

FUSION: Fully Integration of Vision-Language Representations for Deep Cross-Modal Understanding

Zheng Liu^{1,2*}, Mengjie Liu^{1,2}, Jingzhou Chen^{1,2}, Jingwei Xu³, Bin Cui¹, Conghui He^{2†}, Wentao Zhang^{1†}

¹Peking University, ²Shanghai AI Laboratory, ³Nanjing University

Abstract

We introduce *FUSION*, a family of multimodal large language models (MLLMs) with a fully vision-language alignment and integration paradigm. Unlike existing methods that primarily rely on late-stage modality interaction during LLM decoding, our approach achieves deep, dynamic integration throughout the entire processing pipeline. To this end, we propose Text-Guided Unified Vision Encoding, incorporating textual information in vision encoding to achieve pixel-level integration. We further design Context-Aware Recursive Alignment Decoding that recursively aggregates visual features conditioned on textual context during decoding, enabling fine-grained, question-level semantic integration. To guide feature mapping and mitigate modality discrepancies, we develop Dual-Supervised Semantic Mapping Loss. Additionally, we construct a Synthesized Language-Driven Question-Answer (QA) dataset through a new data synthesis method, prioritizing high-quality QA pairs to optimize text-guided feature integration. Building on these foundations, we train *FUSION* at two scales—3B, 8B—and demonstrate that our full-modality integration approach significantly outperforms existing methods with only 630 vision tokens. Notably, *FUSION* 3B surpasses Cambrian-1 8B and Florence-VL 8B on most benchmarks. *FUSION* 3B continues to outperform Cambrian-1 8B even when limited to 300 vision tokens. Our ablation studies show that *FUSION* outperforms LLaVA-NeXT on over half of the benchmarks under same configuration without dynamic resolution, highlighting the effectiveness of our approach. We release our code, model weights, and dataset. <https://github.com/starriver030515/FUSION>

1. Introduction

Research has long shown that human visual perception is not a passive process like a camera capturing reality but rather an active, interpretative mechanism shaped by cognitive factors such as language and environmental context. A critical component of this mechanism is the dynamic in-

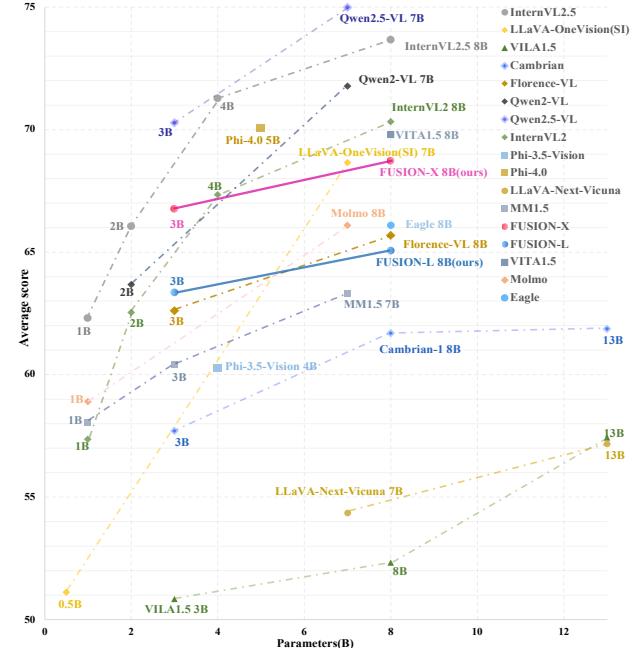


Figure 1. Performance comparison of *FUSION* with leading MLLM models across 18 benchmark dimensions. With only 630 vision tokens, our model (*FUSION*-X) significantly outperforms Cambrian-1 and Florence-VL, achieving overall parity with LLaVA-OneVision, while maintaining a minimal performance gap with top-tier models such as InternVL2 and Qwen2VL. Furthermore, even when the number of vision tokens is reduced to 300, our model (*FUSION*-L) preserves 95% of its original performance, remaining on par with Florence-VL.

teraction between vision and language. In [21], the authors demonstrated that hearing the name of a target object before searching for it significantly improved both the speed and accuracy of object detection. This finding underscores the deep integration of vision and language, suggesting that linguistic input facilitating more effective visual processing by directing attention and helping the brain prioritize relevant features.

Despite the fundamental role of vision-language interaction in human perception, current MLLMs often fail to capture this relationship effectively. In traditional MLLMs[8, 40], visual information is processed independently through

*Email: lz030515123@gmail.com

†Corresponding Author

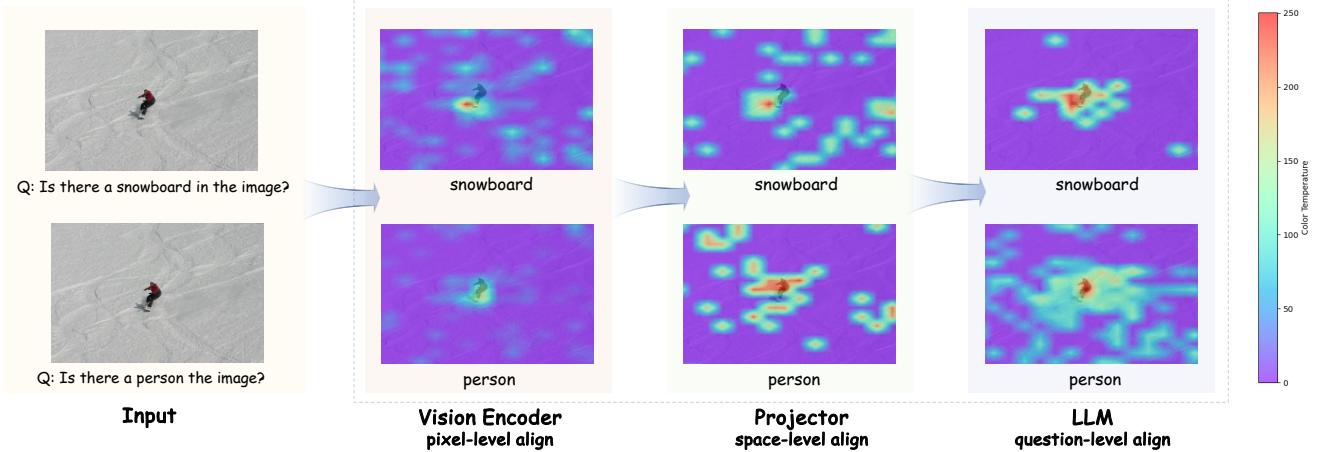


Figure 2. Visualization of modality alignment and integration. At pixel-level, we compute attention maps between image regions and question-relevant keywords within the vision encoder. At space-level, we measure the cosine similarity between vision tokens projected into the LLM embedding space and corresponding keywords. At question-level, we visualize attention maps from question keywords to vision tokens during LLM decoding. The results indicate that our model achieves consistent and progressively enhanced cross-modal alignment throughout the processing pipeline.

isolated vision encoders. These models typically treat vision features as static information, fusing them with language only at a later LLM decoding stage. This decoupled approach results in a lack of centralized encoding of vision features, leading to limited useful information available for textual interaction. Recent advancements[4, 11, 34, 39, 41, 57, 61] have sought to improve visual comprehension capability through techniques such as dynamic resolution encoding or multiple vision encoders, which enhances image details and provides richer representations. However, these improvements predominantly focus on enhancing visual encoding itself, overlooking the deeper, bidirectional integration necessary for effective vision-language fusion. Furthermore, relying solely on vision encoders or more image tokens introduces redundancy and inefficiencies, increasing computational costs while failing to establish a truly interactive fusion between modalities.

This challenge serves as the core motivation for our work: **achieving deep, adaptive, and dynamic interaction and integration between vision and language throughout the entire multimodal processing pipeline**. Previous studies have explored varied integration strategies, such as text-based query integration[13], text-oriented vision encoding[19], and entity-enhanced modality alignment[72]. Nevertheless, these models lack a systematic, comprehensive framework for aligning and integrating visual and textual modalities. A key unresolved issue is embedding misalignment, where inherent discrepancies between vision and text embeddings hinder seamless integration. Moreover, current models lack high-quality training data specifically designed to guide the alignment and integration pro-

cess. As a result, these methods fall short of achieving SOTA performance.

In this paper, we push the boundaries of vision-language fusion by proposing a sophisticated framework that realizes a deeper, more interactive integration of visual and textual modalities. Unlike prior approaches that treat vision and language as separate entities, our approach consistently integrates textual guidance throughout the visual encoding phase. We introduce supervised guidance in modality space mapping and dynamically aggregate vision features based on textual context during LLM decoding. Additionally, we introduce a new data synthesis method that generates QA pairs from text, guiding the model toward more efficient feature integration. Our method enables a more cognitively inspired multimodal learning paradigm. Overall, our contributions are as follows:

- **Text-Guided Unified Vision Encoding:** We project text embeddings into the visual feature space and jointly perform attention with visual representations to enhance pixel-level integration.
- **Context-Aware Recursive Alignment Decoding:** We introduce interaction layers that recursively aggregate visual features based on textual context during LLM decoding, enabling fine-grained, question-level alignment and integration.
- **Dual-Supervised Semantic Mapping Loss:** We propose a novel bidirectional reconstruction loss, ensuring robust consistency and mitigating semantic discrepancies between modalities.
- **Synthesized Language-Driven QA Dataset:** We de-

