Autoregressive Semantic Visual Reconstruction Helps VLMs Understand Better

Dianyi Wang[1,2*] **Wei Song**[2,3,4*] **Yikun Wang**[1,2] **Siyuan Wang**[6]
Kaicheng Yu[3] **Zhongyu Wei**[1,2†] **Jiaqi Wang**[2,5†]
[1]Fudan University [2]Shanghai Innovation Institute [3]AutoLab, Westlake University
[4]Zhejiang University [5]Shanghai AI Lab [6]University of Southern California

Abstract

Typical large vision-language models (LVLMs) apply autoregressive supervision primarily to textual responses, without fully exploiting causal learning over rich visual inputs. As a result, these models often emphasize vision-to-language alignment while potentially overlooking fine-grained visual information. While prior work has explored autoregressive image generation, effectively leveraging autoregressive visual supervision to enhance image understanding remains an open challenge. In this paper, we introduce Autoregressive Semantic Visual Reconstruction (ASVR), which enables joint learning of visual and textual modalities within a unified autoregressive framework. ASVR trains models to autoregressively reconstruct the semantic content of input images, which consistently enhances multimodal comprehension. Notably, we show that even when provided with continuous image features as input, models can effectively reconstruct discrete semantic tokens, resulting in stable and consistent improvements across various multimodal understanding benchmarks. ASVR delivers significant performance gains and scalability across varying data scales, visual input, visual supervision and model architectures. In particular, ASVR generally improves baselines by 2-3% across 14 multimodal benchmarks. The code is available at `https://github.com/AlenjandroWang/ASVR`.

1 Introduction

The success of large language models (LLMs) has demonstrated the tremendous potential and scalability of the autoregressive (AR) paradigm. Recent advances extending LLMs' powerful capabilities to multimodal understanding through bridge-style architectures, exemplified by LLaVA (Liu et al., 2023b, 2024a,b), have achieved remarkable performance across vision-language tasks (Liu et al., 2023c; Yue et al., 2023; Fu et al., 2024a; Goyal et al., 2017; Li et al., 2023b; Hudson and Manning, 2019a; Kembhavi et al., 2016; Li et al., 2025). These models (Bai et al., 2023b; Wang et al., 2024c; Yao et al., 2024; Chen et al., 2024; Lu et al., 2024; Wu et al., 2024c), typically adopt a learnable projector to align features from a pretrained visual encoder into the text embedding space of LLMs.

However, most large vision-language models (LVLMs)(Wang et al., 2024d; Dong et al., 2024; Liu et al., 2024c; Li et al., 2024) supervise only textual outputs through next-token prediction, while overlooking the causal learning of rich visual content within input images. Although this limitation is partially mitigated by training on image–caption pairs that associate visual content with language, the visual modality expresses far more than text alone, capturing spatial relationships, textures, complex compositions, and subtle stylistic cues that language cannot fully convey. For example, LLaVA-1.5(Liu et al., 2023a) represents a single 336×336 image with 576 visual tokens, which collectively encode substantially more information than the associated caption, yet applies no explicit supervision to this visual content.

While recent unified models (Deng et al., 2025; Dong et al., 2025b; Zhang et al., 2025) have explored integrating visual understanding and generation within the autoregressive paradigm (Team, 2024; Wang et al., 2024e; Wu et al., 2024b; Tong et al., 2024b), visual tokens are typically supervised only on the output side through visual generation objectives. In contrast, visual tokens fed as inputs are not explicitly supervised to enhance visual understanding. Effectively supervising autoregressive visual inputs to improve fine-grained visual understanding remains an open challenge. Most recently, Wang et al. (2024a) proposed reconstructing visual inputs via denoising, yet their method relies on external Diffusion Transformer (DiT) modules and lacks a unified next-token pre-

---

[*] Co-first authors.

[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 38101–38115
July 2-7, 2026 ©2026 Association for Computational Linguistics

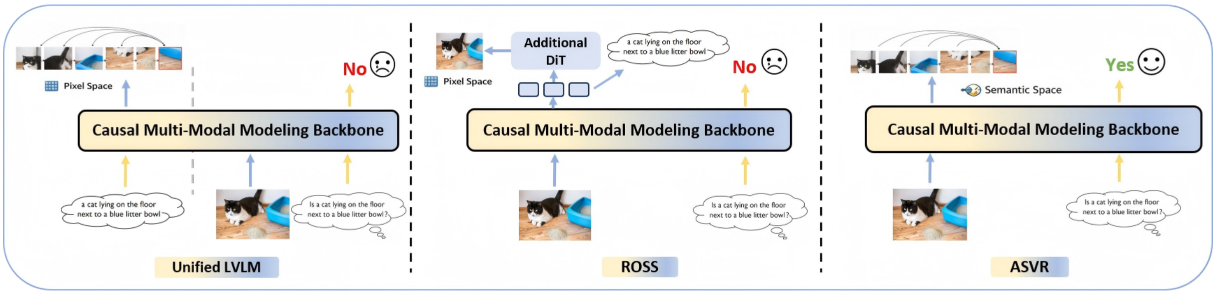

Figure 1: **(Left) Unified LVLMs** aim to implicitly improve visual understanding through additional text-to-image (T2I) objectives. **(Mid) ROSS** introduces an auxiliary diffusion transformer (DiT) to reconstruct pixel-level image information conditioned on latent representations from the LVLM. **(Right) ASVR** directly reconstructs the semantic content of the input image, enhancing semantic-level alignment.

diction framework that explicitly encourage causal learning over detailed visual information and dependencies rather than high-level bidirectional contextual correlations.

In this paper, we introduce **Autoregressive Semantic Visual Reconstruction (ASVR)**, a method that enables joint learning of the visual input and textual output within a unified autoregressive paradigm, without relying on any external modules. Specifically, ASVR allows LVLMs to autoregressively predict the next discrete semantic token of an input image, which is prepared by a pretrained semantic visual tokenizer (Song et al., 2025; Wu et al., 2024b; Qu et al., 2024; Xie et al., 2024). Autoregressively reconstructing semantic visual representations consistently enhances the visual understanding capabilities and further improve reasoning performance. Notably, we find that models can effectively reconstruct discrete semantic tokens even when provided with continuous image features as input. This setting yields substantial gains over approaches that use shared discrete semantic visual tokens for both input and output, and also outperforms the denoising-based visual reconstruction paradigm (Wang et al., 2024b).

Our approach delivers significant and consistent gains across varying data scales (LLaVA-1.5-665K (Liu et al., 2023a), LLaVA-Next-779K (Liu et al., 2024c), Bunny-v1_1-data-2M (He et al., 2024),LLaVA-OV-4M (Li et al., 2024)) as well as across model architectures including Vicuna family (Zheng et al., 2023) and Mistral (Jiang et al., 2023). Specifically, **ASVR** improves baselines by 2-3% on average across 14 multimodal benchmarks with improvements remaining robust across visual feature types, LLM backbone capacities, data scales, and high-resolution scenarios. These results underscore the importance of explicit semantic visual supervision in training LVLMs. ASVR not only improves visual understanding but also introduces a scalable, unified training strategy, offering a new perspective on autoregressive modeling for multimodal systems.

