

HiSVR: Hierarchical Semantic Visual Reconstruction for Unified Multimodal Models

Tailai Chen^{1,2} Weining Wang¹ Yuchen Liu^{1,2} Jun Fu¹

Jie Jiang¹ Yiming Cheng³ Jing Liu^{1,2,*}

¹Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

³Tsinghua University

chentailai26@mails.ucas.ac.cn; liuyuchen2025@ia.ac.cn

{weining.wang, jun.fu, jie.jiang}@nlpr.ia.ac.cn

*Corresponding author: jliu@nlpr.ia.ac.cn

Abstract

Unified multimodal models use a shared Transformer backbone for visual understanding and image generation. Existing post-training signals primarily constrain final outputs, output-side representations, or task-level behavior, providing limited supervision for how visual semantics are organized across intermediate layers. We propose **HiSVR**, a hierarchical semantic visual reconstruction method for post-training unified multimodal models. A frozen semantic visual tokenizer extracts residual semantic codes from each training image, which supervise image-associated hidden states at different Transformer depths. Shallow layers predict shorter semantic-code prefixes, while deeper layers reconstruct longer prefixes, yielding coarse-to-fine supervision over intermediate representations. To balance the differing optimization difficulty across depths and prefix lengths, we introduce an uncertainty-weighted progressive reconstruction loss. HiSVR is used only during training and removed at inference, preserving the original architecture and inference cost. Experiments on unified multimodal backbones show that HiSVR improves visual understanding while maintaining image generation quality. These results indicate that hierarchical semantic reconstruction effectively regularizes intermediate representations in unified multimodal models.

1 Introduction

Unified multimodal models (UMMs) have recently emerged as a promising paradigm for integrating visual understanding and image generation within a single architecture (Zhang et al., 2025; Xie et al., 2025b; Team, 2024; Wang et al., 2024b; Chen et al., 2025b). Instead of relying on separate systems for perception and synthesis, UMMs represent text, images, and task instructions in a unified token space and process them with a shared Transformer backbone. This

formulation provides a unified interface for diverse multimodal tasks, including visual question answering, image captioning, and text-to-image generation, while enabling understanding-oriented and generation-oriented learning signals to interact within the same model.

Recent post-training methods have begun to explicitly leverage the interaction between understanding and generation. Some methods align representations between understanding and generation pathways (Xie et al., 2025a), while others introduce generation-assisted supervision through multi-representation generation (Su et al., 2026) or self-adversarial training (Su et al., 2025). Recent work also explores preference optimization to improve generated images (Jiang et al., 2025). These methods enhance UMMs from different perspectives, but their supervision is typically imposed on final outputs, output-side representations, or task-level behaviors. As a result, the intermediate layers of the shared backbone, where visual information is progressively transformed before final prediction, remain weakly constrained by direct semantic supervision.

This lack of intermediate semantic supervision is particularly important for UMM post-training. Visual understanding requires the model to capture objects, attributes, relations, and scene-level semantics, while image generation requires the model to preserve spatial layouts, local details, and visual consistency. When the same backbone serves both understanding and generation, visual semantics should not only be available at the final prediction stage, but also be organized and preserved across intermediate representations. Otherwise, the model may rely mainly on late-layer states or task-specific output heads, without developing stable visual semantic representations throughout the backbone.

To address this issue, we propose **HiSVR** (**H**ierarchical **S**emantic **V**isual **R**econstruction), a

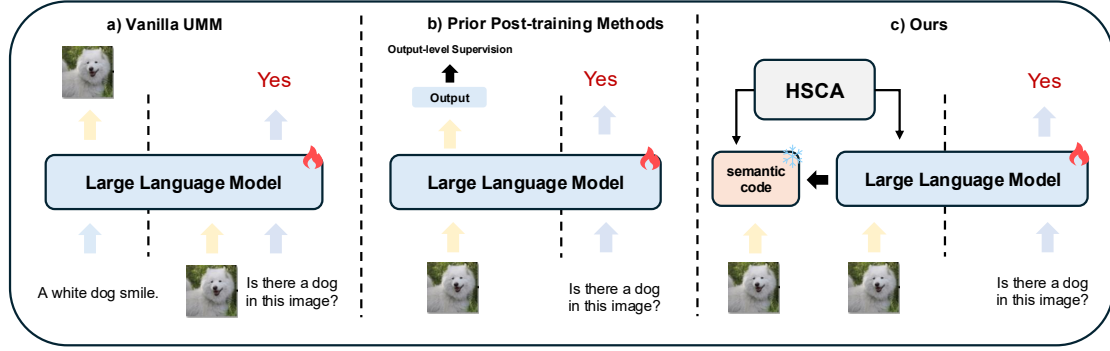


Figure 1: Motivation of HiSVR compared with existing UMM post-training strategies. Prior methods mainly couple visual understanding and image generation through output-level objectives or final representations, providing limited direct supervision to intermediate visual states in the shared backbone. HiSVR instead attaches training-time semantic reconstruction objectives to image-associated hidden states at multiple depths, encouraging hierarchical visual semantics to be organized inside the UMM.

hierarchical semantic reconstruction method for post-training UMMs. HiSVR uses a frozen semantic visual tokenizer to extract residual semantic codes from each training image and employs these codes as auxiliary supervision for intermediate hidden states in the shared backbone. Specifically, we select image-associated hidden states from multiple Transformer depths and attach lightweight reconstruction heads to predict semantic code prefixes of different lengths. Shallow layers are encouraged to predict shorter prefixes, while deeper hidden states predict longer prefixes, forming a coarse-to-fine supervision scheme over intermediate semantic visual representations.

The key idea is that residual semantic code prefixes provide supervision targets with gradually increasing complexity. Shorter prefixes offer relatively coarse semantic information, while longer prefixes contain richer and more detailed visual semantics. Compared with applying semantic supervision only near the final prediction stage, HiSVR directly regularizes image-associated representations at multiple backbone depths, encouraging visual semantics to be preserved more consistently throughout the model. To balance the varying optimization difficulty introduced by different depths and prefix lengths, we further introduce an **uncertainty-weighted progressive reconstruction (UPR)** loss (Kendall et al., 2018), which adaptively weights the hierarchical reconstruction objectives during post-training.