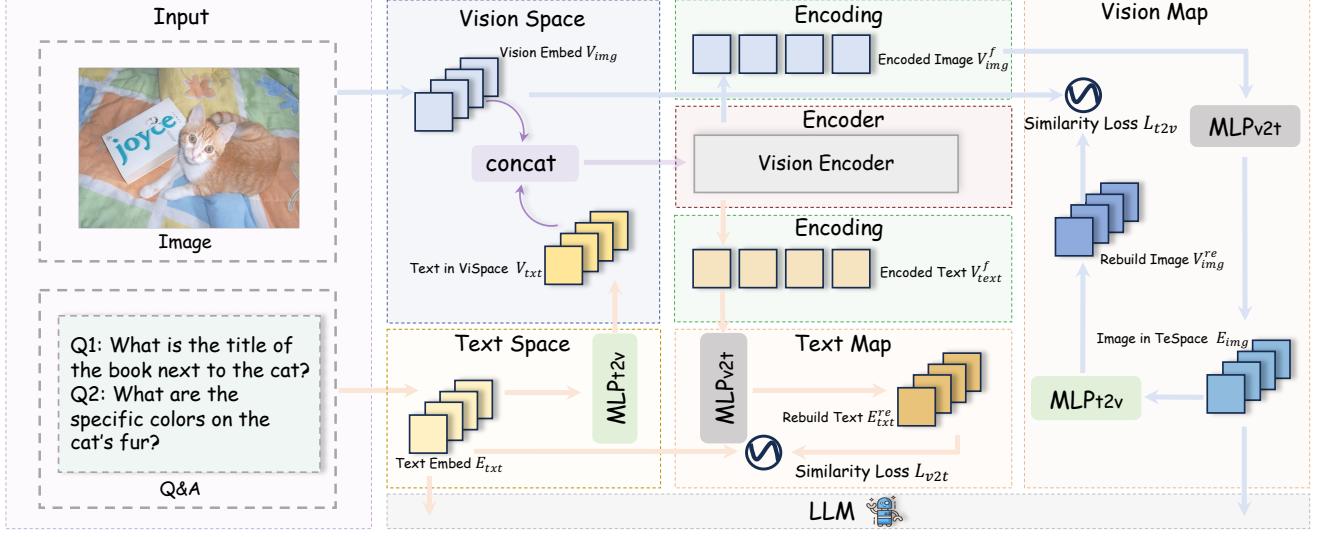


Figure 3. Illustration of our Text-Guided Unified Vision Encoding and Dual-Supervised Semantic Mapping Loss. Given an input image, the corresponding question is first projected into the vision feature space and processed jointly with the image. The extracted visual features are then mapped into the text space and fed into LLM. To ensure the reliability of the mapping MLP, we reconstruct the text and image tokens by reusing the encoded tokens and projecting them back into their original feature spaces, then compute the similarity between the reconstructed and raw tokens to encourage structural alignment between the two spaces.

velop a new method that prioritizes constructing high-quality QA pairs while leveraging generative models to produce images. We synthesize a large-scale, diverse QA dataset to optimize text-guided integration.

Building on our exploration, we present a new family of MLLMs that achieve leading performance across diverse benchmarks. To foster further research, we release our code, model weights, and data. We hope that our work will empower the research community and accelerate advancements in MLLMs.

2. Preliminaries and Related Work

Multimodal LLMs Multimodal Large Language Models (MLLMs) [8, 40, 41] are designed to process and integrate information from multiple modalities, primarily combining visual and textual data. Typically, these models are structured with three main components: a vision encoder, a projection module, and a large language model (LLM). Images are first divided into patches and processed through vision encoders to extract visual representations. These visual features are then mapped into textual embedding space via a projector before being combined with textual data during the language model’s decoding phase. Recent advances [4, 34, 36, 37, 54, 81] have explored dynamic or arbitrary resolution techniques, significantly enhancing the precision and adaptability of visual encoding. Meanwhile, alternative approaches [39, 57, 61] have employed multiple vision encoders to capture a more comprehensive set of visual

features. Despite these innovations, most current methodologies primarily emphasize visual encoding quality, often neglecting deeper and more interactive vision-language dynamics, thus hindering the full potential of multimodal learning.

Modality Alignment and Integration To enhance multimodal integration, some studies have investigated improved methods at various stages within MLLMs. For instance, [13] integrates linguistic guidance by embedding textual queries within the QFormer module, facilitating a language-driven visual feature projection. [72] enhances modality alignment by incorporating a contrastive loss in the projector module during pretraining to better associate visual features with entity representations. More recently, [7] introduced a generative vision encoder that leverages multiple textual prompts to integrate various feature types, achieving SOTA performance. But a fundamental challenge remains—the inherent difference between visual and textual feature spaces makes direct feature alignment difficult, often leading to inconsistencies in multimodal representations.

Instruction Tuning Data High-quality instruction tuning data is a crucial factor in the training and generalization capabilities of MLLMs. Existing works [34, 37, 57, 61] have constructed training recipes with visual question answering (VQA) benchmarks. Other efforts [6, 63, 68] lever-

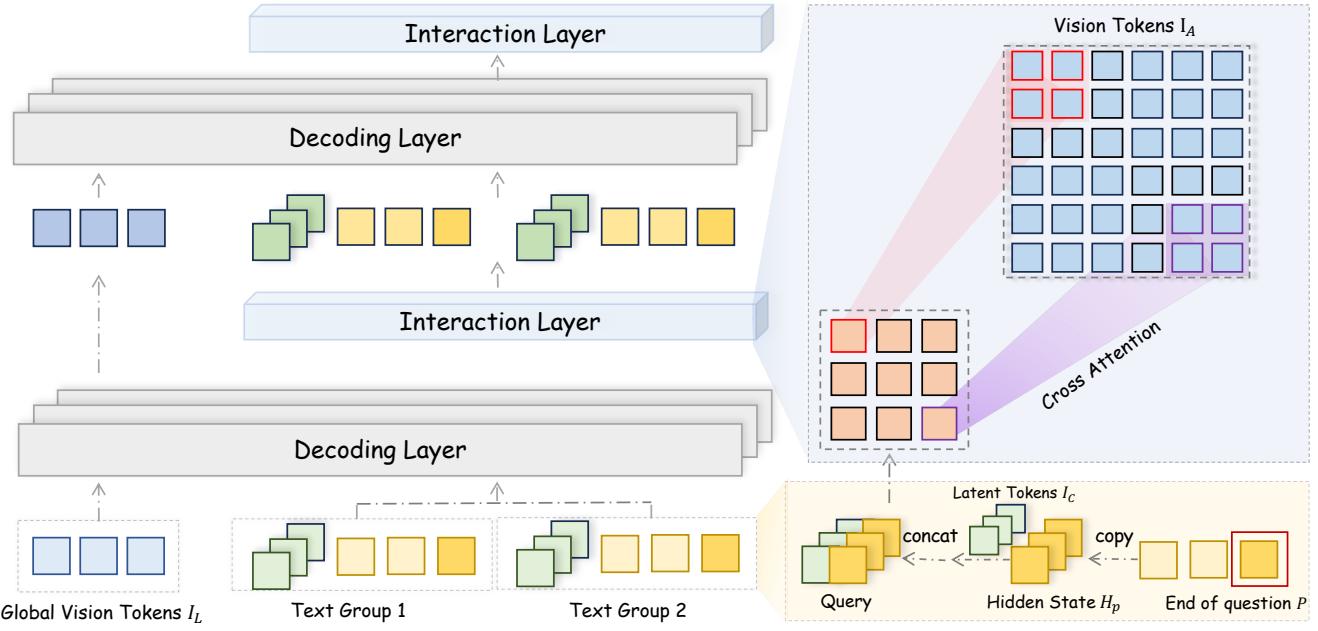


Figure 4. Illustration of our Context-Aware Recursive Alignment Decoding. For each set of question tokens (highlighted in yellow), we prepend a set of context-aware latent tokens (highlighted in green). Additional interaction layers are introduced between decoding layers, where vision tokens interact with both latent tokens and question tokens at a question-level granularity (e.g., Group 1, Group 2, ...).

aged GPT-4V to generate high-quality QA pairs from images. These datasets enhance multimodal understanding but are still predominantly vision-centered, where QA pairs are constructed primarily based on image content. Yet, the inherent limitations of image content restrict their capacity to support complex, instruction-following tasks across diverse domains. A promising alternative [45] attempts to address this limitation by synthesizing high-quality captions first and employing diffusion models to generate corresponding images. Despite its promise, this approach is still constrained to the pretraining stage, and the dataset size remains relatively small, limiting its effectiveness in large-scale multimodal training.

3. Method

In this section, we present our multimodal alignment and integration strategies as well as language-driven dataset construction method. In 3.1, we introduce Text-Guided Unified Vision Encoding for controlled vision-text integration. In 3.2, we propose Context-Aware Recursive Alignment Decoding for adaptive alignment. In 3.3, we define Dual-Supervised Semantic Mapping Loss for cross-modal consistency. In 3.4, we outline our Synthesized Language-Driven QA Dataset framework for high-quality data generation.

3.1. Text-Guided Unified Vision Encoding

Inspired by [19], we incorporate textual information into the vision encoder to enhance multimodal representation learning. However, we adopt a specialized fusion strategy to refine high-level visual features while preserving the integrity of low-level visual information.

As shown in Figure 3, given an image I and textual questions T , we first extract visual embeddings V_{img} and textual embeddings E_{txt} . Then we obtain Text in ViSpace V_{txt} by mapping the textual representation into vision feature space.

$$V_{img} = \text{VisionEmbed}(I) \quad (1)$$

$$E_{txt} = \text{LLMEmbed}(T), \quad V_{txt} = \text{MLP}_{t2v}(E_{txt}) \quad (2)$$

The mapped textual embeddings V_{txt} and visual embeddings V_{img} are then jointly processed through a stack of encoding layers, enabling mutual refinement of multimodal features. Formally, each encoding layer updates these embeddings as follows:

$$(V_{img}^i, V_{txt}^i) = \text{EncoderLayer}(V_{img}^{i-1}, V_{txt}^{i-1}), \quad i = 1, \dots, N \quad (3)$$

where N represents the total number of encoder layers. Initial embeddings are set as $V_{img}^0 = V_{img}$ and $V_{txt}^0 = V_{txt}$.