2 Preliminaries

Large Vision Language Models To process and represent input from different modalities in a unified manner, LVLMs typically comprise three components: a pre-trained LLM core, a projector commonly implemented as a two-layer MLP, and a pre-trained visual encoder with semantic alignment.

Given a input RGB image $I \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the image height and width, a pre-trained visual encoder V_ξ first extracts visual features $\mathbf{z}^I = V_\xi(I)$. These features are then projected into the LLM embedding space through a projector P_ϕ, yielding a sequence of visual embeddings: $\mathbf{H}^I = P_\phi(\mathbf{z}^I) \in \mathbb{R}^{m \times d}$, where $m = h \times w$ denotes the length of visual features, and d is the embedding dimension of LLM. ξ and ϕ are the parameters of the visual encoder and projector, respectively. For a textual input $T \in \mathbb{Z}^L$, the LLM tokenizer produces a sequence of token indices $\mathbf{x}^T = \text{Tokenizer}(T) \in \mathbb{R}^n$. which are then mapped into textual embeddings via the LLM's embedding layer $\mathbf{H}^T = \text{Embedding}(x^T) \in \mathbb{R}^{n \times d}$ where n denotes the text sequence length.

The multimodal input is formed by concatenating the visual and textual embeddings as $[\mathbf{H}^I, \mathbf{H}^T] \in \mathbb{R}^{(m+n) \times d}$, which is then fed into a causal LLM backbone L_θ with parameters θ for unified autoregressive modeling:

$$L_\theta([\mathbf{H}^I, \mathbf{H}^T]) = \prod_{i=1}^{n} L_\theta(x_i^T \mid x_{<i}^T, \mathbf{H}^I) \quad (1)$$

Training Pipeline for LVLMs LVLMs training typically follows a two-stage paradigm (Liu et al., 2023b): pre-training and instruction tuning. During pre-training, the model learns to align visual and textual modalities, enabling joint understanding of multimodal inputs. Instruction tuning further enhances generalization across diverse downstream tasks such as visual question answering (VQA). The training objective is to maximize the likelihood of target textual responses in an autoregressive manner, with supervision applied only to textual responses. In practice, pre-training usually updates only the projector parameters ϕ, while instruction tuning additionally fine-tunes the LLM parameters θ. The visual encoder v_ξ may either remain frozen (Liu et al., 2023b; Tong et al., 2024a) or be jointly optimized (Li et al., 2024; Dong et al., 2024; Wang et al., 2024d; Liu et al., 2024c).

3 Method

In this section, we introduce **ASVR** which learns autoregressive modeling of textual responses while simultaneously reconstructs visual inputs autoregressively to enhance visual understanding. An overview of the method is provided in Section 3.1, followed by detailed descriptions of the visual tokenizer and visual encoder in Sections 3.2 and 3.3, respectively. The training procedure is described in Section 3.4. Figure 2 provides a detailed comparison between conventional LVLMs (e.g., LLaVA) and our ASVR, highlighting the key innovation of incorporating autoregressive visual input supervision to enhance multimodal understanding.

3.1 Overview

We incorporate autoregressive visual supervision into typical LVLMs described in Section 2 by extending the next-token prediction paradigm to reconstruct visual inputs. This unified formulation enables the model to seamlessly integrate visual and textual information, thereby establishing a perceptual foundation for image understanding, alleviating information loss caused by text-only supervision, and ultimately enhancing multimodal understanding and reasoning capabilities.

As illustrated in Figure 2(b), we employ a visual tokenizer to convert the input image I into a discrete sequence of visual token indices $\mathbf{x}^I \in \mathbb{R}^m$, which serve as visual supervision signals and m matches the length of the visual features sequence $\mathbf{H}^I$ extracted by the pre-trained visual encoder and fed into the LLM backbone. A visual prediction head tailored to the visual tokenizer is trained to autoregressively predict the next visual token, analogous to textual supervision:

$$\mathcal{L}_{\text{AR}}^{\text{vision}}(\Theta, I) = -\frac{1}{m}\sum_{i=1}^{m} \log L_\theta(x_i^I \mid x_{<i}^I), \quad (2)$$

where $\Theta = \{\theta, \xi, \phi\}$ denotes the parameters of the LLM backbone, visual encoder and projector. The final training objective jointly optimizes visual and textual autoregressive losses, $\mathcal{L}_{\text{AR}}^{\text{vision}}$ and $\mathcal{L}_{\text{AR}}^{\text{text}}$:

$$\mathcal{L}_{\text{AR}}(\Theta, I, T) = \mathcal{L}_{\text{AR}}^{\text{vision}} + \mathcal{L}_{\text{AR}}^{\text{text}} \quad (3)$$

This design unifies the learning paradigm across modalities, and encourages the model to first develop a coherent visual understanding, which subsequently serves as a foundation for more accurate and contextually grounded multimoda reasoning.

3.2 Visual Tokenizer

Unlike continuous visual encoders, a visual tokenizer converts an input image into a one-dimensional sequence of discrete visual codes through vector quantization (VQ) by learning a fixed-size visual codebook. The corresponding embeddings are retrieved from the codebook based on these codes and used as inputs to the LLM backbone. In this work, we utilize such a visual tokenizer to obtain discrete visual token representations as supervision targets for visual reconstruction. Existing visual tokenizers can be broadly categorized into two types.

Visual Appearance Tokenizer A visual appearance tokenizer (Esser et al., 2021; Team, 2024) is optimized with the objective of reconstructing the appearance of the input image, typically using a combination of pixel-wise L2 loss (Dosovitskiy and Brox, 2016), LPIPS loss(Zhang et al., 2018) and adversarial loss (Isola et al., 2017). The resulting sequence of token indices corresponds to a quantized mapping of the image's pixel-level features. Using appearance tokenizers to provide pixel-level supervision will encourage LVLMs to focus on low-level visual feature reconstruction.

Visual Semantic Tokenizer A visual semantic tokenizer (Qu et al., 2024; Wu et al., 2024b; Xie et al., 2024; Song et al., 2025) is trained to align image features with textual semantics, typically using a contrastive loss (Radford et al., 2021) to enhance