HiSVR is used only as an auxiliary training mechanism and does not alter the native understanding or generation pathway of the underlying UMM. During post-training, the semantic tok-

enizer and reconstruction heads provide additional semantic supervision to the shared backbone. At inference, all auxiliary components are removed, and the model is used with its original architecture and inference cost. Experiments on representative unified multimodal backbones show that HiSVR improves visual understanding while maintaining image generation quality, demonstrating the effectiveness of hierarchical intermediate semantic supervision for UMM post-training.

Our contributions are summarized as follows:

- We propose **HiSVR**, a training-time hierarchical semantic visual reconstruction method for post-training unified multimodal models, introducing no additional inference cost.
- We introduce hierarchical intermediate supervision that aligns image-associated hidden states from multiple Transformer depths with semantic code prefixes of increasing length.
- We design an uncertainty-weighted progressive reconstruction loss to balance reconstruction objectives across different depth and prefix lengths.

2 Related Work

Unified Multimodal Models. Recent unified multimodal models differ mainly in how they represent images and how tightly they couple understanding and generation. Early-fusion models such as Chameleon (Team, 2024) and Emu3 (Wang et al., 2024b) formulate images and text as mixed token sequences and scale next-token prediction to multimodal data. Show-o (Xie

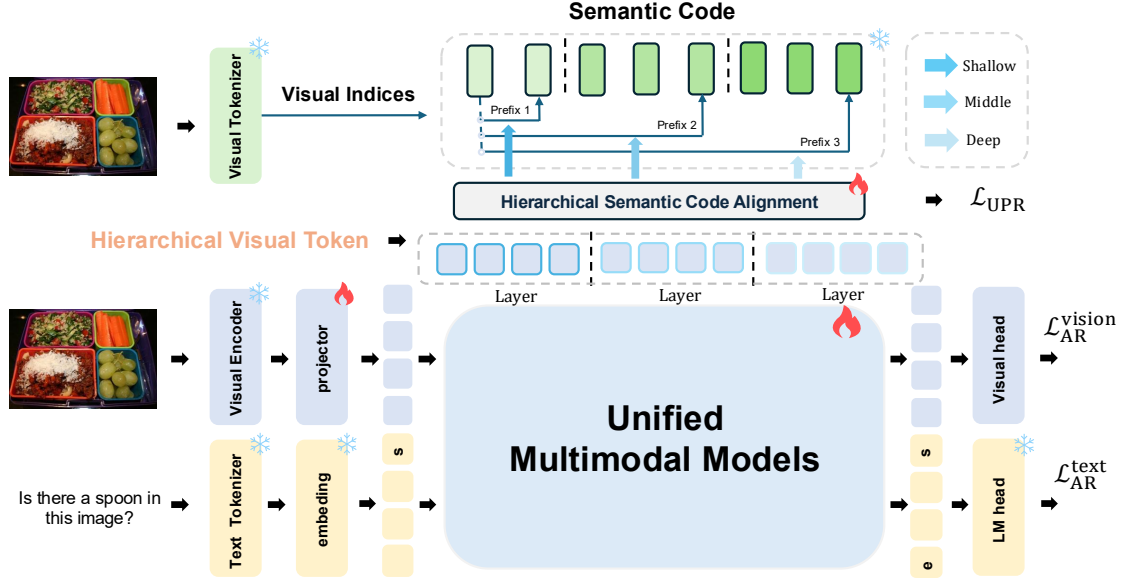


Figure 2: Overview of HiSVR. Given a training image, a frozen semantic tokenizer extracts residual semantic codes. HiSVR supervises image-associated hidden states from selected Transformer depths to reconstruct semantic code prefixes of different lengths through the HSCA module. The UPR loss adaptively balances the resulting layer-level reconstruction objectives. The semantic reconstruction branch is used only during post-training and is removed at inference. Snowflake and flame icons denote frozen and trainable modules, respectively.

et al., 2025b), BLIP3-o (Chen et al., 2025a), IL-LUME (Wang et al., 2024a), VILA-U (Wu et al., 2025d), and Harmon (Wu et al., 2025c) further explore shared backbones or harmonized visual representations for both understanding and generation. Another line improves the visual representation itself: Janus and Janus-Pro decouple visual pathways for understanding and generation (Wu et al., 2025a; Chen et al., 2025b), while TokenFlow, DualToken, and UniWorld study unified tokenizers or semantic encoders that better serve the two tasks (Qu et al., 2025; Song et al., 2025; Lin et al., 2025). More recently, LLaDA2.0-Uni extends unified modeling to diffusion large language models, supporting understanding, generation, editing, and interleaved reasoning within one framework (Bie et al., 2026). These works provide stronger architectures and visual token spaces; our work focuses on how post-training can shape the intermediate visual representations of an existing unified backbone.

Post-training of UMMs. Post-training has become a practical way to improve the interaction between understanding and generation after the base UMM is built. RecA aligns understanding and generation by reconstructing images from understanding-side representations (Xie et al., 2025a). UniMRG uses multi-representation

generation to turn generation-side information into supervision for visual understanding (Su et al., 2026), and UniGame studies self-adversarial post-training within a unified model (Su et al., 2025). Recent work further studies the bidirectional relation between the two abilities: UNO uses understanding supervision to steer visual generation (Liu et al., 2026), while LatentUMM focuses on latent-space consistency and semantic drift between generation and re-encoding (Luo et al., 2026). Preference- or reward-driven methods also optimize generated images with explicit preference objectives (Jiang et al., 2025). These methods mainly operate on final outputs, output-side representations, generated results, or latent transition consistency. HiSVR instead places semantic supervision directly on image-associated hidden states at multiple backbone depths. Related semantic visual reconstruction work shows that reconstructing semantic visual tokens is more beneficial for visual understanding than reconstructing low-level appearance tokens (Wang et al., 2025); we further extend this supervision from final representations to intermediate Transformer depths and use residual semantic code prefixes to construct hierarchical intermediate supervision.

