial redundancy, diluting essential supervisory signals. To validate this, we ablate the visual hint ratio η in Tab. 3, where η represents the percentage of top-scoring patches selected via (12). The results demonstrate that simply providing more visual information is not always

Table 6: Ablation on supervision allocation: We form the corrupted input with a masking ratio $\gamma(t) \sim \mathcal{U}[0.7, 1.0)$, *i.e.* the ratio of visible tokens is $1 - \gamma(t)$. While SFT employs random masking, our method adopts adaptive masking guided by the visual grounding map. In 'Ours (drop-loss)', supervision for the reconstruction loss during training is restricted to the Top-30% grounded targets.

Method	Visual Hints	Masking ratio $\gamma(t)$	Loss targets	GenEval ↑	DPGBench ↑
SFT	0%	$\mathcal{U}[0.7, 1.0)$	All Masked Regions	72.1	82.5
Ours	30%	$\mathcal{U}[0.7, 1.0)$	All Masked Regions	81.5	**85.4**
Ours (drop-loss)	30%	$\mathcal{U}[0.7, 1.0)$	Top-30% Masked Regions	**82.1**	84.7

beneficial; as η increases from our optimal setting of 30% to 100% (full image), the performance steadily drops from 81.5 to 79.2 on GenEval. This confirms that uninformative patches introduce redundancy. Meanwhile, providing no hint ($\eta = 0\%$) or an overly sparse hint ($\eta = 10\%$) fails to provide sufficient structural guidance. Thus, retaining 30% to 50% of the highly-grounded patches strikes the optimal balance, preserving core semantics without capacity waste.

Effectiveness of Text Filtering Components. We evaluate the individual and synergistic effects of intra-modal ($\tilde{s}^{\text{intra}}$) and inter-modal ($\tilde{s}^{\text{inter}}$) affinities (Tab. 4). The first row denotes the no-filtering image–text affinity baseline. Employing only the intra-modal score ($\tilde{s}^{\text{intra}}$) provides a structural linguistic prior that captures core semantic entities, steadily improving alignment (79.8% / 84.8%). Conversely, utilizing only the inter-modal score ($\tilde{s}^{\text{inter}}$) grounds the text tokens to visual patches, significantly boosting GenEval to 81.5%. Finally, the best overall performance is achieved when both are combined (81.5% / 85.4%). Qualitatively, $\tilde{s}^{\text{intra}}$ tends to emphasize tightly coupled phrase-level tokens (e.g., *playing–game* in Fig. 3 (b)), whereas $\tilde{s}^{\text{inter}}$ aggregates attention mass over patches and thus favors visually dominant entities (e.g., *uniform* in Fig. 3 (c)); their combination yields the cleanest visual hints in Fig. 3 (d,h).

Ablation on Hint and Corrupted Input Construction. Tab. 5 ablates the assignment of high- (Top) and low-groundedness (Bot) patches to *Visual Hints* and the unmasked context ($\mathbf{Z}_I^{\text{seen}}$). Assigning Top patches to $\mathbf{Z}_I^{\text{seen}}$ misdirects the reconstruction loss toward irrelevant Bot regions, degrading performance (Rows 1, 3). In contrast, retaining Bot patches as $\mathbf{Z}_I^{\text{seen}}$ correctly concentrates supervision on core semantics (Rows 2, 4). Among these, providing Top rather than Bot patches as Visual Hints further enhances semantic conditioning (Row 4 vs. Row 2). Overall, our default configuration (Row 4)—Top for Hints and Bot for $\mathbf{Z}_I^{\text{seen}}$—is optimal, successfully coupling informative visual guidance with semantically focused reconstruction.