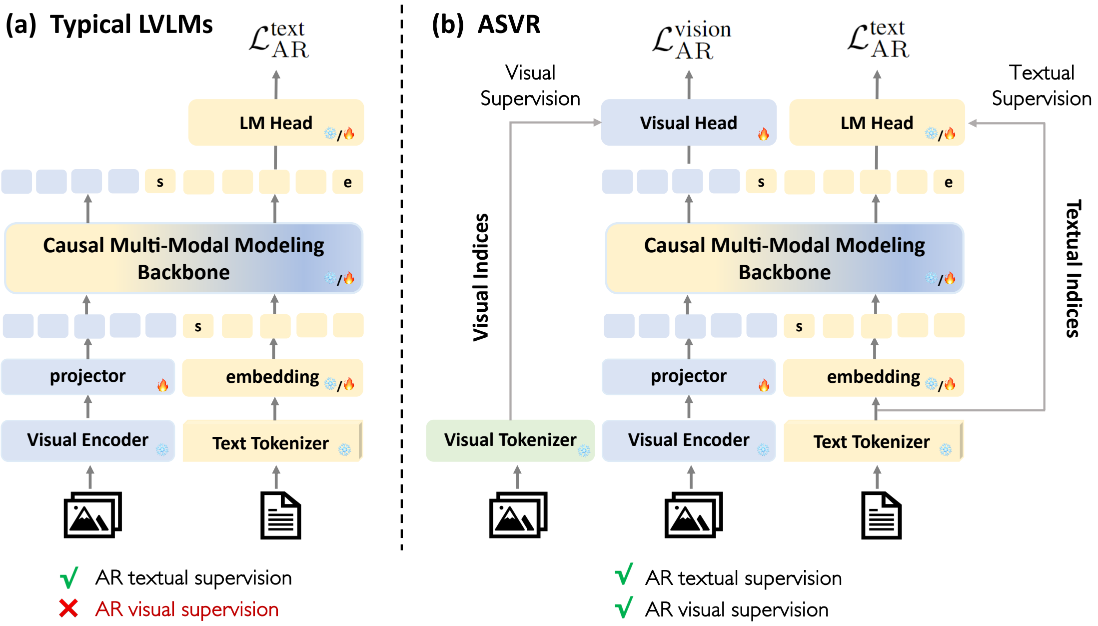

Figure 2: **Left**: The typical LVLM paradigm such as LLaVA (Liu et al., 2023b). **Right**: Overview of **ASVR's** architecture and training pipeline. The input image and text are tokenized into sequences of discrete token indices enabling unified autoregressive supervision over both visual inputs and textual outputs. For each module, the icon before the slash indicates whether it is frozen or tunable during pre-training, while the icon after the slash indicates its configuration during instruction tuning. "s" and "e" denote the start and end of the text tokens, respectively.

cross-modal alignment. The resulting sequence of token indices represents a quantized mapping of the image's high-level semantic features. Using semantic tokenizers to provide semantic visual supervision will guide LVLMs to focus on semantically meaningful structures and relationships of the image, thereby promoting more effective multimodal understanding.

3.3 Visual Encoder

The visual encoder provides continuous visual features as inputs to the LLM backbone, directly affecting the quality of visual modeling. To enhance multimodal understanding, it is crucial to employ a visual encoder that is semantically aligned with textual representations (Wu et al., 2024b; Qu et al., 2024; Wu et al., 2024a), allowing the extraction of high-level and semantically meaningful image features. Typically, such visual encoders adopt transformer-based (Dosovitskiy et al., 2021) architecture, trained with contrastive loss (Radford et al., 2021) to align closely with textual semantics and directly convert input images into one-dimensional sequences of continuous feature vectors.

3.4 Training Recipe

Our training recipe is illustrated in Figure 2, which extends the standard LVLM training by incorporating visual supervision to enable unified autoregressive modeling over both visual inputs and textual responses. Specifically, during the pre-training stage, we focus solely on optimizing the projector and the visual prediction head. This stage aligns visual representations sequence with the LVLM's semantic space, allowing the model to acquire an initial visual perception by learning the mapping from continuous visual features to discrete visual token indices. In the instruction tuning stage, we further fine-tune the LLM backbone parameters. By leveraging diverse vision-language instruction data, the model is encouraged to perform deeper semantic understanding of visual content, thereby enhancing its ability to understand and reason across modalities in a more comprehensive manner.

4 Experiments

In this section, we present a comprehensive set of controlled experiments to evaluate the effectiveness of our method (**ASVR**) within typical LVLM's frameworks (Liu et al., 2023b) across a diverse range of multimoda understanding tasks. We begin by detailing our experimental setup. Then, we analyze the impact of different visual encoders and visual tokenizers on the model's performance. Finally, we further validate the generalization and

Table 1: **Impact of ASVR with different combinations of visual encoders and visual tokenizers across multimoda understanding benchmarks.** "✗" indicates training with textual supervision only, while "✓" denotes the inclusion of visual supervision via an additional $\mathcal{L}_{\text{AR}}^{\text{vision}}$. "Sem." refers to using visual semantic tokenizer to construct visual supervision targets; "App." denotes a visual appearance tokenizer; "App.+Sem." indicates dual supervision, where visual semantic and visual appearance tokenizers are used independently to compute their respective $\mathcal{L}_{\text{AR}}^{\text{vision}}$, which are then summed. Our proposed ASVR utilize semantic supervision.

	$\mathcal{L}_{\text{AR}}^{\text{vision}}$	Visual Tokenizer	OCR				General				Knowledge		Visual-Centric		Hallusion		AVG
			TVQA	DVQA	OCRB	CQA	MMB	MME	SEED	GQA	MMMU	AI2D	RQA	MMVP	Hbench	POPE	
Visual Encoder: VQ-SigLIP-ViT-SO400M/14@384 (Discrete Visual Features)																	
LLaVA	✗	-	49.3	20.0	29.5	12.4	60.4	56.9	63.1	56.2	31.2	50.4	50.2	24.7	21.8	80.7	43.3
ASVR	✓	Sem.	55.5(+6.2)	21.4(+1.4)	32.4(+2.9)	14.7(+2.3)	62.3(+1.9)	57.7(+0.8)	65.4(+2.3)	57.1(+0.9)	32.0(+0.8)	53.5(+3.1)	52.3(+2.1)	26.0(+1.3)	27.7(+5.9)	76.8(-3.9)	45.3
Visual Encoder: SigLIP-ViT-SO400M/14@384 (Continuous Visual Features)																	
LLaVA	✗	-	56.0	21.1	31.3	14.6	64.0	67.2	63.8	60.5	32.7	53.5	52.0	28.7	23.9	85.9	46.8
Appearance Supervise	✓	App.	53.7(-2.3)	17.8(-3.3)	30.2(-1.1)	14.4(-0.2)	61.6(-2.4)	68.7(-1.5)	59.5(-4.3)	57.8(-2.7)	33.1(+0.4)	53.7(+0.2)	49.3(-2.7)	22.0(-6.7)	24.0(+0.1)	84.1(-1.8)	45.0
Dual Supervise	✓	App.+Sem.	59.4(+3.4)	23.7(+2.6)	33.5(+2.2)	16.1(+1.5)	65.6(+1.6)	70.2(+3.0)	66.1(+2.3)	61.5(+1.0)	**34.0**(+1.3)	56.3(+2.8)	53.5(+1.5)	22.0(-6.7)	30.7(+6.8)	86.3(+0.4)	48.5
ASVR	✓	Sem.	**59.5**(+3.5)	**24.3**(+3.2)	**35.4**(+4.1)	**16.4**(+1.8)	**66.1**(+2.1)	**72.8**(+5.6)	**66.4**(+2.6)	**61.5**(+1.0)	33.9(+1.2)	**57.0**(+3.5)	**54.1**(+2.1)	**30.0**(+1.3)	**33.7**(+9.8)	**86.3**(+0.4)	**49.8**

adaptability of our method across various LLM backbones with different parameter scales and under varying amounts of training data.