3 Method

3.1 Overview

We propose **HiSVR**, a training-time hierarchical semantic visual reconstruction method for post-training unified multimodal models (UMMs), as shown in Figure 2. Given a training image, a frozen semantic visual tokenizer extracts residual semantic codes. HiSVR then supervises image-associated hidden states from selected Transformer depths to predict semantic code prefixes of different lengths through a **Hierarchical Semantic Code Alignment (HSCA)** module. The reconstruction targets become progressively longer at deeper layers, forming a coarse-to-fine schedule over the backbone depth. The resulting losses are balanced by an **Uncertainty-weighted Progressive Reconstruction (UPR)** loss. All HiSVR components are used only during post-training and are removed at inference.

3.2 Unified Multimodal Backbone

A unified multimodal model processes visual and textual signals through a shared Transformer backbone. Given a multimodal input sequence X , the backbone produces hidden states layer by layer:

$$H_l = F_l(X), \quad l = 1, \dots, L, \quad (1)$$

where L denotes the total number of Transformer layers. We denote the image-associated hidden states at layer l as:

$$H_l^I \in \mathbb{R}^{N \times d}, \quad (2)$$

where N is the number of image-associated tokens and d is the hidden dimension.

The definition of image-associated hidden states depends on the native UMM input format. For visual understanding samples, H_l^I corresponds to the hidden states at input visual token positions. For image generation samples, H_l^I corresponds to the hidden states at target image token positions. Thus, HiSVR can be added to different post-training samples without changing the original UMM formulation.

We denote the original UMM post-training objective as $\mathcal{L}_{\text{UMM}}$, which may include textual response supervision, image generation loss, or both. HiSVR adds an auxiliary representation-level loss on top of this native objective.

3.3 Hierarchical Semantic Visual Codes

For each training image I , we use a frozen semantic visual tokenizer $\mathcal{T}_{\text{sem}}$ to extract discrete residual semantic codes:

$$Q^{\text{sem}} = \mathcal{T}_{\text{sem}}(I) \in \mathbb{Z}^{P \times D}, \quad (3)$$

where P denotes the number of spatial positions and D denotes the number of residual code levels. In our implementation, the tokenizer is an RQ-VAE with a SigLIP encoder that produces codes with spatial resolution $P = 27 \times 27 = 729$ and $D = 8$ residual levels.

Each spatial position p is represented by a residual code sequence $(q_{p,1}, q_{p,2}, \dots, q_{p,D})$. We use prefixes of this sequence as reconstruction targets:

$$Q_{1:K}^{\text{sem}} = \{q_{p,j} \mid 1 \leq p \leq P, 1 \leq j \leq K\}. \quad (4)$$

The prefix length K controls the amount and difficulty of semantic information to be reconstructed, enabling auxiliary targets of different complexity.

3.4 Hierarchical Semantic Code Alignment

The HSCA module maps intermediate hidden states from selected Transformer depths to the semantic code space. It consists of spatial alignment, depth-specific projection, and prefix reconstruction.

Spatial alignment. The number of image-associated tokens N in the UMM may differ from the spatial size P of the semantic tokenizer. For example, the UMM may use $N = 32 \times 32 = 1024$ visual tokens while the semantic tokenizer operates at $P = 27 \times 27 = 729$ positions. To bridge this gap, we reshape H_l^I into a 2D spatial grid and apply bilinear interpolation:

$$\tilde{H}_l^I = \text{Interpolate}(H_l^I, N \rightarrow P) \in \mathbb{R}^{P \times d}. \quad (5)$$

Depth-specific projection. Let $\mathcal{S} = \{l_1, l_2, l_3\}$ denote the selected Transformer layers corresponding to shallow, middle, and deep positions. For each layer group $g \in \{\text{shallow}, \text{mid}, \text{deep}\}$, HSCA maintains an independent projection head $\mathcal{P}_g$ implemented as a two-layer MLP:

$$Z_l = \mathcal{P}_g(\tilde{H}_l^I) = W_2 \text{GELU}(W_1 \tilde{H}_l^I + b_1) + b_2, \quad (6)$$

where $Z_l \in \mathbb{R}^{P \times d_r}$ and d_r is the projection dimension.

Table 1: Comparison with understanding-only and unified multimodal models on visual understanding benchmarks. We further report our post-training results on both Show-o and Harmon backbones; baseline rows in the post-training block are filled from our server-side VLMEvalKit evaluations.