To ensure a balanced interaction between textual and visual features, we restrict textual influence at lower layers by masking the textual-to-visual attention in the first half of the encoder. After encoding, we aggregate visual and textual

embeddings separately into early and late-stage representations to capture both coarse and fine-grained multimodal features:

$$V_{img}^e = \frac{2}{N} \sum_{i=1}^{\frac{N}{2}} V_{img}^i, \quad V_{img}^l = \frac{2}{N} \sum_{i=\frac{N}{2}+1}^N V_{img}^i \quad (4)$$

$$V_{txt}^e = \frac{2}{N} \sum_{i=1}^{\frac{N}{2}} V_{txt}^i, \quad V_{txt}^l = \frac{2}{N} \sum_{i=\frac{N}{2}+1}^N V_{txt}^i \quad (5)$$

The final encoded visual and textual representations are obtained via channel-wise concatenation:

$$V_{img}^f = \text{Concat}(V_{img}^e, V_{img}^l; \text{dim}=\text{channel}) \quad (6)$$

$$V_{txt}^f = \text{Concat}(V_{txt}^e, V_{txt}^l; \text{dim}=\text{channel}) \quad (7)$$

Ultimately, the visual representation V_{img}^f is mapped into the LLM embedding space, yielding Image in TeSpace E_{img} , which facilitates direct interaction with textual features.

$$E_{img} = \text{MLP}_{v2t}(V_{img}^f) \quad (8)$$

3.2. Context-Aware Recursive Alignment Decoding

Unlike conventional approaches that position vision tokens solely at the beginning of the input sequence, we introduce additional context-aware latent tokens I_C that are specifically dedicated to each question. As shown in Figure 4, we introduce interaction layers between the LLM’s decoding layers. Within each interaction layer, latent tokens I_C recursively refine visual-textual alignment based on the evolving textual context, thus ensuring question-level, fine-grained multimodal interaction.

To enhance computational efficiency and enable effective fusion of multimodal features, we employ a localized windowed attention mechanism. Specifically, we define a grid of latent tokens $I_C \in \mathbb{R}^{n \times n \times D}$, where n denotes the spatial resolution of the token grid and D represents the embedding dimension. A predefined window size w is used to limit attention operations within manageable regions. To minimize biases from the text-guided encoding and to maintain the efficacy of the window attention, we construct an auxiliary visual representation I_A , which is achieved by partitioning the original image into four sub-images, encoding each without textual conditioning, and concatenating them. I_A is later utilized as keys and values. To ensure compatibility with both the predefined window size and latent tokens grid, we interpolate I_A to a resolution of $(n \cdot w, n \cdot w)$ and finally get $I_A \in \mathbb{R}^{(n \cdot w) \times (n \cdot w) \times D}$

Before decoding, we initialize I_C by interpolating I_A onto the (n, n) grid. During decoding, given a textual question T , we determine the position P corresponding to the end of this question within the token sequence. The hidden

state representation $H_P \in \mathbb{R}^{1 \times D}$ at position P , which aggregates context information up to the current position, is extracted and concatenated with each latent token $I_C[i, j]$, forming context-aware query representations:

$$I_Q[i, j] = \text{MLP}(\text{Concat}(H_P, I_C[i, j])) \in \mathbb{R}^{1 \times D} \quad (9)$$

Subsequently, each latent token $I_C[i, j]$ is updated through localized attention as follows:

$$I_C[i, j] = \text{softmax} \left(\frac{Q[i, j]K[i, j]^T}{\sqrt{D}} \right) V[i, j] \in \mathbb{R}^{1 \times D} \quad (10)$$

where:

$$\begin{aligned} Q[i, j] &= W^Q \cdot I_Q[i, j], \quad Q[i, j] \in \mathbb{R}^{1 \times D}, \\ K[i, j] &= W^K \cdot \text{Patch}(I_A, i, j, w), \quad K[i, j] \in \mathbb{R}^{w^2 \times D}, \\ V[i, j] &= W^V \cdot \text{Patch}(I_A, i, j, w), \quad V[i, j] \in \mathbb{R}^{w^2 \times D} \end{aligned}$$

Here, W^Q, W^K, W^V represent learned projection matrices, and the Patch function extracts localized regions from I_A defined as:

$$\text{Patch}(I_A, i, j, w) = I_A[i \cdot w : (i+1) \cdot w, ; j \cdot w : (j+1) \cdot w] \in \mathbb{R}^{w^2 \times D} \quad (11)$$

During the decoding process, we consistently leverage the textual context to interact back with the vision features, thereby updating I_C . Through this recursive and context-aware alignment strategy, we dynamically enhance integration between visual and textual modalities.

3.3. Dual-Supervised Semantic Mapping Loss

In the previous sections, visual and textual features are frequently mapped and interact. To better guide feature mapping, we introduce Dual-Supervised Semantic Mapping Loss based on two complementary transformations: MLP_{v2t} for vision-to-text transformations and MLP_{t2v} for text-to-vision transformations.

For MLP_{v2t} , it is crucial to ensure that the mapped visual features align closely with the feature space of LLM. We shift our focus to V_{txt}^f , which, as discussed earlier, represents textual tokens processed by the vision encoder in the visual feature space. Consequently, after mapping V_{txt}^f through MLP_{v2t} , its rebuilt textual representation should closely resemble the LLM-based textual representation E_{txt} . To quantify the alignment quality of MLP_{v2t} , we compute the cosine similarity-based loss between these two feature representations:

$$\mathcal{L}_{v2t} = 1 - \frac{E_{txt} \cdot \text{MLP}_{v2t}(V_{txt}^f)}{|E_{txt}| \cdot |\text{MLP}_{v2t}(V_{txt}^f)|} \quad (12)$$

Analogous to \mathcal{L}_{v2t} , we use E_{img} to evaluate the effectiveness of MLP_{t2v} . An optimal MLP_{t2v} should transform

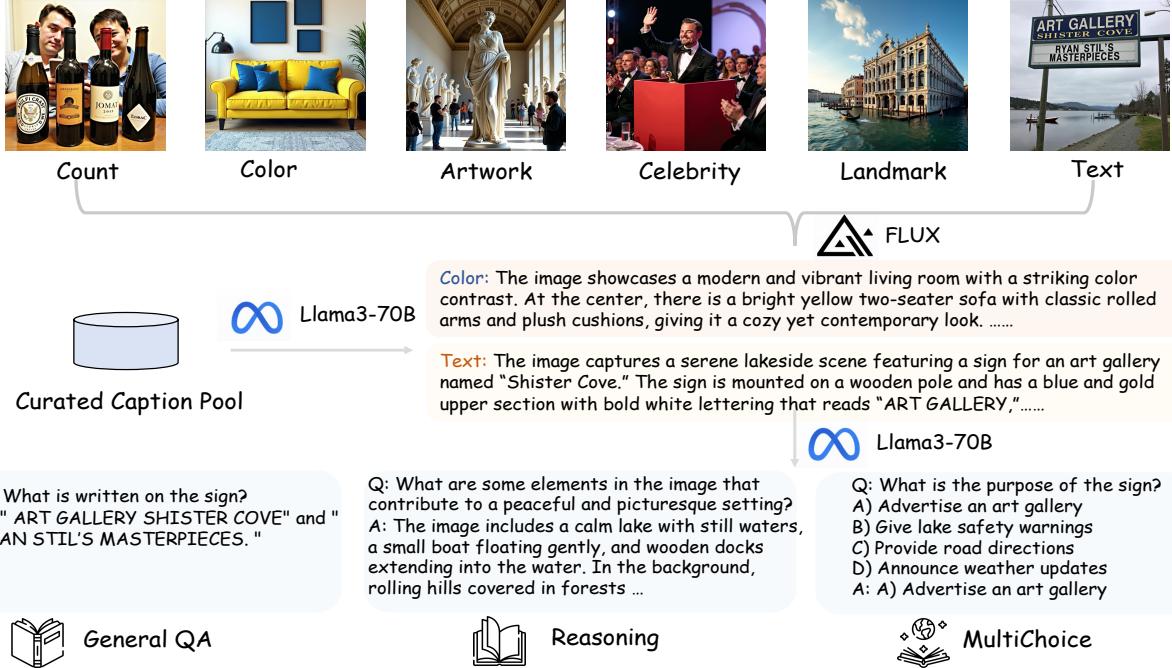


Figure 5. Overview of our Text-Centered QA Dataset framework. Our approach shifts the focus from visual content to textual richness by leveraging high-quality captions, enriching them with LLMs, and using them as the foundation for both image generation via diffusion models and diverse QA pair construction.

E_{img} into a rebuilt image representation that closely resembles the vision feature V_{img} in the visual space. Thus, we again employ cosine similarity loss to assess the quality of MLP_{t2v} :

$$\mathcal{L}_{t2v} = 1 - \frac{\text{MLP}_{t2v}(E_{img}) \cdot V_{img}}{|\text{MLP}_{t2v}(E_{img})| \cdot |V_{img}|} \quad (13)$$

In addition to the supervised feature space mapping loss, we incorporate a fundamental cross-entropy loss.

$$\mathcal{L}_{CE} = - \sum_{i=1}^n \log P(T_i | T_{<i}, V) \quad (14)$$

The complete training objective combines these losses with a balancing parameter :

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}). \quad (15)$$

The illustration of our reconstructed loss is shown in Figure 3. This dual-supervised approach explicitly reinforces semantic correspondence between the visual and textual representations, significantly boosting the effectiveness and accuracy of multimodal integration.