4.1 Experimental Setup

Implementation Details. We implement our experiments baseline on the LLaVA-1.5 (Liu et al., 2023a) with only textual supervision as discussed in sec 2. We utilize Vicuna-v1.5-7B (Zheng et al., 2023) as the LLM backbone and initialize visual encoder with the pretrained weights from SigLIP-SO400M-patch14-384 (Alabdulmohsin et al., 2023) to support continuous visual features for LVLMs. For visual tokenizer, we employ both visual appearance tokenizer and visual semantic tokenizer(VQ-SigLIP)proposed in DualToken (Song et al., 2025) to construct visual supervision targets, which convert input images into $27 \times 27 \times 8$ visual semantic or appearance token sequences, with a residual depth of $D = 8$. The visual head also derived from DualToken, is integrated and aligned with the chosen visual tokenizer to ensure architectural compatibility. Training is conducted on LLaVA-558K (Liu et al., 2023b) for pretraining and LLaVA-1.5-665K (Liu et al., 2023b) for instruction tuning.

Evaluation Details We conduct a comprehensive evaluation of model's capabilities on 14 widely used vision-language understanding benchmarks. Specifically, the general multimodal benchmarks include MMBench (Liu et al., 2024d) English dev split(MMB), GQA (Hudson and Manning, 2019b), SEED-Image(SEED) (Li et al., 2023a) and MME sum (Fu et al., 2024b). For OCR-based question answering, we assessed performance on TextVQA(TVQA) (Singh et al., 2019), ChartQA(CQA) (Masry et al., 2022), DocVQA(DVQA) (Mathew et al., 2021) and OCRBench(OCRB) (Liu et al., 2024e) . For knowledge-based question answering, we utilize MMMU validation split (Yue et al., 2024), AI2D (Kembhavi et al., 2016). Additionally, we evaluated hallucination robustness on POPE (Li et al., 2023c), Hallusionbench(Hbench) (Guan et al., 2024) and visual-centric tasks on MMVP (Tong et al., 2024c) and RealworldQA(RQA) (xAI, 2024). Evaluation prompts can be found in Appendix A.3.

4.2 Main Results

The Effectiveness of ASVR As shown in Table 1, with the configuration of the continuous-based visual encoder (SigLIP), we observe ASVR consistent and significant performance improvements across all 14 benchmarks, increasing the average score from **46.8** to **49.8**, with 3%. Notably, the gains are evident even on knowledge-based QA such as MMMU (Yue et al., 2024) and AI2D (Kembhavi et al., 2016), suggesting that reconstructing and and perceiving visual inputs can enhance the model's cognitive reasoning abilities. Furthermore, substantial improvements are also observed on fine-grained tasks such as OCRBench (Liu et al., 2024e), MMVP (Tong et al., 2024c), and HallusionBench (Guan et al., 2024). In particular, HallusionBench sees an increase of nearly 10 points, further validating the effectiveness of our method. Moreover, under the configuration with a discrete-based visual encoder(VQ-SigLIP), semantic visual supervision also yields notable performance gains over the baseline. This further demonstrates the

Table 2: **Generalizability of ASVR to different training data scales and LLM backbones across benchmarks.** The same visual encoder (SigLIP-ViT-SO400M/14@384) is used for ASVR and the baseline. "/" separates data scales for pre-training (left), mid-training (middle, if applicable) and instruction tuning (right).

	$\mathcal{L}_{\mathrm{AR}}^{\mathrm{vision}}$	LLM backbone	Data Scale	OCR				General				Knowledge		Visual-centric		Hallusion		AVG
				TVQA	DVQA	OCRB	CQA	MMB	MME	SEED	GQA	MMMU	AI2D	RQA	MMVP	Hbench	POPE	
With Different Data Scales																		
Baseline	✗	Vicuna-v1.5-7B	2M/2M	**61.6**	**43.8**	35.4	38.7	68.4	74.9	67.9	61.7	40.6	64.6	**56.1**	34.8	36.9	85.6	55.1
ASVR	✓	Vicuna-v1.5-7B	2M/2M	60.6(-1.0)	43.1(-0.7)	**36.2**(+0.8)	**38.9**(+0.2)	**68.6**(+0.2)	**76.2**(+1.3)	**68.7**(+0.8)	**62.0**(+0.3)	**41.4**(+0.8)	**64.8**(+0.2)	55.9(-0.2)	**35.9**(+1.1)	**42.2**(+5.3)	**85.7**(+0.1)	**55.7**
Baseline	✗	Vicuna-v1.5-7B	558K/4M/4M	57.2	44.1	49.6	39.2	71.7	71.7	68.7	58.2	37.9	70.7	56.7	40.0	36.1	85.5	56.2
ASVR	✓	Vicuna-v1.5-7B	558K/4M/4M	**60.0**(+2.8)	**46.5**(+2.4)	**51.3**(+1.7)	**41.7**(+2.5)	**72.2**(+0.5)	**73.2**(+1.5)	**69.9**(+1.2)	**59.8**(+1.6)	**39.7**(+1.8)	**71.8**(+1.1)	**57.5**(+0.8)	**42.0**(+2.0)	**37.9**(+1.8)	**86.9**(+1.4)	**57.9**
With Different LLM Backbones																		
Baseline	✗	Mistral-7B	558K/665K	50.8	15.7	**34.6**	15.2	65.9	66.9	67.9	62.4	32.0	53.0	55.0	35.3	32.7	86.6	48.1
ASVR	✓	Mistral-7B	558K/665k	**54.9**(+4.1)	**17.9**(+2.2)	34.1(-0.5)	**15.6**(+0.4)	**67.1**(+1.2)	**71.5**(+4.6)	**68.3**(+0.4)	**62.5**(+0.1)	**32.6**(+0.6)	**54.5**(+1.5)	**55.4**(+0.4)	**35.7**(+0.4)	**35.0**(+2.3)	**86.8**(+0.2)	**49.4**
Baseline	✗	Vicuna-v1.5-13B	558K/665k	57.2	22.1	32.4	15.1	67.1	68.9	65.6	60.4	35.6	54.9	54.8	34.0	32.9	86.8	49.1
ASVR	✓	Vicuna-v1.5-13B	558k/665K	**61.6**(+4.4)	**27.3**(+5.2)	**37.1**(+4.7)	**18.4**(+3.3)	**70.8**(+3.7)	**74.9**(+6.0)	**68.7**(+3.1)	**62.8**(+2.4)	**36.4**(+0.8)	**60.0**(+5.1)	**56.0**(+1.2)	**35.3**(+1.3)	**36.8**(+3.9)	**87.5**(+0.7)	**52.4**

generalizability and robustness of our method.

Semantic v.s. Appearance Specifically, ASVR incorporating semantic supervision alone yields the highest average performance across benchmarks, outperforming even the dual supervision setting that combines both appearance and semantic visual indices. In contrast, applying appearance-only supervision degrades model performance compared to the baseline. These results highlight that guiding the LVLM to reconstruct and perceive high-level semantic visual information of the input image, rather than low-level appearance details, more effectively enhances its multimoda understanding capabilities.

Continuous vs. Discrete We adopt SigLIP-ViT-SO400M/14@384 (Zhai et al., 2023) to provide continuous visual features, while employing visual semantic tokenizer VQ-SigLIP (Song et al., 2025) to generate discrete visual features; both approaches aligned with textual semantics. Our experimental results indicate that, regardless of whether autoregressive semantic visual supervision is applied, the configuration of using continuous visual features consistently outperforms its discrete features counterpart arcoss all benchmarks. This performance gap may be attributed to image feature degradation introduced by vector quantization in discrete encoding, which can lead to loss of fine-grained visual information crucial for downstream multimoda understanding. More ablations on semantic tokenizers, training strategies and visual supervision are provided in Appendix A.5.