Type	Model	Encoder	LLM Scale	POPE↑	MME-P↑	MME-C↑	MMB↑	SEED↑	GQA↑	MMMU↑	MMVP↑	VSR↑	RWQ↑
Und. Only	LLaVA-v1.5 (Liu et al., 2023b)	CLIP ViT-L	1.3B	84.1	1128	—	—	—	56.5	30.7	—	—	—
	MobileVLM (Chu et al., 2023)	CLIP ViT-L	1.4B	84.5	1196	—	53.2	—	56.1	—	—	—	—
	MobileVLM-V2 (Chu et al., 2024)	CLIP ViT-L	1.4B	84.3	1303	—	57.7	—	59.3	—	—	—	—
	DeepSeekVL (Lu et al., 2024)	SigLIP-Large	1.3B	88.3	1307	225	64.6	—	59.3	33.8	—	—	—
	MiniGemini (Li et al., 2024)	CLIP ViT-L + ConvNeXt-L	2B	83.9	1341	312	59.8	—	59.9	—	—	—	—
Unified	ILLUME (Wang et al., 2024a)	OpenCLIP ViT-H	7B	88.5	1445	—	65.1	72.9	—	38.2	—	—	—
	TokenFlow-XL (Qu et al., 2025)	CLIP ViT-B(VQ)	13B	86.8	1546	—	68.9	68.7	62.7	38.7	—	—	—
	LWM (Liu et al., 2024a)	VQGAN	7B	75.2	—	—	—	—	44.8	—	—	—	—
	VILA-U (Wu et al., 2025d)	SigLIP-Large(VQ)	7B	85.8	1402	—	—	59.0	60.8	—	—	—	—
	Chameleon (Team, 2024)	VQGAN	7B	—	—	—	—	—	—	22.4	—	—	—
	D-DiT (Li et al., 2025)	VAE	2.0B	84.0	1125	—	—	—	59.2	—	—	—	—
	Janus (Wu et al., 2025a)	SigLIP-Large	1.3B	87.0	1338	222	69.4	63.7	59.1	30.5	—	—	—
	Janus-Pro (Chen et al., 2025b)	SigLIP-Large	1.5B	86.2	1444	268	75.5	68.3	59.3	36.3	—	—	—
	Harmon-0.5B (Wu et al., 2025c)	MAR-B	0.5B	86.5	1148	260	59.8	62.5	56.3	34.2	—	—	—
	Post-training Results												
Post-training on Show-o-1.3B													
Post-train	Show-o baseline	MAGVIT-v2	1.3B	80.8	1004	242.9	49.9	56.4	58.6	26.7	50.0	54.3	38.2
	+ SFT	MAGVIT-v2	1.3B	81.1	998	244.1	48.0	56.8	58.3	27.0	50.7	51.6	37.7
	+ RecA (Xie et al., 2025a)	MAGVIT-v2	1.5B	—	—	—	—	—	—	—	47.3	54.4	36.2
	+ UniMRG (Su et al., 2026)	MAGVIT-v2	1.5B	—	—	—	47.4	—	—	—	48.1	51.3	39.8
	+ Ours	MAGVIT-v2	1.3B	82.4	1048	258.4	51.3	59.1	60.3	29.4	52.2	55.3	40.7
Post-training on Harmon-1.5B													
Post-train	Harmon-1.5B baseline	MAR-H	1.5B	83.8	1158	266.4	49.2	65.0	58.9	35.6	59.3	61.1	47.2
	+ SFT	MAR-H	1.5B	84.1	1149	269.2	49.0	65.3	58.4	35.9	58.7	61.7	46.8
	+ RecA (Xie et al., 2025a)	MAR-H	1.5B	—	1223	—	—	65.2	58.8	35.7	61.0	59.0	49.5
	+ UniMRG (Su et al., 2026)	MAR-H	1.5B	—	—	—	52.2	—	—	—	62.7	61.2	51.9
	+ Ours	MAR-H	1.5B	85.2	1205	281.8	51.8	67.4	60.4	38.2	62.4	63.6	50.4

Prefix reconstruction. HSCA assigns a semantic code prefix length to each selected layer through a function $K(l)$. Our default configuration uses:

$$K(l_1) = 2, \quad K(l_2) = 5, \quad K(l_3) = 8, \quad (7)$$

where l_1, l_2, l_3 correspond to layers 8, 16, and 24 of a 24-layer Transformer, respectively. The selected layers reconstruct cumulative prefixes rather than disjoint code ranges. Thus, deeper layers retain the targets assigned to earlier layers while reconstructing additional residual levels.

For each selected layer l , the projected features Z_l are passed through level-specific classification heads to predict the residual codes:

$$\begin{aligned} \hat{Q}_{l,j} &= \mathcal{D}_j(Z_l), \\ \hat{Q}_{l,j} &\in \mathbb{R}^{P \times |V_j|}, \quad j = 1, \dots, K(l). \end{aligned} \quad (8)$$

where $|V_j|$ is the vocabulary size of the j -th residual semantic codebook.

3.5 Uncertainty-weighted Progressive Reconstruction (UPR) Loss

For each selected layer l and residual level $j \leq K(l)$, the reconstruction loss is:

$$\mathcal{L}_{l,j}^{\text{rec}} = -\frac{1}{P} \sum_{p=1}^P \log \hat{q}_{l,p,j}[q_{p,j}]. \quad (9)$$

Different layers and residual levels can have different reconstruction difficulty. We adopt homoscedastic uncertainty weighting (Kendall et al.,

2018) to balance these objectives. For each layer-level pair (l, j) , we introduce a learnable log-variance parameter $s_{l,j}$ and define:

$$\mathcal{L}_{\text{UPR}} = \sum_{l \in \mathcal{S}} \sum_{j=1}^{K(l)} \left(e^{-s_{l,j}} \mathcal{L}_{l,j}^{\text{rec}} + \frac{1}{2} s_{l,j} \right), \quad (10)$$

where the exponential term adaptively scales each reconstruction objective and the log-variance term prevents the model from trivially increasing all uncertainties.

3.6 Training Objective

The final training objective combines the original UMM loss with the UPR loss:

$$\mathcal{L} = \mathcal{L}_{\text{UMM}} + \lambda \mathcal{L}_{\text{UPR}}, \quad (11)$$

where λ controls the weight of the hierarchical semantic reconstruction.

During training, the semantic visual tokenizer is frozen. The UMM backbone, the HSCA projection heads, and the semantic decoder are optimized jointly. During inference, all auxiliary components (HSCA module, semantic decoder, and uncertainty parameters) are removed, introducing no additional inference cost.

4 Experiments

4.1 Experimental Setup

Model Architectures. We evaluate HiSVR on two representative UMM architectures with different generation paradigms:

Table 2: Evaluation of text-to-image generation on GenEval (Ghosh et al., 2023) benchmark. We compare representative generation-only and unified multimodal models, and further report post-training results on the Show-o backbone.

Type	Method	Single Obj.↑	Two Obj.↑	Counting↑	Colors↑	Position↑	Color Attri.↑	Overall↑
<i>Gen. Only</i>	SDXL (Podell et al., 2024)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 (Betker et al., 2023)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium (Esser et al., 2024)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
<i>Unified</i>	SEED-X (Ge et al., 2024)	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	D-DiT (Li et al., 2025)	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	TokenFlow-XL (Qu et al., 2025)	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	OmniGen2 (Wu et al., 2025b)	0.99	0.86	0.64	0.85	0.31	0.55	0.70
	Chameleon (Team, 2024)	—	—	—	—	—	—	0.39
	Janus-Pro (Chen et al., 2025b)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Ours Post-training Results								
<i>Post-training on Show-o-1.3B</i>								
	Show-o baseline	0.97	0.84	0.65	0.80	0.27	0.45	0.66
	+ SFT	0.96	0.84	0.67	0.77	0.30	0.54	0.68
<i>Post-train</i>	+ RecA (Xie et al., 2025a)	—	—	—	—	—	—	0.72
	+ UniMRG (Su et al., 2026)	—	—	—	—	—	—	0.71
	+ Ours	0.98	0.88	0.67	0.81	0.34	0.56	0.71

- **AR: Show-o (Xie et al., 2025b).** Show-o is a unified multimodal Transformer that supports both visual understanding and image generation. We apply HiSVR to intermediate hidden states at visual-token positions in its multimodal backbone.
- **AR+MAR: Harmon (Wu et al., 2025c).** Harmon uses a shared masked autoregressive backbone for visual comprehension and generation. We apply the same HiSVR objective to its unified backbone to examine whether the proposed supervision transfers across generation paradigms.