3.4. Synthesized Language-Driven QA Dataset

We propose a novel language-driven approach to constructing QA datasets, fundamentally shifting the traditional emphasis from visual content to the richness and diversity

of textual information. Unlike conventional methods, our framework prioritizes detailed textual descriptions, making text the central element guiding both image generation and QA pair creation.

As shown in Figure 5, our process begins with carefully selecting high-quality captions from an extensive caption pool. These initial captions are then enriched with LLMs, producing detailed and nuanced textual descriptions that capture various visual and contextual attributes. These enriched descriptions serve as prompts for diffusion model, which generates images closely aligned with the provided textual context. Simultaneously, we utilize the enriched descriptions as input content, leveraging the LLM again to construct diverse QA pairs, ensuring broad coverage across multiple-choice questions, multi-turn dialogues, reasoning-based tasks. This step distinguishes our method from traditional visual-first or MLLM-based approaches. Directly leveraging comprehensive textual data significantly enhances the model's understanding and representation of visual contexts.

Generative models inherently introduce challenges, including ambiguity, missing and inconsistencies. To address these issues, we implement a rigorous, multi-stage filtering process, with clearly defined criteria detailed in Appendix D.2. Examples of our prompts and QA pairs are provided in Appendix D.1.

By placing textual content at the heart of the dataset construction process, our approach offers a scalable, adaptable, and robust method for synthesizing diverse and high-quality QA datasets. This language-driven framework thus provides a fresh perspective on multimodal data creation, highlighting the essential role of textual richness in enhancing model comprehension and performance.

4. Experiments

Backbone

- *LLM*: We select two strong language models, Phi-3.5-mini-instruct[1] and LLaMA3.1-8B-instruct[3], due to their robust language modeling capabilities.
- *Vision Encoder*: We utilize SigLIP-SO400M-Patch14-384[76] and the most advanced SigLIP2-Giant-OPT-Patch16-384[62]. By leveraging the powerful SigLIP2-Giant-OPT-Patch16-384, we achieve superior performance with fewer tokens.
- *Projector*: MLP_{t2v} and MLP_{v2t} are both implemented as two-layer MLPs.

Training Strategies We introduce a three-stage training framework, distinct from traditional two-stage paradigms, ensuring comprehensive alignment and integration between visual and linguistic modalities.

- *Stage1: Foundational Semantic Alignment*: We pre-train the vision encoder using extensive image-caption datasets to establish precise semantic alignment between visual and textual representations.
- *Stage1.5: Contextual Multimodal Fusion*: In contrast to Stage 1, this intermediate stage incorporates various types of QA data along with image-caption pairs. This phase is designed to enhance the model’s adaptability in aligning vision and language representations across a broad spectrum of scenarios.
- *Stage2: Visual Instruction Tuning*: At this stage, we expose the model to various visual tasks, enabling it to answer downstream vision-related questions effectively.

Data In Stage 1, we utilize a large-scale collection of image-caption pairs, such as PixelProse[59], LLaVA 558k[40], and ShareGPT4V Captions[8]. Additionally, we use our pipeline produce 1.5M image-caption pairs across various scenarios. In Stage 1.5, the primary objective is to enhance question diversity. We generate 1M distinct QA samples spanning reasoning, conversational interactions, and multiple-choice questions. We also use publicly available datasets such as PixmoAskModelAnything[14], Q-Instruct[68] and LVIS-Instruct[63]. Additionally, we

borrow some domain-specific data, including Math, OCR, and Science from Stage 2 to further enrich the dataset. In Stage 2, the focus shifts towards vision-centric instruction tuning. We utilize part of domain-specific data from Stage 1.5. We incorporate datasets specifically designed for downstream visual tasks, such as LLaVA 665K[40], MMathCot[50], and Cambrian-7M[61], while also generating approximately 1M additional task-oriented samples. Details of dataset composition are provided in Appendix A.

Implementation Details For the Text-Guided Unified Vision Encoding component, we constrain the maximum supported text length to 256 tokens, with any excess truncated to ensure computational efficiency. In the Context-Aware Recursive Alignment Decoding module, we employ a fixed attention window size of 3 to strike a balance between performance and efficiency. Furthermore, we adopt a dynamic token mechanism that enables the model to adapt to varying lengths of visual token sequences. We randomly sample the number of latent tokens from the set {4, 16, 64, 144, 256} for each batch. In Appendix B.4, we show that this strategy can improve training stability while maintaining representational effectiveness. For the Dual-Supervised Semantic Mapping Loss, the balancing coefficient of λ is set to 0.1. For the main results, we use a total of 9M image-caption pairs in Stage 1, and 12M instruction data (5M in Stage 1.5 and 7M in Stage 2). In each stage, we unfreeze all components to ensure comprehensive optimization and seamless integration. In Appendix B.5, we observe that unfreezing all parameters in each stage accelerates model convergence and enhances performance. For all models, the global batch size during the pretraining stage is 256, with a cosine decay learning rate with maximum value 1e-4. In Stage 1.5, the global batch size is 128 and the learning rate is 2e-5. In Stage 2, we maintain a global batch size of 128 with a learning rate of 1e-5.

Evaluation. Follow [7], We evaluate the performance of different MLLM models on 21 benchmarks with four different categories:

- General multimodal benchmarks: MMBench (EN and CN)[43], VisWiz[23], POPE[16], MM-Vet[73], MME Perception[17], MME Cognition[17], SeedBench[33], HallusionBench[22], LLaVA in the Wild[40] and MMStar[9].
- OCR & Chart benchmark: TextVQA[58], OCRBench[44], ChartQA[51] and DocVQA[53].
- Knowledge based benchmark: AI2D[25], MathVista[49], MMMU[74] and ScienceQA[46].
- Vision Centric benchmark: MME-RealWorld[82], MMVP[79], RealworldQA[70] and CV-Bench[61].

Table 1. Model comparison on 22 Benchmarks. We provide model variants with different Vision Encoders: SigLIP-SO400M-Patch14-384 (FUSION) and SigLIP2-Giant-OPT-Patch16-384 (FUSION-X). FUSION-L is built upon FUSION-X by employing interpolation on image tokens during inference. FUSION-X achieves superior performance with fewer tokens.

(a) Results on general multimodal benchmarks.													
Model		# Vis tok.	General Benchmarks										
Method			MMB _{E,N}	MMB _{C,N}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	Seed-Image	HallB	LLaVA ^W	MMStar
<=4B Model Comparison													
Qwen2.5VL 3B	-	79.1	78.1	-	85.9	61.4	1592.4	607.5	74.0	46.6	-	56.3	
InternVL2 4B	-	78.5	73.9	-	84.6	50.5	1532.8	531.8	73.2	42.4	-	53.9	
DeepSeek-VL2-Tiny	-	74.6	72.1	-	-	52.5	1548.3	357.1	72.3	39.6	-	45.9	
MM1.5 3B	-	-	-	-	88.1	41.0	1478.4	319.6	72.4	-	73.0	-	
Phi 3.5-Vision	-	75.5	64.2	58.2	82.2	46.5	1473.4	412.1	69.9	53.3	68.8	49.0	
Florence-VL 3B	576	71.6	60.8	59.1	88.3	51.0	1498.7	403.9	70.6	58.1	71.1	44.9	
FUSION 3B (ours)	780	79.5	71.7	64.6	88.9	57.2	1595.9	416.5	74.6	51.4	84.7	52.4	
FUSION-X 3B (ours)	620	80.3	74.8	66.1	88.7	60.3	1582.1	440.0	75.3	51.9	85.2	50.9	
FUSION-L 3B (ours)	308	77.6	70.8	65.3	88.3	56.7	1573.7	406.8	74.1	48.7	77.6	47.7	
>=7B Model Comparison													
Qwen2VL 7B	-	83.0	80.5	-	88.4	62.0	1639.2	637.1	76.0	50.6	-	60.7	
InternVL2 8B	-	81.7	81.2	-	86.9	54.2	1639.7	575.3	75.4	45.2	-	61.5	
LLaVA-OneVision 8B	-	81.7	78.0	-	87.2	58.8	1626.0	483.0	74.8	47.5	86.9	60.9	
MM1.5 7B	-	-	-	-	88.6	42.2	1514.9	346.4	73.4	-	74.2	-	
Cambrain 8B	576	75.9	67.9	-	87.4	48.0	1547.1	-	74.7	48.7	71.0	50.0	
Florence-VL 8B	576	76.2	69.5	59.1	89.9	56.3	1560.0	381.1	74.9	57.3	74.2	50.0	
Eagle 8B	1024	75.9	-	-	-	-	1559.0	-	76.3	-	-	-	
FUSION 8B (ours)	780	80.5	74.9	59.5	89.3	60.0	1592.3	396.1	77.2	52.6	86.9	52.4	
FUSION-X 8B (ours)	620	82.0	76.2	62.9	88.8	60.0	1607.5	337.2	78.2	51.4	88.0	52.7	
FUSION-L 8B (ours)	308	80.0	73.6	59.9	88.5	57.3	1601.7	338.9	75.9	46.7	82.1	49.3	
(b) Results on Vision centric, Knowledge based, and OCR & Chart benchmarks.													
Model		# Vis tok.	Vision centric				Knowledge based				OCR & Chart		
Method			MME-RW	RWQA	CV-Bench	MMVP	A12D	MathVista	MMMU	SQA	TexVQA	OCRBench	ChartQA
<= 4B Model Comparison													
Qwen2.5VL 3B	-	53.1	65.5	-	-	81.4	61.2	51.2	79.3	-	82.8	84.0	93.9
InternVL2 4B	-	52.1	60.5	-	-	79.0	58.5	48.3	96.0	74.7	78.4	81.5	89.2
DeepSeek-VL2-Tiny	-	-	64.2	-	-	71.6	53.6	40.7	-	80.7	80.5	81.0	86.9
MM1.5 3B	-	-	56.9	-	-	65.7	44.4	37.1	85.8	76.5	65.7	74.2	87.5
Phi 3.5 Vision	-	-	53.5	69.3	67.7	77.4	-	43.3	89.0	61.1	59.8	72.0	75.9
Florence-VL 3B	576	-	60.4	70.2	64.7	73.8	52.2	41.8	84.6	69.1	63.0	70.7	-
FUSION 3B (ours)	780	41.5	65.1	76.4	76.0	78.9	54.3	44.7	87.1	71.8	60.0	75.7	70.9
FUSION-X 3B (ours)	620	41.7	63.7	78.3	78.1	79.2	54.9	44.2	87.3	73.9	63.7	75.8	71.1
FUSION-L 3B (ours)	308	39.5	61.8	76.2	77.0	77.3	48.6	43.4	85.6	71.4	56.9	67.7	63.5
>= 7B Model Comparison													
Qwen2VL 7B	-	57.4	70.1	-	-	83.0	58.2	54.1	85.5	84.3	86.6	83.0	94.5
InternVL2 8B	-	53.5	64.4	-	-	83.6	58.3	52.6	96.3	77.4	79.4	83.3	91.6
LLaVA-OneVision 7B	-	57.5	65.5	-	-	81.6	56.1	47.7	96.6	78.5	69.7	78.8	87.5
MM1.5 7B	-	-	62.5	-	-	72.2	47.6	41.8	89.6	76.5	63.5	88.1	78.2
Cambrain 8B	576	-	64.2	72.2	51.3	73.0	49.0	42.7	80.4	71.7	62.4	73.3	77.8
Florence-VL 8B	576	-	64.2	73.4	73.3	74.2	55.5	43.7	85.9	74.2	63.4	74.7	-
Eagle 8B	1024	-	66.5	-	71.6	76.1	52.7	43.8	84.3	77.1	62.6	80.1	86.6
FUSION 8B (ours)	780	46.0	65.2	78.7	78.7	80.4	56.6	43.1	89.2	77.3	63.8	80.3	78.6
FUSION-X 8B (ours)	620	44.7	66.1	79.2	79.7	81.4	59.4	42.2	90.3	74.7	66.6	79.8	77.8
FUSION-L 8B (ours)	308	42.3	65.1	78.2	76.7	79.2	55.2	41.8	88.3	72.8	59.5	73.0	66.0