Validating the Supervisory Allocation. We hypothesize that when constructing the corrupted input, the reconstruction loss should be strictly concentrated

Fig. 5: Qualitative T2I generation results of SeGroS with OpenUni-3.6B [41]. Refer to the supplementary material for the corresponding text prompts.

on semantically relevant regions, rather than being diluted across redundant areas. To validate this, we fix the masking schedule $\gamma(t) \sim \mathcal{U}(0.7, 1.0)$ for all settings and vary only *where* the reconstruction loss is applied (Tab. 6). The SFT baseline randomly samples masked positions and supervises all masked tokens with the reconstruction loss. In contrast, SeGroS uses the grounding map to select semantically aligned masked targets, yielding substantial gains in both GenEval and DPGBench. We further test a drop-loss variant that restricts supervision to only the Top-30% grounded targets (assigning zero weight to the rest). Despite backpropagating through only a small subset of masked positions, it remains strong—improving GenEval and still substantially outperforming SFT—while only mildly reducing DPGBench. Overall, this suggests that grounding-aware allocation can retain most of the gains with far fewer supervised targets.

5 Related work

Unified Multimodal Models (UMMs). UMMs [31, 34, 43] unify multimodal understanding and generation within a shared modeling framework. Early token-based approaches [7, 38] encode images as vector quantized tokens, enabling unified mixed-modal sequence modeling. To improve efficiency and representation quality, Show-o [47] integrates discrete denoising diffusion with autoregressive modeling, while Harmon [42] adopts a MAR-based unified encoder for enhanced semantic consistency. OpenUni [41] bridges an autoregressive MLLM and a diffusion generator via a lightweight connector rather than fully sharing a single backbone. Some UMMs also leverage pretrained vision encoders (e.g., CLIP) as semantic priors for improved alignment [32, 37, 43].

Fine-tuning Strategies for UMMs. Fine-tuning plays a crucial role in improving UMM generation by refining text-to-visual alignment. Reca [46] strengthens alignment through image-conditioned reconstruction, where the model reconstructs masked visual tokens from dense visual prompts. Reward-driven alignment methods have also been explored, including self-rewarding frameworks and reinforcement-learning-based optimization [22, 35]. BLIP3-o [6] further studies unified training recipes that incorporate supervised instruction tuning to improve alignment and aesthetics. In contrast, SeGroS improves fine-tuning efficiency by grounding supervision in text–image correspondence and redesigning both conditioning cues and reconstruction targets to reduce prompt-irrelevant updates.

Adaptive Masking Strategies. Masked image modeling [1, 8–10, 18, 36, 48, 49] has become a prominent self-supervised paradigm for learning visual representations via masked reconstruction. However, random masking often allocates capacity to low-information background regions rather than object-level semantics [10, 23, 28]. Prior work mitigates this limitation by masking salient or semantically coherent regions [23, 28] or by adaptively adjusting masking ratios based on token salience [10]. OneRef [45] further explores text-guided masking in referring-centric pretraining settings to strengthen vision–language grounding. While these methods improve representation learning through adaptive masking, they do not explicitly address how reconstruction supervision should be allocated in multimodal generation.

6 Conclusion

In this paper, we introduced SeGroS, a semantically-grounded fine-tuning framework that addresses the granularity mismatch and supervisory redundancy in UMMs. By leveraging a visual grounding map, SeGroS strategically formulates visual hints and corrupted input, concentrating the Text-to-Image loss exclusively on core text-aligned regions. As a result, SeGroS significantly improves cross-modal alignment across major benchmarks.

Acknowledgment

This work was supported by the NRF of Korea grant (RS-2025-24803204), IITP under the Leading Generative AI Human Resources Development (IITP-2026-RS-2026-25546026) grant funded by the Korea government (MSIT), and the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (RQT-25-090192).