For both models, the HSCA module and UPR loss are used only during post-training and are removed at inference time, leaving the original inference pathway unchanged.

Implementation Details. We use LLaVA-Mix-665K as visual instruction data and BLIP3o-60K image-text data for generation-side post-training. For each training image, we pre-extract $27 \times 27 \times 8$ hierarchical semantic visual codes using a frozen RQ-VAE semantic tokenizer (Lee et al., 2022). Unless otherwise specified, HSCA is attached to three Transformer depths (layers 8, 16, and 24), following the default configuration in Section 3.4. More implementation details are provided in Appendix A.

Evaluation Metrics. For visual understanding, we use VLMEvalKit to evaluate POPE, MME, MMBench, SEEDBench, GQA, MMMU, MMVP, VSR, RealWorldQA, TextVQA, and ScienceQA.

For image generation, we report GenEval (Ghosh et al., 2023). Appendix B gives the detailed evaluation protocols.

4.2 Results on Visual Understanding

Table 1 presents the main visual understanding results. On the Show-o backbone, HiSVR improves the post-trained model on the majority of evaluated benchmarks, with particularly consistent gains on perception-oriented tasks. On Harmon, the same training objective also yields improvements, suggesting that the benefit is not tied to a single generation paradigm. We further provide qualitative comparisons in Figure 3.

4.3 Results on Image Generation

Table 2 reports the text-to-image generation results on the GenEval benchmark. On the Show-o backbone, HiSVR achieves an overall score of 0.71, improving over both the original Show-o baseline (0.66) and the SFT baseline (0.68). Compared with SFT, HiSVR improves on several composition- and attribute-sensitive categories, including Two Object (0.88 vs. 0.84), Position (0.34 vs. 0.30), and Color Attribution (0.56 vs. 0.54), while matching Counting (0.67). These results indicate that intermediate semantic reconstruction preserves image generation ability while improving several relation- and attribute-sensitive generation cases.

4.4 Case Study

Figure 3 presents representative cases for visual understanding and image generation. For visual understanding, HiSVR better preserves fine-

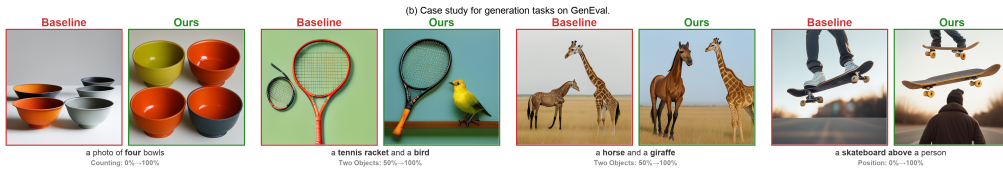
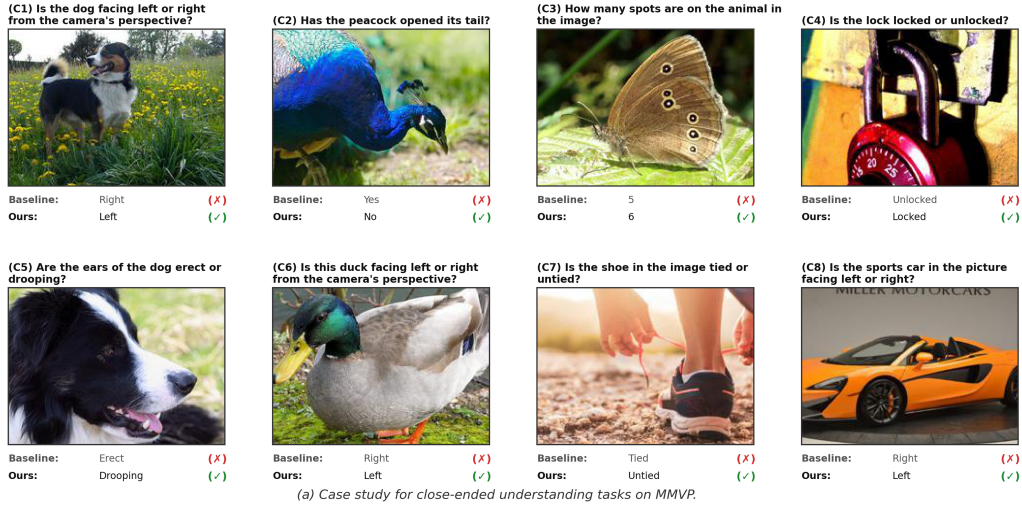


Figure 3: Qualitative case studies on visual understanding and image generation. (a) Eight challenging examples from the MMVP benchmark where the baseline model fails but our model succeeds, spanning spatial orientation (C1, C6, C8), object state recognition (C2, C4, C7), counting (C3), and attribute perception (C5). ✓ and ✗ denote correct and incorrect answers, respectively. (b) Four representative text-to-image generation examples from GenEval where our method outperforms the baseline on counting, two-object composition, and spatial relationship tasks.

grained cues such as object state, spatial orientation, counting, and attributes. For image generation, HiSVR follows compositional prompts more faithfully, especially when multiple objects, color attributes, or spatial relations are involved. These examples complement the quantitative results and illustrate how intermediate semantic supervision improves sensitivity to visual structure without changing the inference architecture.

4.5 Ablation Studies

We conduct ablation studies to isolate the contribution of the main design choices in HiSVR. Table 3 summarizes the results.

Table 3: Ablation studies of HiSVR on Show-o visual understanding benchmarks.