Table 2. Performance analysis of model components. We evaluate the contribution of different components, including Text-Guided Unified Vision Encoding(TUNE), Context-Aware Recursive Alignment Decoding(CARD), and Dual-supervised Semantic Mapping(DSM)

TUNE	DSM	CARD	MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	SeedImage	LLaVA ^W	CVBench	MMVP	A12D	MMMU	SQA	TextVQA	ChartQA	
✓	✓	✓	LLaVA-NeXT	74.4	70.4	52.0	87.8	43.7	1552.0	331.4	74.4	81.4	71.2	66.0	71.2	41.1	77.3	69.8	69.7
				72.8	69.2	59.8	86.9	41.3	1529.6	323.7	72.4	77.0	66.2	66.5	70.2	37.8	74.9	63.0	54.0
				73.2	70.4	60.2	87.4	43.1	1589.8	340.4	72.0	80.5	66.7	68.0	70.6	39.2	75.8	62.2	54.1
				73.7	70.5	61.1	87.5	43.4	1538.3	351.2	72.7	81.1	67.4	68.0	71.0	39.8	76.0	62.8	55.9
				74.0	71.6	60.7	87.6	45.0	1506.1	330.4	72.3	83.9	66.8	69.0	71.4	38.6	77.1	63.3	55.6
				73.8	71.4	58.9	87.6	45.6	1568.6	350.4	72.3	78.9	67.0	68.3	71.2	40.0	76.2	63.0	54.9
				74.9	72.2	61.2	87.9	44.8	1558.3	361.8	73.0	82.7	69.7	70.0	71.7	39.2	78.3	63.6	55.5

Table 3. Impact of synthetic data on performance. We gradually incorporate our generated data into different stage, showing the improvement in performance on various tasks.

	Stage1	Stage1.5	Stage2	MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	SeedImage	LLaVA ^W	CVBench	MMVP	A12D	MMMU	SQA	TextVQA	ChartQA
baseline				77.2	71.4	66.6	88.1	57.4	1544.9	379.4	73.4	84.4	74.8	74.6	77.2	42.5	84.6	73.4	74.8
+1.5M	✓			77.5	71.6	65.7	88.2	58.2	1567.6	386.1	73.6	83.5	74.6	74.5	77.4	42.6	85.8	73.7	75.2
+1.0M	✓	✓		78.0	72.1	66.2	88.9	59.4	1567.2	408.5	74.1	81.9	86.3	74.2	78.8	43.9	85.7	73.5	74.9
+0.9M	✓	✓	✓	80.3	74.8	66.1	88.7	60.3	1582.1	440.0	75.3	85.2	78.3	78.1	79.2	44.2	87.3	73.9	75.8

Baselines. We evaluate our approach against a wide range of open-source baseline models. For the 3B model size, we choose InternVL2-4B[66], Qwen2.5VL-3B[4], DeepSeek-VL2-Tiny[69], Florence-VL 3B[7], Phi3.5-Vision 3B[1] and MM1.5 3B[78]. For the 8B model size, we choose InternVL2-8B[66], Qwen2VL-7B[65], Eagle-8B [57], Florence-VL 8B[7], Cambrian-8B[61] and LLaVA-OneVision-7B[34].

Result. As shown in Table 1, our model demonstrates leading performance across the majority of benchmarks. It significantly outperforms Cambrian-1. When compared to Florence-VL under a comparable data scale, our model achieves a substantial performance gain. Remarkably, our 3B model surpasses both Cambrian-1 8B and Florence-VL 8B. Furthermore, thanks to our fully integrated fusion and alignment mechanisms, our model requires only 620 tokens to achieve performance that is competitive with or superior to the current SOTA models. In nearly half of the benchmarks, our model outperforms InternVL2, which leverages high-resolution inputs with a substantially larger number of vision tokens and significantly more training data. Additionally, in certain metrics, our model’s performance approaches that of Qwen2.5VL. Moreover, our 3B model achieves the best result in MMBench, underscoring

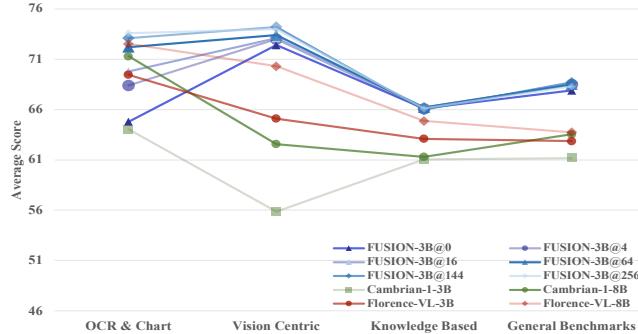
the effectiveness of our model architecture and the advantages conferred by our synthetic data strategy. We provide a comparative analysis with additional models in Figure 1. Additionally, a detailed visualization of our model’s alignment mechanisms is presented in Figure 2, illustrating how cross-modal integration is progressively refined throughout the pipeline.

5. Ablation Study

In this section, we first compare our model to the LLaVA-NeXT architecture to validate the effectiveness of each component within our framework. We present an additional ablation study that demonstrates the significant impact of latent token on model performance. Subsequently, we examine the contribution of synthetic data to modality alignment and downstream task performance. Lastly, to further highlight the superiority of our model’s full- modality integration, we reduce the number of tokens and analyze the model’s performance under increasingly con- strained conditions. Unless otherwise specified, the model used in this section is FUSION-X with its respective configuration.

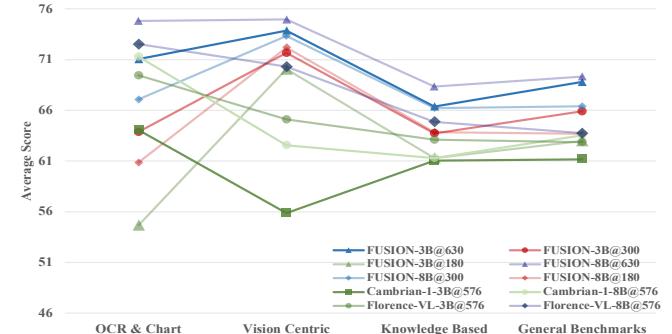
5.1. Component-wise Analysis of Modality Fusion

We select LLaMA3-8B as the language model and CLIP as the vision encoder, with Open LLaVA-NeXT as the base-



(a) Model performance under varying numbers of latent vision tokens.

Figure 6. Comparative analysis of performance with different latent and global vision token configurations.



(b) Model performance under varying numbers of global vision tokens.

line for comparison. Both models are trained on identical datasets comprising 558K pretraining samples and 1M Open LLaVA-NeXT finetuning dataset. Experimental results in Table 2 indicate that each component of our model significantly contributes to overall performance enhancement. Specifically, Text-Guided Unified Vision Encoding and Dual-supervised Semantic Mapping notably improve general multimodal tasks. We observe that the integration of Dual-supervised Semantic Mapping with Text-Guided Unified Vision Encoding results in a significant performance boost, highlighting the importance of modality alignment and the effectiveness of our Dual-supervised Semantic Mapping design. Furthermore, Context-Aware Recursive Alignment Decoding substantially boosts OCR accuracy and effectively reduces hallucinations. When integrated collectively, these components enable our model to outperform Open LLaVA-NeXT across more than half of benchmarks, without relying on dynamic resolution techniques.