References

1. Ahn, J., Choi, H., Kim, S., Min, D.: Madis-stereo: Enhanced stereo matching via distilled masked image modeling. IEEE Access **13**, 8912–8923 (2025)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. CoRR **abs/2308.12966** (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K.P., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. In: International Conference on Machine Learning. pp. 4055–4075. PMLR (2023)

5. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
7. 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 preprint arXiv:2501.17811 (2025)
8. Choi, H., Kim, D., Cha, S., Yi, K.M., Min, D.: Improving generative pre-training: An in-depth study of masked image modeling and denoising models. arXiv preprint arXiv:2412.19104 (2024)
9. Choi, H., Lee, H., Joung, S., Park, H., Kim, J., Min, D.: Emerging property of masked token for effective pre-training. In: European Conference on Computer Vision. pp. 272–289. Springer (2024)
10. Choi, H., Park, H., Yi, K.M., Cha, S., Min, D.: Salience-based adaptive masking: revisiting token dynamics for enhanced pre-training. In: European Conference on Computer Vision. pp. 343–359. Springer (2024)
11. CortexLM: Cortexlm/midjourney-v6 (2024), `https://huggingface.co/datasets/CortexLM/midjourney-v6`
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
13. 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. Advances in Neural Information Processing Systems **38** (2026)
14. Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. In: International Conference on Learning Representations. pp. 35206–35231 (2024)
15. 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 preprint arXiv:2404.14396 (2024)
16. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
17. Hahn, M., Zeng, W., Kannen, N., Galt, R., Badola, K., Kim, B., Wang, Z.: Proactive agents for multi-turn text-to-image generation under uncertainty. In: International Conference on Machine Learning. pp. 21591–21628. PMLR (2025)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
19. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
20. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
21. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)

22. Jin, W., Niu, Y., Liao, J., Duan, C., Li, A., Gao, S., Liu, X.: Srum: Fine-grained self-rewarding for unified multimodal models. arXiv preprint arXiv:2510.12784 (2025)
23. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. In: European Conference on Computer Vision. pp. 300–318. Springer (2022)
24. Kim, J., Esmaeili, E., Qiu, Q.: Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8031–8040 (2025)
25. Koh, J.Y., Fried, D., Salakhutdinov, R.R.: Generating images with multimodal language models. Advances in Neural Information Processing Systems **36**, 21487–21506 (2023)
26. Lezama, J., Chang, H., Jiang, L., Essa, I.: Improved masked image generation with token-critic. In: European Conference on Computer Vision. pp. 70–86. Springer (2022)
27. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
28. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. Advances in Neural Information Processing Systems **35**, 14290–14302 (2022)
29. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
30. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
31. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025)
32. Lin, H., Cho, J., Zadeh, A., Li, C., Bansal, M.: Bifrost-1: Bridging multimodal llms and diffusion models with patch-level clip latents. In: Conference on Neural Information Processing Systems (2025)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
34. Liu, Z., Ren, W., Liu, H., Zhou, Z., Chen, S., Qiu, H., Huang, X., An, Z., Yang, F., Patel, A., et al.: Tuna: Taming unified visual representations for native unified multimodal models. arXiv preprint arXiv:2512.02014 (2025)
35. Mao, W., Yang, Z., Shou, M.Z.: Unirl: Self-improving unified multimodal models via supervised and reinforcement learning. arXiv preprint arXiv:2505.23380 (2025)
36. Son, S., Choi, H., Min, D.: Sg-mim: Structured knowledge guided efficient pre-training for dense prediction. arXiv preprint arXiv:2409.02513 (2024)
37. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: The Twelfth International Conference on Learning Representations
38. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
39. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)

40. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. In: Forty-first International Conference on Machine Learning (2024)
41. 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 preprint arXiv:2505.23661 (2025)
42. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17739–17750 (2025)
43. 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. In: International Conference on Learning Representations. vol. 2025, pp. 93620–93638 (2025)
44. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024)
45. Xiao, L., Yang, X., Peng, F., Wang, Y., Xu, C.: Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. Advances in Neural Information Processing Systems **37**, 139854–139885 (2024)
46. Xie, J., Darrell, T., Zettlemoyer, L., Wang, X.: Reconstruction alignment improves unified multimodal models. In: The Fourteenth International Conference on Learning Representations. (2026)
47. 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. In: The Thirteenth International Conference on Learning Representations (2025)
48. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14475–14485 (2023)
49. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
50. 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 understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)