Method	MMMU	MMBench	POPE	Overall
<i>Post-training on Show-o-1.3B</i>				
Final-layer only	27.9	49.4	80.6	52.5
Inverted prefix schedule	28.6	48.7	81.0	52.9
Disjoint prefix ranges	28.2	50.3	80.8	53.4
Uniform loss weighting	28.9	50.7	81.6	53.8
HiSVR (Ours)	29.4	51.2	82.3	54.3

Effect of multi-depth supervision. We first compare HiSVR with a final-layer-only variant

that applies semantic reconstruction only to the last Transformer layer using the full code prefix. This comparison tests whether semantic supervision should be concentrated near the output side or distributed across intermediate depths.

Effect of prefix schedule. We then compare the proposed cumulative prefix schedule with two alternatives. The inverted prefix schedule assigns the longest prefix to the shallow layer and the shortest prefix to the deepest layer. The disjoint prefix schedule assigns non-overlapping residual code ranges to different depths. These variants test whether the coarse-to-fine cumulative design matters beyond simply adding reconstruction losses at multiple layers.

Effect of UPR loss. We compare the UPR loss (Eq. 10) with a uniform weighting baseline where all reconstruction objectives are equally weighted. This comparison evaluates whether adaptive loss balancing is useful when reconstruction targets differ in depth, prefix length, and optimization difficulty.

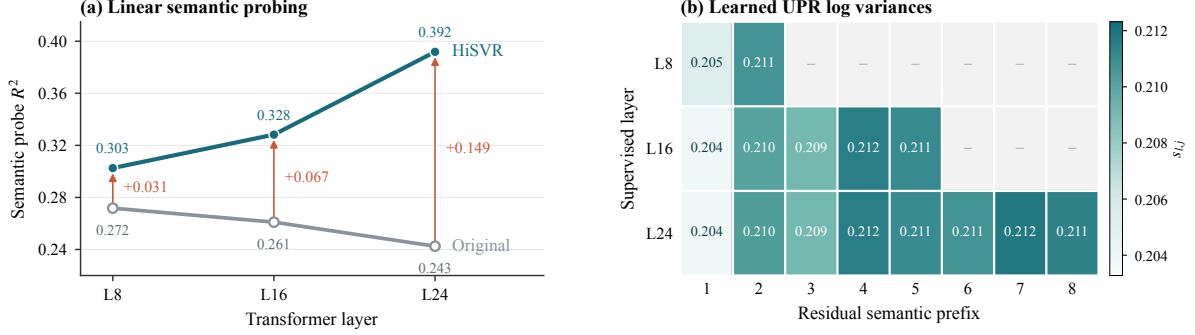


Figure 4: Layer-wise semantic diagnostics and learned UPR balancing on the Show-o backbone. (a) A lightweight linear probe predicts frozen RQ-VAE semantic-token features from image-associated hidden states at layers 8, 16, and 24. Higher R^2 indicates that semantic visual structure is more linearly recoverable from the corresponding Transformer layer. (b) Learned UPR log variances $s_{l,j}$ over assigned layer-prefix objectives; blank cells denote prefix levels not used by the cumulative schedule.

4.6 Analysis

The benchmark results show that HiSVR improves downstream understanding and generation, but they do not by themselves reveal where the improvement comes from inside the UMM. Since HiSVR explicitly supervises image-associated hidden states at multiple Transformer depths, the most direct diagnostic is to ask whether visual semantics become easier to recover from those intermediate representations. We therefore focus our analysis on two complementary questions: whether frozen hidden states expose stronger semantic information after HiSVR, and whether the learned UPR weights behave as a stable balancing mechanism across heterogeneous layer-prefix objectives.

Layer-wise semantic probing. To directly test whether HiSVR improves intermediate visual representations, we freeze the backbone and train a lightweight linear probe on image-associated hidden states from layers 8, 16, and 24. The probe predicts patch-level semantic features extracted by the frozen RQ-VAE semantic tokenizer, using unlabeled POPE images only as natural image inputs. We sample 96 semantic-token positions per image from 240 images and use an image-level 75/25 train-test split, so no patch from a test image is seen during probe training. As shown in Figure 4(a), HiSVR consistently improves semantic recoverability over the original Show-o backbone, increasing R^2 from 0.272 to 0.303 at layer 8, from 0.261 to 0.328 at layer 16, and from 0.243 to 0.392 at layer 24. The same trend appears under cosine similarity, which improves from

0.713/0.709/0.696 to 0.726/0.740/0.748 across the three layers. The gain is largest at the deepest probed layer, but improvements also appear at the shallow and middle layers, indicating that HiSVR does not merely alter the final decoding interface. Instead, it makes frozen visual semantics more accessible throughout the supervised depth range, which directly supports our main hypothesis that hierarchical semantic reconstruction improves the organization of intermediate UMM representations.

Learned UPR weights. Figure 4(b) shows the learned UPR log variances for all assigned layer-prefix objectives. The learned values remain within a narrow range, but they are target-specific across layers and residual prefix levels rather than manually fixed. This mild adaptive reweighting is consistent with the role of UPR: it balances heterogeneous reconstruction targets while keeping the auxiliary supervision numerically stable. We provide the corresponding reconstruction-loss curves in Appendix C.

5 Conclusion

We propose HiSVR, a training-time hierarchical semantic visual reconstruction method for unified multimodal model post-training. HiSVR supervises image-associated hidden states with residual semantic code prefixes, providing intermediate semantic supervision beyond conventional output-side objectives. Since the auxiliary reconstruction branch is removed at inference, HiSVR preserves the original model architecture and inference cost. Experiments on unified multimodal

backbones show that HiSVR improves visual understanding while maintaining image generation quality, suggesting that coarse-to-fine semantic reconstruction is an effective way to regularize intermediate multimodal representations.

Limitations

Our work has several limitations. First, HiSVR relies on a frozen semantic visual tokenizer, so its supervision quality is bounded by the tokenizer’s capacity. Second, we evaluate HiSVR on two unified multimodal backbones with up to 1.5B parameters, and its effectiveness on larger models remains to be studied. Third, the current layer selection and prefix schedule are fixed, while adaptive strategies may further improve performance. Finally, although HiSVR adds no inference cost, it introduces extra training-time computation. Its real-world use may also inherit risks from the underlying model and data, including hallucinated visual answers, biased generations, and failures under out-of-distribution inputs. Therefore, high-stakes applications require additional safety evaluation and human verification.