5.2. Impact of Latent Token on Performance

To further underscore the pivotal role of latent tokens in our modality alignment strategy, we evaluate our 3B model’s performance under varying latent token configurations. As depicted in Figure 6a, even a little increase in latent token count leads to substantial improvements, particularly in OCR & Chart and Vision-Centric tasks. This indicates that latent tokens, while fewer in number, play a crucial role in boosting multi-modal understanding and alignment with minimal resource overhead. Further case studies and detailed analyses on the role of latent tokens are provided in the Appendix C.2.

5.3. Influence of Synthetic Data on Modality Fusion

To assess the value of synthetic data, we conduct experiments on our 3B model by removing generated samples utilized in each stage. Our results in Table 3 confirm that synthetic data significantly facilitate modality alignment and enhance performance in targeted downstream tasks. Specif-

ically, incorporating synthetic data during Stage 2 training markedly improves performance on tasks such as MM-Bench, SQA, CVBench and MMVP.

5.4. Robustness Under Reduced Token Constraints

To further showcase the effectiveness of our comprehensive modality alignment strategy, we test the model’s resilience by progressively reducing the number of image tokens through interpolation. Experimental outcomes in Figure 6b demonstrate that our model maintains superior performance compared to Florence-VL, even when the token count is reduced to 300 (we have named this version of the model FUSION-L). This robust performance under constrained conditions underscores the efficiency and adaptability of our modality fusion approach.

6. Conclusion

In this paper, we present FUSION, a family of MLLMs that rethinks how vision and language should be integrated. Rather than treating vision as static input fused only at the decoding stage, our model enables interaction at every phase—starting from vision encoding guided by text, to decoding that recursively aligns with visual features. To address semantic mismatch between modalities, we propose a dual-supervised loss that enforces consistency through bidirectional reconstruction. Beyond architecture, we also introduce a scalable data generation strategy that synthesizes diverse and informative QA pairs, providing strong supervision for vision-language alignment.

Our experiments show that FUSION achieves superior performance with far fewer vision tokens than existing models, highlighting both its efficiency and accuracy. Even with reduced input, the model matches or surpasses larger baselines, proving the value of deeper modality interaction.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Ben-haim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sam-budha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyra Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. [arXiv preprint arXiv:2404.14219](#), 2024. [7](#), [9](#)
- [2] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. [arXiv preprint arXiv:1810.12440](#), 2018. [1](#)
- [3] AI@Meta. Llama 3 model card. 2024. [7](#)
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. [arXiv preprint arXiv:2502.13923](#), 2025. [2](#), [3](#), [9](#)
- [5] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. [arXiv preprint arXiv:1905.13648](#), 2019. [1](#)
- [6] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. [arXiv preprint arXiv:2402.11684](#), 2024. [3](#), [1](#)
- [7] Juhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jian-feng Gao, Tianyi Zhou, and Bin Xiao. Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion. [arXiv preprint arXiv:2412.04424](#), 2024. [3](#), [7](#), [9](#)
- [8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. [arXiv preprint arXiv:2311.12793](#), 2023. [1](#), [3](#), [7](#)
- [9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? [arXiv preprint arXiv:2403.20330](#), 2024. [7](#)
- [10] Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. [arXiv preprint arXiv:2109.00122](#), 2022. [1](#)
- [11] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. [arXiv preprint arXiv:2312.14238](#), 2024. [2](#)
- [12] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation. [arXiv preprint arXiv:2108.06712](#), 2022. [1](#)
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. [arXiv preprint arXiv:2305.06500](#), 2023. [2](#), [3](#)
- [14] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. [arXiv preprint arXiv:2409.17146](#), 2024. [7](#), [1](#)
- [15] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In [Proceedings](#)

- of the 32nd ACM International Conference on Multimedia, pages 11198–11201, 2024. 7
- [16] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, Hanwen Jiang, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object, in any scene, with one reference. [arXiv preprint arXiv:2305.15727](#), 2023. 7
- [17] 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. Mme: A comprehensive evaluation benchmark for multimodal large language models. [arXiv preprint arXiv:2306.13394](#), 2024. 7
- [18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Or-gad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. [arXiv preprint arXiv:2304.14108](#), 2023. 10
- [19] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. [arXiv preprint arXiv:2402.05472](#), 2024. 2, 4, 7
- [20] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model. [arXiv preprint arXiv:2312.11370](#), 2023. 1
- [21] Luyan Gary. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 2012. 1
- [22] 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. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. [arXiv preprint arXiv:2310.14566](#), 2024. 7
- [23] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. [arXiv preprint arXiv:1802.08218](#), 2018. 7, 1
- [24] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. [arXiv preprint arXiv:2003.10286](#), 2020. 1
- [25] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. <http://dx.doi.org/10.1007/s10579-020-09517-1>, 55(3):661–688, 2020. 7, 1
- [26] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1
- [27] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. [arXiv preprint arXiv:2203.06486](#), 2022. 1
- [28] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. [arXiv preprint arXiv:2312.12241](#), 2023. 1
- [29] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyo Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. In *ECCV*, 2022. 1
- [30] Black Forest Labs. Flux. 2024. 1
- [31] LAION. <https://huggingface.co/datasets/laion/gpt4v-dataset>, 2023. 1
- [32] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. 2024. 1
- [33] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. [arXiv preprint arXiv:2307.16125](#), 2023. 7
- [34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. [arXiv preprint arXiv:2408.03326](#), 2024. 2, 3, 9, 7
- [35] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. [arXiv preprint arXiv:2403.00231](#), 2024. 1
- [36] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. [arXiv preprint arXiv:2403.18814](#), 2024. 3
- [37] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. [arXiv preprint arXiv:2311.06607](#), 2024. 3
- [38] Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichek Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. [HuggingFace repository](#), 2023. 1
- [39] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. [arXiv preprint arXiv:2311.07575](#), 2023. 2, 3
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. [arXiv preprint arXiv:2304.08485](#), 2023. 1, 3, 7

- [41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>, 2024. 2, 3
- [42] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024. 7
- [43] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahu Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2024. 7
- [44] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. <http://dx.doi.org/10.1007/s11432-024-4235-6>, 67(12), 2024. 7
- [45] Zheng Liu, Hao Liang, Bozhou Li, Tianyi Bai, Wentao Xiong, Chong Chen, Conghui He, Wentao Zhang, and Bin Cui. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*, 2025. 4, 9
- [46] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*, 2022. 7, 1
- [47] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2022. 1
- [48] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2023. 1
- [49] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2024. 7
- [50] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursu: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025. 7, 1
- [51] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 7, 1
- [52] Minesh Mathew, Viraj Bagal, Rubén Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infograph-icvqa. *arXiv preprint arXiv:2104.12756*, 2021. 1
- [53] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. *arXiv preprint arXiv:2007.00398*, 2021. 7, 1
- [54] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 3
- [55] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *ACL*, 2015. 1
- [56] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. *arXiv preprint arXiv:1912.03098*, 2020. 1
- [57] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2025. 2, 3, 9
- [58] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *arXiv preprint arXiv:2003.12462*, 2020. 7
- [59] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024. 7, 1
- [60] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 1
- [61] 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. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 3, 7, 9, 1
- [62] Michael Tschanen, 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. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 7
- [63] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 3, 7, 1

- [64] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 1
- [65] 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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 9
- [66] Weiyun Wang, Zhe Chen, Wenhui Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 9
- [67] Chris Wendler. [wendlerc/renderedtext](https://huggingface.co/datasets/wendlerc/RenderedText). <https://huggingface.co/datasets/wendlerc/RenderedText>, 2023. 1
- [68] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. *arXiv preprint arXiv:2311.06783*, 2023. 3, 7, 1
- [69] 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, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 9
- [70] x.ai. Grok 1.5v: The next generation of ai. 2023. Accessed: 2024-07-26. 7
- [71] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 1
- [72] Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, and Wentao Zhang. Sea: Supervised embedding alignment for token-level visual-textual integration in mllms. *arXiv preprint arXiv:2408.11813*, 2024. 2, 3
- [73] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2024. 7
- [74] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi 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, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2024. 7
- [75] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mammoth: Building math generalist models through hybrid instruction tuning. In *ICLR*, 2024. 1
- [76] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 7
- [77] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. *arXiv preprint arXiv:1903.02741*, 2019. 1
- [78] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, Zirui Wang, Afshin Dehghan, Peter Grasch, and Yinfei Yang. Mm1.5: Methods, analysis insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 9
- [79] He Zhang, Shenghao Ren, Haolei Yuan, Jianhui Zhao, Fan Li, Shuangpeng Sun, Zhenghao Liang, Tao Yu, Qiu Shen, and Xun Cao. Mmvp: A multimodal mocap dataset with vision and pressure sensors. *arXiv preprint arXiv:2403.17610*, 2024. 7
- [80] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2024. 1
- [81] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 3
- [82] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2025. 7
- [83] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*, 2017. 1
- [84] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*, 2021. 1
- [85] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2016. 1

FUSION: Fully Integration of Vision-Language Representations for Deep Cross-Modal Understanding