Discussion The combination of visual encoder for provide visual features and visual semantic tokenizer for constructing semantic visual supervision targets proves to the most effective model configuration. The visual encoder avoids the visual information loss typically introduced by vector quantization, thereby providing better visual inputs for the LMM. Meanwhile, semantic supervision guides the LVLM reconstruct high-level, semantically meaningful aspects of the image, which are benefit for multimoda understanding.Notably, our findings demonstrate that continuous visual inputs with discrete semantic visual supervision targets can be seamlessly integrated into the unified autoregressive next-token prediction paradigm in the same manner as language. This formulation enables the LVLM to reconstruct and perceive visual semantic information, enhancing LVLM's capacity for comprehensive multimoda understanding, We also observe that using stronger visual supervision, such as SigLIP-2 (Tschannen et al., 2025), leads to larger performance gains. Moreover, ASVR benefits from having the input and supervision in the same feature space, outperforming setups where they come from different feature spaces in Appendix A.5. We further demonstrate that the unified autoregressive modeling paradigm consistently surpasses its denoising-based counterpart (Wang et al., 2024b), with results provided in the Section 4.5.

4.3 Method Generalizability

We validate the generalization and robustness of ASVR under different data scales and LLM backbone configurations, as summarized in Table 2.

The Impact of Data Scaling To investigate the effectiveness of ASVR under varying training data scales, we follow Bunny (Bunny-pretrain-LAION-2M (He et al., 2024) for pre-training and Bunny-v1_1-data-2M (He et al., 2024) for instruction tun-

Table 3: **High resolution adaptation of ASVR across multimoda understanding benchmarks.** We follow LLaVA-Next (Liu et al., 2024c) that utilize the visual encoder(SigLIP-ViT-SO400M/14@384) and high resolution input (1152 × 1152) for ASVR and baseline.

	$\mathcal{L}_{\text{AR}}^{\text{vision}}$	LLM backbone	Data Scale	OCR				General				Knowledge		Visual-centric		Hallusion		AVG
				TVQA	DVQA	OCRB	CQA	MMB	MME	SEED	GQA	MMMU	AI2D	RQA	MMVP	Hbench	POPE	
LLaVA	✗	Vicuna-v1.5-7B	558K/779k	58.1	44.1	39.5	47.5	66.6	74.1	66.8	62.0	35.8	62.8	**57.8**	30.0	40.6	84.5	55.0
ASVR	✓	Vicuna-v1.5-7B	558K/779k	**58.9**(+0.8)	**48.9**(+4.8)	**45.6**(+6.1)	**49.3**(+1.8)	**68.0**(+1.4)	**76.7**(+2.6)	**67.2**(+0.4)	**62.4**(+0.4)	**36.9**(+1.1)	**65.4**(+2.6)	57.6(-0.2)	**31.9**(+1.9)	**43.7**(+3.1)	**86.5**(+2.0)	**57.1**

ing) and LLaVA-OV (Li et al., 2024) (558K for pretraining, 4M for midtraining and 4M for instruction tuning follwing LLaVA-OV (Li et al., 2024) training recipe). As shown in Table 1 and Table 2, ASVR consistently yields substantial improvements over the baseline across different training data scales, demonstrating its ability to effectively leverage additional data through autoregressive semantic visual reconstruction.

The Impact of LLM Backbone Capacities We further evaluate the generalizability of ASVR across different LLM backbones. Specifically, we extend our experiments to Mistral-7B(Jiang et al., 2023) and Vicuna-v1.5-13B, which differ from Vicuna-v1.5-7B in model family and scale (Zheng et al., 2023). As shown in Table 2, ASVR consistently surpasses the baseline across multimodal understanding benchmarks, maintaining strong performance advantages regardless of backbone variations. These results demonstrating both its robustness and adaptability in diverse LLM configurations. The backbone scaling experiment and clear scaling law table will provide in Appendix A.4.

4.4 High-resolution Adaptation

ASVR is also compatible with existing high-resolution strategies and can further enhance the multimodal understanding capabilities of LVLMs. To evaluate the effectiveness of ASVR under high-resolution configurations, we upscale the input resolution of both ASVR and the baseline models to 1152 × 1152, while keeping the training conditions identical. We use LLaVA-558K(Liu et al., 2023b) for the pre-training stage and LLaVA-Next-779K(Liu et al., 2024c) for instruction tuning following LLaVA-Next settings (Liu et al., 2024c). As shown in Table 3, under high-resolution configurations, ASVR consistently outperforms the baseline by 2% in average scores across 14 multimodal benchmarks, further demonstrating its flexibility and robustness across different input resolutions.

4.5 Comparison with ROSS

ROSS (Wang et al., 2024a) reconstructs continuous, appearance-level visual features (VAE features) through denoising, whereas our ASVR reconstructs discrete, semantic-level visual indices (such as discretized SigLIP features) via autoregression. We conduct experiments using the LLaVA-Next dataset (Liu et al., 2024c) under identical training settings, clearly demonstrating that the ASVR-trained model consistently outperforms the ROSS ablation variants across multiple multimodal evaluation metrics. We also implement an additional variant of ROSS that reconstructs continuous semantic-level features (SigLIP features) through denoising. The result is shown in the table below shown in Table 4.

Our ASVR-trained model achieves the best performance, indicating that autoregressive semantic visual reconstruction (ASVR) is superior to both denoising semantic visual reconstruction ablation variants and even denoising appearance visual reconstruction (ROSS) ablation variants. We attribute this performance gap to a fundamental alignment principle: LLMs are inherently trained to model high-level semantic information. Therefore, when the visual supervision is semantically aligned with textual inputs—as in ASVR—it naturally leads to better integration and understanding. In contrast, reconstructing low-level visual features (as in appearance-based ROSS) lacks semantic alignment and can even hinder comprehension. Since tasks such as VQA rely heavily on semantic reasoning, reconstructing semantic visual information is more effective for enhancing multimodal understanding.

4.6 Qualitative Comparison

We visualize attention-score maps from several cases, illustrating the attention distribution of the last token with respect to all visual tokens, as shown in Figure 3. Compared to the baseline (LLaVA), our ASVR method consistently demonstrates more precise focus on image regions relevant to the given textual query. This highlights that incorporating

Table 4: **The detailed comparision between ASVR and ROSS ablation variant. ASVR achieves the best performance under identical training conditions.** ROSS models visual information through a denoising approach, whereas ASVR adopts unified autoregressive paradigm. The SigLIP-ViT-SO400M/14@384 is utilized for semantic visual supervision and VAE features is appearance visual supervision.