References

- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions. <https://cdn.openai.com/papers/dall-e-3.pdf>.
- Tiwei Bie, Haoxing Chen, Tiejuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, Jianguo Li, Tao Lin, Qi Qin, Hongjun Wang, Xiaomei Wang, Haoyuan Wu, Yi Xin, and Junbo Zhao. 2026. Llada2.0-uni: Unifying multimodal understanding and generation with diffusion large language model. *arXiv preprint arXiv:2604.20796*.
- Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. 2025a. BLIP3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. 2023. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.
- Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.
- Drew A. Hudson and Christopher D. Manning. 2019. 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*, pages 6700–6709.
- Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. *arXiv preprint arXiv:2505.00703*.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.
- Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532.

- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. SEED-bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.
- Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. 2025. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2779–2790.
- Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. 2025. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. **Visual spatial reasoning**. *Transactions of the Association for Computational Linguistics*, 11:635–651.
- Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Zeyu Liu, Zanlin Ni, Yang Yue, Cheng Da, Huan Yang, Di Zhang, Kun Gai, and Gao Huang. 2026. Steering visual generation in unified multimodal models with understanding supervision. *arXiv preprint arXiv:2605.05781*.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Yinyi Luo, Wenwen Wang, Hayes Bai, Marios Savvides, and Jindong Wang. 2026. Latentmm: Dual latent alignment for unified multimodal models. *arXiv preprint arXiv:2605.17766*.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*.
- Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. 2025. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555.
- Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. 2025. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*.
- Zhaolong Su, Wang Lu, Hao Chen, Sharon Li, and Jindong Wang. 2025. Unigame: Turning a unified multimodal model into its own adversary. *arXiv preprint arXiv:2511.19413*.
- Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. 2026. Generation enhances understanding in unified multimodal models via multi-representation generation. In *International Conference on Learning Representations*.
- Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.
- Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. 2024a. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*.
- Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. 2025. Autoregressive semantic visual reconstruction helps vlms understand better. *arXiv preprint arXiv:2506.09040*.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao,

- Bowen Zhang, Liangdong Wang, Guang Liu, and 6 others. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2025a. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12966–12977.
- Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xian Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, and 3 others. 2025b. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*.
- Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. 2025c. Harmonizing visual representations for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17739–17750.
- Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2025d. VILA-U: A unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*.
- Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. 2025a. [Reconstruction alignment improves unified multimodal models](#). *Preprint*, arXiv:2509.07295.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025b. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.

Appendix Contents

A	Implementation Details	12
B	Evaluation Protocols	12
C	Additional Analysis Details	13

A Implementation Details

A.1 Post-training Data

We use publicly available research artifacts, including Show-o, Harmon, LLaVA-style instruction data, LLaVA-NeXT-738K, BLIP3o-60K, GenEval, and VLMEvalKit, following their released licenses and terms of use. The semantic tokenizer is used only for training-time supervision and is not redistributed as part of the submitted paper.

For Show-o post-training, we use LLaVA-style visual instruction data for image-to-text learning and BLIP3o-60K text-image shards for text-to-image learning. The Show-o configuration uses the full BLIP3o-60K shard set, including DALL-E 3, GenEval-train, human gesture, JourneyDB, MSCOCO-human, object, occupation, and text-oriented subsets. For Harmon post-training, we use LLaVA-NeXT-738K for multimodal understanding and BLIP3o-60K for text-to-image generation. Unless otherwise stated, we follow the official training and evaluation splits of each benchmark.

A.2 Semantic Code Extraction

For each training image, we extract hierarchical semantic visual codes offline using a frozen RQ-VAE semantic tokenizer with a SigLIP encoder. The tokenizer produces a spatial code map with resolution 27×27 , corresponding to $P = 729$ spatial positions, and $D = 8$ residual code levels. These residual levels form the semantic code hierarchy used by HiSVR. The semantic tokenizer is frozen throughout post-training and is not used at inference.

A.3 Post-training Hyperparameters

Tables 4–6 summarize the post-training setup used for the two UMM backbones. We split the configuration into optimization settings, data/tokenizer settings, and HiSVR-specific module settings to keep the appendix readable. The Show-o setup follows the discrete HiSVR configuration with a 15K-step schedule, while the Harmon setup follows the

Table 4: Optimization hyperparameters for HiSVR post-training.

	Show-o-1.3B	Harmon-1.5B
Optimizer	AdamW	AdamW
Learning rate	5×10^{-7}	1×10^{-5}
HiSVR branch learning rate	5×10^{-5}	1×10^{-5}
β_1, β_2	(0.9, 0.999)	(0.9, 0.95)
Weight decay	0.01	0.02
Learning-rate schedule	Cosine	Linear warmup + cosine
Warmup	500 steps	1% of total iterations
Training length	15K steps	15K iterations
Gradient accumulation	5	40
Batch size	T2I: 1, MMU: 2	32
Mixed precision	bfloat16	bfloat16
Max gradient norm	1.0	1.0
GPUs	$1 \times \text{NVIDIA A800}$	$1 \times \text{NVIDIA A800}$

Table 5: Data and tokenizer setup for HiSVR post-training.

	Show-o-1.3B	Harmon-1.5B
Understanding data	LLaVA-style instruction data	LLaVA-NeXT-738K
Generation data	BLIP3o-60K shards	BLIP3o-60K shards
Generation resolution	512×512	512×512
Max text length	128	128
Generation tokenizer	MAGVIT-v2	MAR tokenizer
Semantic tokenizer	Frozen RQ-VAE semantic tokenizer	Frozen RQ-VAE semantic tokenizer
Tokenizer encoder	SigLIP	SigLIP
Semantic code resolution	27×27	27×27
Residual code levels	8	8

MMU+T2I HiSVR configuration launched on a single NVIDIA A800 GPU.

A.4 Inference-time Auxiliary Branch Removal

HiSVR is used only during post-training. At inference, the semantic tokenizer, HSCA projection heads, semantic decoder, and uncertainty parameters are removed. Therefore, HiSVR does not change the native model architecture, decoding procedure, or inference cost of the underlying UMM.