Appendix

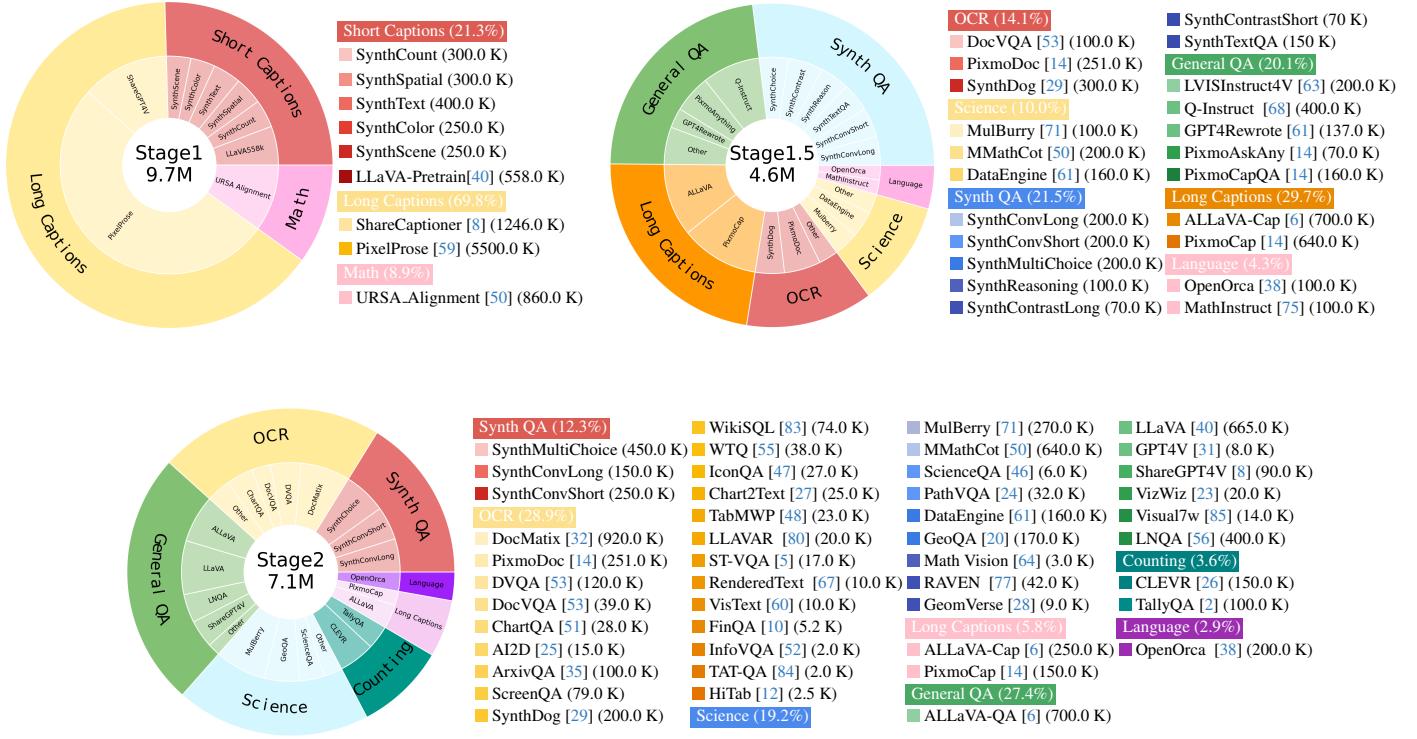


Table 4. Performance analysis of model components. We evaluate the contribution of different components, including Text-Guided Unified Vision Encoding(TUNE), Context-Aware Recursive Alignment Decoding(CARD), and Dual-supervised Semantic Mapping(DSM)

TUNE	DSM	CARD	MMB _{EN}	MMB _{CN}	VizWiz	POPE	MM-Vet	MME _P	MME _C	SeedImage	LLaVA _W	CVBench	MMVP	AID2	MMMU	SQA	TextVQA	ChartQA
<i>Vicuna-7B + Clip</i>																		
LLaVA-NeXT	68.5	60.7	59.7	87.1	40.6	1540.0	319.6	70.6	79.0	64.9	62.3	64.5	39.0	70.7	67.2	64.2		
✓			68.2	59.7	58.8	86.9	35.6	1526.4	298.4	68.5	73.1	63.1	64.1	63.7	37.0	69.1	58.7	52.7
✓		✓	68.9	60.7	59.7	87.1	36.7	1548.9	317.3	68.9	75.7	63.7	64.3	64.1	36.9	69.7	59.4	52.0
✓	✓		68.6	60.0	59.1	87.3	37.4	1543.1	309.4	68.7	75.2	64.8	64.8	64.9	37.1	70.4	60.9	52.6
✓	✓		69.4	61.6	60.2	87.2	38.2	1575.4	347.9	69.4	77.8	64.4	65.2	65.1	37.8	70.5	59.8	51.9
✓	✓		68.8	60.4	59.5	87.0	37.9	1550.3	322.1	69.0	74.2	65.0	64.6	64.4	36.4	70.9	60.8	52.4
✓	✓	✓	70.0	62.1	59.9	87.2	38.9	1564.1	326.4	69.8	76.4	65.8	65.7	64.9	36.8	71.4	61.2	52.2
<i>LLaMA3-8B + Clip</i>																		
LLaVA-NeXT	74.4	70.4	52.0	87.8	43.7	1552.0	331.4	74.4	81.4	71.2	66.0	71.2	41.1	77.3	69.8	69.7		
✓			72.8	69.2	59.8	86.9	41.3	1529.6	323.7	72.4	77.0	66.2	66.5	70.2	37.8	74.9	63.0	54.0
✓		✓	73.2	70.4	60.2	87.4	43.1	1589.8	340.4	72.0	80.5	66.7	68.0	70.6	39.2	75.8	62.2	54.1
✓	✓		73.7	70.5	61.1	87.5	43.4	1538.3	351.2	72.7	81.1	67.4	68.0	71.0	39.8	76.0	62.8	55.9
✓	✓		74.0	71.6	60.7	87.6	45.0	1506.1	330.4	72.3	83.9	66.8	69.0	71.4	38.6	77.1	63.3	55.6
✓	✓		73.8	71.4	58.9	87.6	45.6	1568.6	350.4	72.3	78.9	67.0	68.3	71.2	40.0	76.2	63.0	54.9
✓	✓	✓	74.9	72.2	61.2	87.9	44.8	1558.3	361.8	73.0	82.7	69.7	70.0	71.7	39.2	78.3	63.6	55.5

Table 5. Impact of the Interaction Layer. We investigate the effect of the Interaction Layer on model performance, specifically in scenarios where Latent Tokens are utilized.

Latent Tokens	Interaction Layers	MMB _{EN}	MMB _{CN}	VizWiz	POPE	MM-Vet	MME _P	MME _C	SeedImage	LLaVA _W	CVBench	MMVP	AID2	MMMU	SQA	TextVQA	ChartQA
<i>Vicuna-7B + Clip</i>																	
✓		69.4	61.6	60.2	87.2	38.2	1575.4	347.9	69.4	77.8	64.4	65.2	65.1	37.8	70.5	59.8	51.9
✓		69.1	62.3	58.8	87.0	38.7	1537.6	311.4	69.7	77.2	64.9	64.6	65.2	37.4	69.9	60.7	52.1
✓	✓	70.0	62.1	59.9	87.2	38.9	1564.1	326.4	69.8	76.4	65.8	65.7	64.9	36.8	71.4	61.2	52.2
<i>LLaMA3-8B + Clip</i>																	
✓		74.0	71.6	60.7	87.6	45.0	1506.1	330.4	72.3	83.9	66.8	69.0	71.4	38.6	77.1	63.3	55.6
✓		74.2	71.5	59.4	87.5	43.7	1567.8	331.8	72.6	79.4	68.8	68.0	72.2	38.2	77.4	63.6	55.8
✓	✓	74.9	72.2	61.2	87.9	44.8	1558.3	361.8	73.0	82.7	69.7	70.0	71.7	39.2	78.3	63.6	55.5

B. Additional Ablation Studies

In this section, we conduct additional ablation experiments to further validate the effectiveness and generalization ability of our proposed model. We also perform deeper investigations into our design choices. Unless otherwise specified, we evaluated two configurations throughout the following experiments: Vicuna 7B + Clip and LLaMA3 8B + Clip.

B.1. Extended Component-wise Analysis

We conducted additional experiments in Vicuna 7B configuration to further supplement the component-wise analysis presented in the main paper. The comprehensive results are

summarized in Table 4. The findings indicate that Context-Aware Recursive Alignment Decoding plays a crucial role in OCR-related tasks, achieving superior performance with only a minimal number of latent tokens introduced. Moreover, our experiments reveal that the combined use of Text-Guided Unified Vision Encoding and Dual-Supervised Semantic Mapping leads to significant performance gains, highlighting the effectiveness of modality alignment. The results consistently demonstrate that our model outperforms LLaVA-NeXT by a larger margin under this configuration.

Table 6. Impact of localized window size. We investigate the effect of varying window sizes within the Interaction Layer on model performance. Specifically, we experiment with window sizes of 2, 3, and 4.

Window Size																
	MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	SeedImage	LLaVA ^W	CVBench	MMVP	A12D	MMMU	SQA	TextVQA	ChartQA
<i>Vicuna-7B + Clip</i>																
2	69.5	61.2	59.0	86.9	38.1	1544.4	322.6	69.9	78.1	64.7	64.9	64.2	37.4	70.7	60.5	51.6
3	70.0	62.1	59.9	87.2	38.9	1564.1	326.4	69.8	76.4	65.8	65.7	64.9	36.8	71.4	61.2	52.2
4	69.8	61.7	59.3	87.3	39.2	1557.9	319.2	70.0	77.6	65.3	63.1	64.5	36.6	71.1	60.9	51.7
<i>LLaMA3-8B + Clip</i>																
2	74.0	71.5	59.8	87.4	46.2	1563.2	323.9	72.1	81.9	66.1	69.7	71.2	38.8	76.2	62.1	53.3
3	74.9	72.2	61.2	87.9	44.8	1558.3	361.8	73.0	82.7	69.7	70.0	71.7	39.2	78.3	63.6	55.5
4	74.3	71.7	59.4	87.0	44.5	1565.7	337.9	72.2	79.2	66.2	67.3	70.2	38.7	76.8	62.6	54.2

Table 7. Impact of the Number of Latent Tokens. We examine how different configurations of Latent Tokens affect model performance. Our experiments include both fixed-length settings (64, 144, 256) and dynamic-length configurations (ranging from 4 to 256, and 16 to 576 tokens).