Method	Visual Supervision	**LLM backbone**	**Visual Modeling**	**Data**	**TVQA**	**DVQA**	**OCRB**	**CQA**	**MMB**	**MME**	**SEED**	**GQA**	**MMMU**	**AI2D**	**RQA**	**MMVP**	**Hbench**	**POPE**	**AVG**
ROSS	Apperance VAE	Vicuna-v1.5–7B	Denoising	LLaVA–Next	56.3	39.6	35.9	41.0	65.6	71.7	65.9	61.6	34.4	65.5	55.0	33.3	28.9	85.9	52.9
ROSS	Semantic Siglip	Vicuna-v1.5–7B	Denoising	LLaVA–Next	57.5	40.2	37.4	42.5	67.0	70.5	66.2	62.1	**34.9**	64.6	**55.7**	30.1	31.2	85.8	53.3
ASVR	Semantic Siglip	Vicuna-v1.5–7B	Autoregressive	LLaVA–Next	**58.6**	**40.6**	**39.7**	**43.4**	**67.9**	**73.0**	**67.5**	**62.9**	34.2	**65.8**	55.4	**36.8**	**39.2**	**85.9**	**55.1**

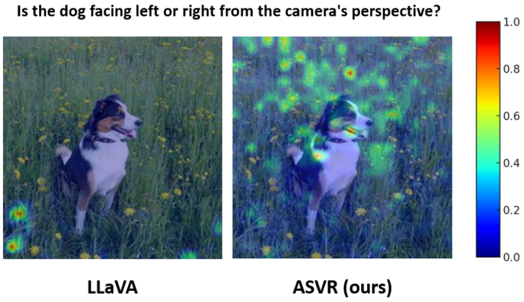

Figure 3: We further provide qualitative comparisons of attention maps under a controlled setting where the LLM backbone and training data remain unchanged. ASVR consistently encourages the model to allocate higher attention weights to image regions that are directly relevant to the question.

semantic visual supervision via the autoregressive semantic visual reconstruction objective $\mathcal{L}_{\text{AR}}^{\text{vision}}$ effectively enhance its ability to accurately associate textual descriptions with corresponding visual elements. More comparison on attention maps are shown in Appendix A.2.

5 Related Work

Large Vision Language Models The rapid progress in large language models (LLMs)(Bai et al., 2023a; AI@Meta, 2024; Touvron et al., 2023; Bi et al., 2024; OpenAI, 2023b,a; Chen et al., 2025) has showcased their strong generalization and remarkable instruction-following capabilities. To further expand these strengths for interpreting and interacting with the world through both visual and linguistic channels. There has been growing interest in Large Vision-Language Models (LVLMs)(Liu et al., 2023b,a, 2024c; Dong et al., 2025a), typically trained using a straightforward two-stage visual instruction tuning paradigm (Liu et al., 2023b), and align visual features extracted by visual encoder with the knowledge and reasoning capabilities of LLMs through the lightweight projector. This process involves jointly training the projector and the LLM on visual instruction datasets, with optional fine-tuning of the visual encoder. However, supervision is limited to text outputs. ASVR introduces a novel autoregressive visual semantic supervision mechanism that encourages the LVLM to reconstruct semantic visual tokens, enhancing its multimodal understanding capabilities.

Visual Autoregression for LVLMs Recent approaches (Team, 2024; Qu et al., 2024; Wang et al., 2024e; Wu et al., 2024b,a) introduce autoregressive visual supervision via visual tokenizers, such as VQGAN (Esser et al., 2021) and VQ-VAE (van den Oord et al., 2018), enabling LVLMs to support both multimodal understanding and image generation by predict relevant next visual tokens, which are then decoded into images. In contrast, ASVR focuses specifically on enhancing the multimodal understanding capability of LVLMs. Rather than generating images, ASVR employs autoregressive visual supervision to reconstruct semantic visual tokens within the given continuous image features as input. While prior methods are generative, ASVR adopts the reconstructive approach aimed at promoting perception of visual information.

Reconstructive Objectives for LVLMs ROSS(Wang et al., 2024b) introduces visual supervision for LVLMs by applying denoising objective to reconstruct continuous, appearance-level visual features (VAE features). In contrast, ASVR proposes a unified approach by employing autoregressive objective—analogous to that used for text—to reconstruct semantic visual tokens. This design enables seamless integration of visual and textual information under a unified next-token prediction paradigm.

6 Conclusion

In summary, we introduced **Autoregressive Semantic Visual Reconstruction (ASVR)**, enabling joint learning of visual and textual modalities within a unified autoregressive framework and effectively improving multimodal understanding capability of LVLMs. ASVR explicitly integrates semantic visual supervision on visual inputs to foster deep perception. Our findings indicate that autoregressively reconstructing semantic visual representations of images consistently enhances performance across diverse multimodal tasks and also outperform its denoising-based counterpart. This effectiveness is robust across different visual feature types, LLM backbone capacities, data scales, and high-resolution scenarios, underscoring ASVR's adaptability, scalability and versatility.

Acknowledgements

The research is supported by the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (Grant No.2025-GZL-RGZN-BTBX-02028).The project's computational resources are partially supported by CFFF platform of Fudan University and Shanghai Innovation Institute.

Limitations

Due to computational resource constraints, the current implementation of ASVR does not incorporate video understanding capabilities or dynamic resolution functionalities. However, these features are readily transferable and can be integrated in future iterations. In subsequent work, we plan to scale the model to support these functionalities, while also expanding its capabilities to include additional modalities. Furthermore, we aim to extend ASVR's capacity to enable multimodal generation, thereby broadening its applicability in various domains.

References

AI@Meta. 2024. Llama 3 model card.

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Kedi Chen, Dezhao Ruan, Yuhao Dan, Yaoting Wang, Siyu Yan, Xuecheng Wu, Yinqi Zhang, Qin Chen, Jie Zhou, Liang He, Biqing Qi, Linyang Li, Qipeng Guo, Xiaomin Shi, and Wei Zhang. 2025. A survey of inductive reasoning for large language models. *ArXiv*, abs/2510.10182.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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

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

Shuai Dong, Siyuan Wang, Xingyu Liu, and Zhongyu Wei. 2025a. Interleaved latent visual reasoning with selective perceptual modeling. *ArXiv*, abs/2512.05665.

Shuai Dong, Jie Zhang, Guoying Zhao, Shiguang Shan, and Xilin Chen. 2025b. Semantic mismatch and perceptual degradation: A new perspective on image editing immunity. *ArXiv*, abs/2512.14320.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, and 5 others. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Preprint*, arXiv:2404.06512.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024b. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *Preprint*, arXiv:2402.11530.

Drew A Hudson and Christopher D Manning. 2019a. 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.

Drew A. Hudson and Christopher D. Manning. 2019b. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Chenglin Li, Feng Han, Yikun Wang, Ruilin Li, Shuai Dong, Haowen Hou, Haitao Li, Qianglong Chen, Feng Tao, Jingqi Tong, and 1 others. 2025. Videopro: Adaptive program reasoning for long video understanding. *arXiv preprint arXiv:2509.17743*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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. 2024d. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024e. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

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. *Preprint*, arXiv:2403.05525.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *Preprint*, arXiv:2203.10244.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

OpenAI. 2023a. Chatgpt (august 3 version).

OpenAI. 2023b. Gpt-4 technical report. *arXiv:2303.08774*.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. 2024. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. DINOv3. *Preprint*, arXiv:2508.10104.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

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*.