B Evaluation Protocols

B.1 Visual Understanding Benchmarks

We evaluate visual understanding with VLMEvalKit when the benchmark is supported. The benchmark suite includes POPE, MME, MMBench, SEEDBench, GQA, MMMU, MMVP, VSR, RealWorldQA, TextVQA, and ScienceQA. We use the official split and metric for each benchmark unless otherwise specified.

- **POPE** (Li et al., 2023b): object-hallucination evaluation with a balanced, image-grounded design; the test split contains 9000 question-answer pairs.
- **MME** (Fu et al., 2023): a comprehensive

Table 6: HiSVR module and loss-weight setup.

	Show-o-1.3B	Harmon-1.5B
Selected Transformer layers	8, 16, 24	9, 18, 28
Depth ranges	[0, 2), [0, 5), [0, 8)	[0, 2), [0, 5), [0, 8)
Prefix schedule $K(l)$	2, 5, 8	2, 5, 8
Projection head	2-layer MLP	2-layer MLP
Projection dimension	2048	2048
Semantic decoder	Level-specific classifiers	Level-specific classifiers
Frozen modules	Semantic tokenizer	Semantic tokenizer
Trainable modules	Backbone, HSCA, decoder	Backbone, HSCA, decoder
Inference-time auxiliary modules	Removed	Removed
Text-to-image loss	1.0	1.0
Image-to-text/MMU loss	1.0	1.0
HiSVR loss	1.0	1.0
UPR uncertainty weights	Learned	Learned

benchmark covering perception and cognition abilities of multimodal models.

- **MMBench** (Liu et al., 2024b): a curated multiple-choice benchmark; it contains 1164 dev questions and 1784 test questions.
- **SEEDBench** (Li et al., 2023a): a multi-dimensional benchmark for evaluating generative comprehension across diverse visual understanding dimensions.
- **GQA** (Hudson and Manning, 2019): a compositional visual question answering benchmark requiring multi-step reasoning over scene graphs.
- **MMMU** (Yue et al., 2024): a college-level multimodal reasoning benchmark with 11 500 questions across multiple disciplines.
- **MMVP** (Tong et al., 2024): a benchmark targeting fine-grained visual patterns that are difficult for CLIP-based vision-language models.
- **VSR** (Liu et al., 2023a): a visual spatial reasoning benchmark focused on spatial relationships between objects.
- **RealWorldQA**: a visual question answering benchmark built from diverse real-world images.
- **TextVQA**: a text-centric visual question answering benchmark that requires reading and reasoning over image text.
- **ScienceQA**: a multimodal science question answering benchmark covering visual and textual reasoning.

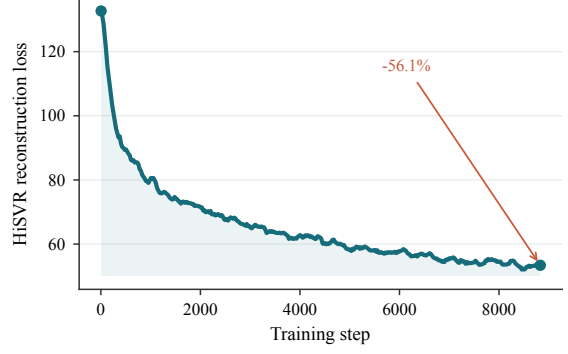


Figure 5: Overall auxiliary semantic reconstruction loss during HiSVR post-training on the Show-o backbone. The curve is smoothed only for visualization.

B.2 Text-to-image Generation Benchmark

We evaluate text-to-image generation with GenEval (Ghosh et al., 2023). Following the official protocol, generated images are scored on six compositional dimensions: single object, two objects, counting, colors, position, and color attribution. We report both category-level scores and the overall score in Table 2.

C Additional Analysis Details

C.1 Diagnostic Probing Protocol

For the layer-wise semantic probing experiment in Figure 4(a), we freeze the original and HiSVR-tuned Show-o backbones and extract image-associated hidden states from Transformer layers 8, 16, and 24. We use 240 images from POPE as unlabeled natural image inputs and do not use their question or answer annotations. For each image, we sample 96 positions from the 27×27 semantic-token grid and align the corresponding Show-o image-token hidden states to this grid by bilinear interpolation from the native 32×32 VQ-token layout. The probing target is the continuous patch-level semantic feature from the frozen RQ-VAE semantic tokenizer. We train an independent ridge-linear probe for each model and layer with an image-level 75/25 train-test split, preventing patches from the same image from appearing in both splits. We report test-set R^2 in the main figure and use cosine similarity as an additional diagnostic statistic in the text.

C.2 Training Diagnostics

Training dynamics. Figure 5 shows the auxiliary reconstruction objective from the Show-o

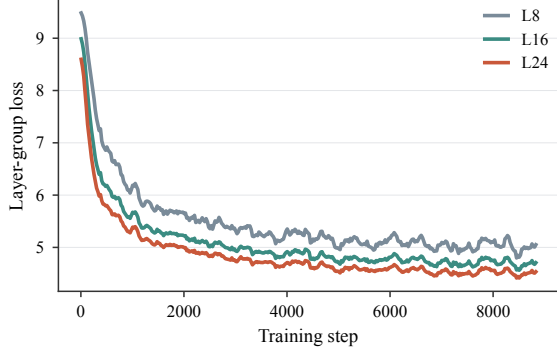


Figure 6: Layer-group reconstruction losses at the selected Show-o Transformer depths. The curves are smoothed only for visualization.

HiSVR training log. Using the average of the first and last 40 logged points, the HiSVR reconstruction loss decreases from 120.84 to 53.09, corresponding to a relative reduction of 56.1%. This confirms that the semantic reconstruction branch is actively optimized during post-training rather than serving as a static regularizer.

Layer-wise reconstruction behavior. We further inspect the reconstruction losses at the selected Transformer depths in Figure 6. The layer groups at depths 8, 16, and 24 reduce their reconstruction losses by 44.4%, 43.1%, and 41.7%, respectively, but they maintain different magnitudes throughout training. This supports supervising multiple depths: the intermediate targets are optimized successfully, but they exhibit distinct difficulty profiles rather than collapsing to a single shared trend.