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

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2024b. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024c. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural discrete representation learning. *Preprint*, arXiv:1711.00937.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. 2024a. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. 2024b. Reconstructive visual instruction tuning. *Preprint*, arXiv:2410.09575.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024d. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024e. 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 1 others. 2024a. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, and 1 others. 2024b. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024c. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

xAI. 2024. Grok. `https://x.ai`. Developed by xAI.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. 2024. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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. *Preprint*, arXiv:2311.16502.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Jie Zhang, Shuai Dong, Shiguang Shan, and Xilin Chen. 2025. Towards transferable defense against malicious image edits. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Use of LLMs in in Paper Writing

In preparing this paper, Large Language Models (LLMs) were employed to support the refinement of writing. Their role was limited to enhancing the linguistic presentation of the paper by improving readability, clarity, and stylistic consistency. Specifically, the models were used for tasks such as rephrasing sentences, checking grammar, and streamlining the flow of the text. We emphasize that the LLMs were not involved in generating research ideas, designing methodologies, or conducting experiments. All conceptual development, methodological design, and analytical work were carried out solely by the authors. The contribution of the LLMs was restricted to language-level improvements and did not extend to the scientific substance of the work. The authors retain complete responsibility for the content of this paper, including passages revised with LLM assistance. Care has been taken to ensure that the use of LLMs complies with ethical standards and does not give rise to plagiarism or any form of scientific misconduct.

A.2 Qualitative Comparison

We visualize attention-score maps from several cases, illustrating the attention distribution of the last token with respect to all visual tokens, as shown in Figure 4. Compared to the baseline (LLaVA), our ASVR method consistently demonstrates more precise focus on image regions relevant to the given textual query. This highlights that incorporating semantic visual supervision via the autoregressive semantic visual reconstruction objective $\mathcal{L}_{\text{AR}}^{\text{vision}}$ effectively enhance its ability to accurately associate textual descriptions with corresponding visual elements.

A.3 Evaluation Prompts

All prompts used for evaluation benchmarks are released and summarized in Table5 following

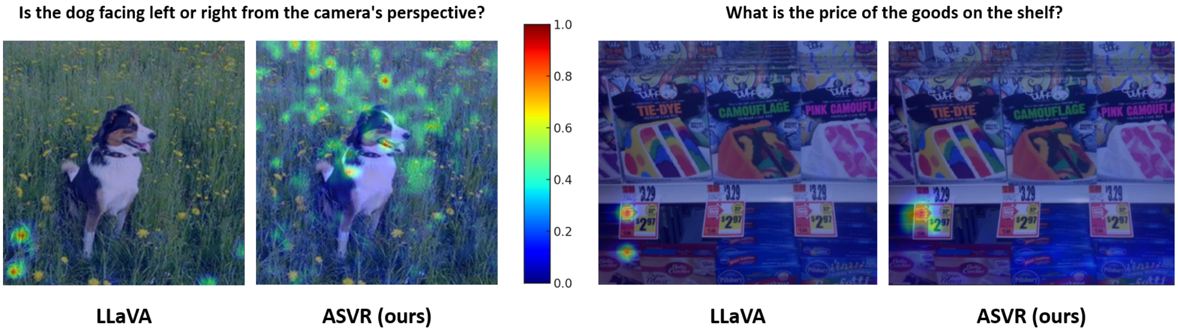

Figure 4: We further provide qualitative comparisons of attention maps under a controlled setting where the LLM backbone and training data remain unchanged. ASVR consistently encourages the model to allocate higher attention weights to image regions that are directly relevant to the question.

Cambrian-1 (Tong et al., 2024a).

A.4 The Scalability of ASVR

we show the clear scaling study along two axes in Table 6 and Table 7:

Data scaling We train on four datasets—LLaVA-1.5-665K (Liu et al., 2023b), LLaVA-Next-779K (Liu et al., 2024b), Bunny-2M (He et al., 2024), and LLaVA-OV-4M (Li et al., 2024) to isolate the effect of data volume.

Backbone scaling Using the same Vicuna family, we vary only the parameter count and scale up to a 13B model (the maximum allowed by our computational budget).

A.5 Ablation Study

The Impact of Semantic Tokenizer Increasing the degree of alignment with text for semantic tokenizer leads to performance of ASVR. We use VQ-SigLip (Song et al., 2025) trained on different data scales to construct semantic visual supervision targets: with 3M data trained varient, which achieves zero-shot ImageNet classification accuracy of 78.6% (Deng et al., 2009), and with 12M data trained varient, which achieves 81.6% and thus exhibits stronger semantic alignment. As shown in Table 8, ASVR equipped with the better-aligned with 12M data trained varient consistently outperforms the variant using 3M data trained across the majority of multimodal benchmarks, with the average performance improving by more than 2%. These results demonstrate that employing better semantically aligned visual tokenizer provides semantic visual supervision targets with more meaningful aspects of the image, and further support our claim that Semantic Visual Reconstruction plays a key role in enhancing the multimodal understanding capabilities of LVLMs. Moreover, when the supervised visual tokenizer provides richer semantic information, ASVR achieves stronger performance. We present the results obtained using discrete SigLIP2 (Tschannen et al., 2025) as visual tokenizer which contain richer semantic visual information in the Appendix A.5.

The Impact of Training Strategy We explore different training strategies for ASVR, comparing whether to apply semantic visual supervision in both the pre-training and instruction tuning stages, or to apply it only during instruction tuning, while keeping the pre-training stage purely with text-based autoregressive training. As shown in Table 8, incorporating semantic visual supervision to support visual autoregressive training in both the pre-training and instruction tuning stages consistently outperforms the single-stage variant across all benchmarks, achieving an average performance gain of nearly 6%. This further underscores the importance of Semantic Visual Reconstruction during the pre-training phase, as it enables the model to develop a more complete perception of visual information. By doing so, it enhances vision-language alignment and mitigates the information loss associated with relying solely on textual supervision.

The Impact of visual supervision We further extend our experiments by employing discrete VQ-SigLIP2 (Tschannen et al., 2025) as visual supervision, which provides richer and stronger semantic information, to verify that enhanced visual-semantic supervision can better scale the effectiveness of ASVR. To ensure fair comparison, we use LLaVA-Next (Liu et al., 2024c) as the training dataset under identical conditions, evaluating ASVR against the baseline with SigLIP (Zhai et al.,

Table 5: **Listing the prompts used in the evaluation of each benchmark.**

Benchmark	Prompt
TextVQA (Singh et al., 2019)	Answer the question using a single word or phrase.
DocVQA (Mathew et al., 2021)	Answer the question using a single word or phrase.
OCRBench (Liu et al., 2024e)	Give the short answer directly.
ChartQA (Masry et al., 2022)	Answer the question using a single number or phrase.
MMBench (Liu et al., 2024d)	Answer with the option's letter from the given choices directly.
MME (Fu et al., 2024b)	Answer the question using a single word or phrase.
SEED-Image (Li et al., 2023a)	Answer with the option's letter from the given choices directly.
GQA (Hudson and Manning, 2019b)	Answer the question using a single word or phrase.
MMMU (Yue et al., 2024)	Answer with the option's letter from the given choices directly.
AI2D (Kembhavi et al., 2016)	Answer with the option's letter from the given choices directly.
RealworldQA (xAI, 2024)	Please answer directly with only the letter of the correct option and nothing else.
MMVP (Tong et al., 2024c)	Answer with the option's letter from the given choices directly.
Hallusionbench (Guan et al., 2024)	Answer the question using a single word or phrase.
POPE (Li et al., 2023c)	Answer the question using a single word or phrase.

Table 6: **The scaling relationship between computational cost (FLOPs) and average performance score across different scale dataasets with the same LLM backbone-Vicuna-v1.5-7B.**

Data	FLOPs (×1e19)	Avg Score
LLaVA-1.5-665K (Liu et al., 2023b)	1.53	49.8
LLaVA-Next-779K (Liu et al., 2024b)	2.49	55.1
Bunny-2M (He et al., 2024)	4.52	55.7
LLaVA-OV-4M (Li et al., 2024)	7.51	57.9

Table 7: **The scaling relationship between computational cost (FLOPs) and average performance score across different scale backbone parameters with the same training dataset-LLaVA-1.5-665k.**

LLM Backbone	FLOPs (×1e19)	Avg Score
Vicuna-v1.5-7B	1.53	49.8
Vicuna-v1.5-13B	2.58	52.4

2023) as both visual input and supervision, as well as with SigLIP2 (Tschannen et al., 2025) serving the same roles.

The results shown in Table 9 clearly demonstrate that stronger visual semantic encoders lead to better performance when used for supervision. Specifically, ASVR with SigLIP-2 outperforms the baseline (LLaVA) with SigLIP-2 by an average of +2.2 points across 14 benchmarks. In comparison, ASVR with SigLIP improves over its baseline by +1.3 points. These results indicate that ASVR benefits more from stronger semantic supervision, and that pairing ASVR with more powerful semantic vision supervision further enhances its ability to improve visual understanding.

More importantly, we conducted an additional ablation to directly the impact of having the input and supervision in different feature spaces. We kept SigLIP as the visual encoder training on LLaVA-Next (Liu et al., 2024c) but replaced the supervision tokenizer with tokenizers derived from different feature spaces (e.g., DINO-v3 (Siméoni et al., 2025) and DepthAnything (Yang et al., 2024)). Our results shown in table 10, demonstrate that using mismatched feature spaces does not produce larger gains, and in some benchmarks it even degrades performance. This suggests that ASVR learning a grounding-like alignment within the same semantic feature space: the LLM learns to map similar-but-not-identical continuous visual representations to shared discrete semantic identifiers (i.e., clustering/grounding in the tokenizer space).

Table 8: **Ablation study for various ASVR configurations.** This table presents a comparison of various ASVR settings, including semantic tokenizer, varied the degree of alignment with text, and the training strategy, where "PT/IT" denotes that semantic visual supervision is applied during both the pre-training and instruction tuning stages, while "IT" indicates that semantic visual supervision is applied only during instruction tuning.

Ablated Aspects	Original	Ablated Setting	OCR				General				Knowledge		Visual-centric		Hallusion		AVG
			TVQA	DVQA	OCRB	CQA	MMB	MME	SEED	GQA	MMMU	AI2D	RQA	MMVP	HBench	POPE	
Semantic Tokenizer	12M	3M	57.8(-1.7)	**25.4**(+1.1)	33.1(-2.3)	16.2(-0.2)	**67.2**(+1.1)	70.3(-2.5)	64.8(-1.6)	60.0(-1.5)	31.8(-2.1)	55.9(-1.1)	**54.3**(+0.2)	24.7(-5.3)	33.0(-0.7)	86.1(-0.2)	48.6
Training Strategy	PT/IT	IT	55.3(-4.2)	18.9(-5.4)	29.5(-5.9)	14.0(-2.4)	61.2(-4.9)	67.8(-5.0)	60.5(-5.9)	58.3(-3.2)	33.4(-0.5)	52.6(-4.4)	52.3(-1.8)	20.8(-9.2)	30.0(-3.7)	84.9(-1.4)	45.7
ASVR	-	-	**59.5**	24.3	**35.4**	**16.4**	66.1	**72.8**	**66.4**	**61.5**	**33.9**	**57.0**	54.1	**30.0**	**33.7**	**86.3**	**49.8**

Table 9: **Extend experiments on LLaVA-Next dataset, LLaVA indicates the baseline (typically LVLM framework), ASVR builds upon the baseline by introducing autoregressive semantic visual supervision.** "✗" indicates the use of textual supervision only. Visual encoder(SigLIP-ViT-SO400M/14@384 and SigLIP2-ViT-SO400M/14@384) are both utilized for ASVR and baseline to get different visual input and visual supervision.

	Visual Encoder	Visual Supervision	LLM Backbone	Data	TVQA	DVQA	OCRB	CQA	MMB	MME	SEED	GQA	MMMU	AI2D	RQA	MMVP	Hbench	POPE	AVG
LLaVA	Siglip-so400m-384	✗	Vicuna-v1.5–7B	LLaVA–Next	57.7	40.7	37.9	42.6	67.4	71.5	67.2	61.8	34.3	65.3	54.6	32.8	33.1	86.4	53.8
ASVR	Siglip-so400m-384	Semantic Siglip	Vicuna-v1.5–7B	LLaVA–Next	58.6	40.6	39.7	43.4	67.9	73.0	67.5	62.9	34.2	65.8	55.4	36.8	**39.2**	85.9	55.1
LLaVA	Siglip2-so400m-384	✗	Vicuna-v1.5–7B	LLaVA–Next	59.2	41.8	40.5	46.3	66.9	74.0	68.3	62.7	34.7	66.4	**56.9**	33.3	35.1	86.1	55.2
ASVR	Siglip2-so400m-384	Semantic Siglip-2	Vicuna-v1.5–7B	LLaVA–Next	**61.0**	**43.7**	**44.8**	**49.9**	**70.2**	**76.8**	**69.5**	**63.4**	**36.3**	**67.3**	56.7	**42.0**	35.8	**86.8**	**57.4**

Table 10: **Extend experiments on LLaVA-Next dataset, ASVR utlizes VQ-SigLIP as visual supervision, which shares the same feature space as the visual encoder (SigLIP-So400m-384). We further extend this by introducing VQ-DINOv3 and VQ-DepthAnything as alternative visual supervision signals.**

Visual Encoder	Visual Supervision	TVQA	DVQA	OCRB	CQA	MMB	MME	SEED	GQA	MMMU	AI2D	RQA	MMVP	Hbench	POPE	AVG
Siglip-so400m-384	✗	58.1	44.1	39.5	47.5	66.6	74.1	66.8	62.0	35.8	62.8	57.8	30.0	40.6	84.5	55.0
Siglip-so400m-384	VQ-DINO-v3	58.0	46.2	45.0	43.0	66.2	75.2	**68.0**	61.8	36.0	65.2	57.5	31.8	37.9	**87.4**	55.6
Siglip-so400m-384	VQ-DepthAnything	52.8	40.7	38.6	42.2	67.2	75.9	67.2	62.2	32.0	59.8	**58.0**	30.3	40.7	84.6	53.7
Siglip-so400m-384	VQ-SigLIP	**58.9**	**48.9**	**45.6**	**49.3**	**68.0**	**76.7**	67.2	**62.4**	**36.9**	**65.4**	57.6	**31.9**	**43.7**	86.5	**57.1**