Latent Token																
	MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	SeedImage	LLaVA ^W	CVBench	MMVP	A12D	MMMU	SQA	TextVQA	ChartQA
<i>Vicuna-7B + Clip</i>																
64	69.2	61.5	60.1	86.6	38.8	1548.2	322.7	69.6	78.4	66.3	64.8	65.4	36.9	70.5	59.9	51.1
144	69.5	61.9	60.3	86.9	38.4	1557.1	335.8	70.2	73.8	65.4	65.6	63.7	37.7	70.2	60.7	52.0
256	68.9	60.8	58.7	87.0	39.2	1546.9	320.4	68.9	74.2	64.7	64.5	64.2	37.4	69.9	61.5	52.4
4,16,64,144,256	70.0	62.1	59.9	87.2	38.9	1564.1	326.4	69.8	76.4	65.8	65.7	64.9	36.8	71.4	61.2	52.2
16,64,144,256,576	68.9	61.4	57.6	87.0	38.1	1543.2	319.6	69.3	70.5	64.4	62.8	64.3	35.9	68.8	59.2	51.7
<i>LLaMA3-8B + Clip</i>																
64	73.8	70.3	61.4	87.5	45.6	1535.9	356.8	72.7	79.1	68.8	69.6	71.1	41.0	77.1	62.9	54.0
144	74.0	71.1	59.7	87.7	45.5	1579.8	351.4	72.6	81.8	69.0	69.3	70.6	41.5	77.5	62.6	54.3
256	73.2	71.5	60.7	87.5	46.5	1560.4	348.2	72.4	77.2	68.5	69.4	71.0	40.0	76.9	64.2	56.7
4,16,64,144,256	74.9	72.2	61.2	87.9	44.8	1558.3	361.8	73.0	82.7	69.7	70.0	71.7	39.2	78.3	63.6	55.5
16,64,144,256,576	73.5	71.1	60.6	87.3	43.8	1544.3	320.7	72.2	82.2	68.1	68.0	70.1	38.0	75.5	62.6	54.2

B.2. Effectiveness of Context-Aware Decoding

In the above experiments, we demonstrated the effectiveness of Context-Aware Recursive Alignment Decoding for OCR tasks. However, a natural question arises: does the performance improvement stem merely from the additional image tokens introduced by this decoding strategy? To isolate this factor, we conduct an ablation study in which we still introduce Latent Tokens but remove the Interaction Layers. In this setting, the Latent Tokens serve solely as holistic representations of the image tokens, without engaging in contextual interaction. As shown in Table 5, introducing Latent Tokens alone yields no significant performance improvement compared to the baseline, with only a marginal gain observed on OCR tasks. In contrast, when the Interaction Layers are included, the model achieves a substantial overall performance boost, underscoring their critical role in effective context-aware decoding.

B.3. Effect of Localized Window Size

In the main experiments, we fix the localized window size to 3 as a trade-off between performance and computational efficiency. Here, we vary the window size to further investigate its impact. As shown in Table 6, increasing the window size does not lead to significant performance improvements, while it introduces a noticeable increase in computational cost. Conversely, when the window size is too small, interpolation is required to match the dimensions between the windowed features and the latent tokens. This interpolation leads to substantial compression of global visual information, resulting in a decline in performance.

B.4. Effect of the Number of Latent Tokens

In the main results, we set the number of latent tokens to {4, 16, 64, 144, 256}, randomly sampling from this set for each training batch. We find that this random selection strat-

Table 8. Effectiveness of Our Training Strategy. We validate the effectiveness of our training strategy by comparing different parameter-freezing schemes. Specifically, in Stage 1, when parameters are not fully unfrozen, only the projector and interaction layer are updated. In Stages 1.5 and 2, partial unfreezing involves training the projector, interaction layer, and the LLM, while keeping encoder fixed.

Stage1	Stage1.5	Stage2	MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	LLaVA ^W	CVBench	MMVP	AI2D	MMMU	SQA	TextVQA	ChartQA
✓	✓	✓	70.4	62.6	51.9	85.6	41.1	1396.8	272.4	72.7	63.2	66.8	69.5	38.9	73.4	63.9	62.2
	✓	✓	70.0	62.4	51.4	85.9	40.2	1379.1	307.9	70.6	62.5	66.9	69.2	38.1	72.6	63.3	61.9
✓	✓		69.3	62.8	51.2	85.1	39.8	1365.8	288.2	71.8	61.4	67.3	68.8	39.2	72.1	62.9	61.6
✓			68.9	62.7	51.3	85.1	37.7	1354.2	276.4	71.5	61.6	66.7	68.7	38.7	71.6	62.1	61.9
			68.7	62.8	51.1	85.3	35.9	1368.6	275.1	69.1	60.7	66.5	68.4	38.9	71.2	61.9	61.6

Table 9. Comprehensive evaluation of FUSION models with varying total visual token counts (180-630) across General, Vision-Centric, Knowledge-Based, and OCR & Chart benchmarks.

(a) Results on general multimodal benchmarks.

Model	# Vis tok.	General Benchmarks											
		MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	Seed-Image	HallB	LLaVA ^W	MMStar	
Method													
FUSION-X 3B	630	80.3	74.8	66.1	88.7	60.3	1582.1	440.0	75.3	51.9	85.2	50.9	
FUSION-X 3B	300	77.6	70.8	65.3	88.3	56.7	1573.7	406.8	74.1	48.7	77.6	47.7	
FUSION-X 3B	180	75.6	68.6	63.8	87.0	48.5	1485.3	375.0	72.1	45.5	74.1	45.6	
FUSION-X 8B	630	82.0	76.2	62.9	88.8	60.0	1607.5	337.2	78.2	51.4	88.0	52.7	
FUSION-X 8B	300	80.0	73.6	59.9	88.5	57.3	1601.7	338.9	75.9	46.7	82.1	49.3	
FUSION-X 8B	180	76.7	69.7	58.5	88.4	47.2	1561.6	302.9	73.6	45.5	77.8	46.4	
Cambrian-1 3B	576	74.6	64.7	58.5	87.7	44.1	1497.0	355.8	72.6	46.6	68.1	47.3	
Cambrian-1 8B	576	75.9	67.9	60.2	87.4	48.0	1547.1	372.6	74.7	48.7	71.0	50.0	
Florence-VL 3B	576	71.6	60.8	59.1	88.3	51.0	1498.7	403.9	70.6	58.1	71.1	44.9	
Florence-VL 8B	576	76.2	69.5	59.1	89.9	56.3	1560.0	381.1	74.9	57.3	74.2	50.0	

(b) Results on Vision centric, Knowledge based, and OCR & Chart benchmarks.

Model	# Vis tok.	Vision Centric				Knowledge Based				OCR & Chart			
		MME-RW	RealWorldQA	CVBench	MMVP	AI2D	MathVista	MMMU	SQA	TextVQA	OCRBench	ChartQA	DocVQA
Method													
FUSION-X 3B	630	41.7	63.7	78.3	78.1	79.2	54.9	44.2	87.3	73.9	63.7	75.8	71.1
FUSION-X 3B	300	39.5	61.8	76.2	77.0	77.3	48.6	43.4	85.6	71.4	56.9	67.7	63.5
FUSION-X 3B	180	37.7	61.2	75.3	73.7	74.6	44.2	42.7	83.7	67.8	47.3	56.1	47.5
FUSION-X 8B	630	44.7	66.1	79.2	79.7	81.4	59.4	42.2	90.3	74.7	66.6	79.8	77.8
FUSION-X 8B	300	42.3	65.1	78.2	76.7	79.6	55.2	41.8	88.3	72.8	59.5	73.0	66.0
FUSION-X 8B	180	41.2	63.2	77.1	76.3	76.5	51.6	40.6	86.7	71.6	54.2	63.2	54.4
Cambrian-1 3B	576	38.4	59.0	68.6	40.0	73.9	48.4	42.7	79.2	68.7	54.6	66.8	66.1
Cambrian-1 8B	576	41.2	64.2	72.2	51.3	73.0	49.0	42.7	80.4	71.7	62.4	73.3	77.8
Florence-VL 3B	576	40.6	60.4	70.2	64.7	73.8	52.2	41.8	84.6	69.1	63.0	70.7	70.8
Florence-VL 8B	576	43.2	64.2	73.4	73.3	74.2	55.5	43.7	85.9	74.2	63.4	74.7	75.1

egy balances the strengths of different token quantities and accelerates model convergence. As shown in Table 7, we evaluate both fixed numbers of latent tokens and dynamic

sampling from different ranges. Our results show that increasing the number of latent tokens leads to notable improvements in OCR tasks, which is consistent with our ear-

Table 10. Comprehensive evaluation of FUSION models with varying latent visual token counts (0-256) across General, Vision-Centric, Knowledge-Based, and OCR & Chart benchmarks.

(a) Results on general multimodal benchmarks.

Model	# Latent Vis tok.	General Benchmarks											
		MMB ^{EN}	MMB ^{CN}	VizWiz	POPE	MM-Vet	MME ^P	MME ^C	Seed-Image	HalIB	LLaVA ^W	MMStar	
Method	# Latent Vis tok.												
FUSION-X 3B	0	79.9	74.0	59.6	87.2	58.5	1563.4	422.1	74.1	51.0	81.4	50.9	
FUSION-X 3B	4	80.3	74.3	66.0	88.2	60.3	1577.9	429.8	74.7	51.4	86.2	50.5	
FUSION-X 3B	16	80.4	74.8	66.3	88.5	59.5	1584.3	431.5	74.3	51.6	83.4	49.6	
FUSION-X 3B	64	80.4	75.0	64.8	88.7	59.7	1585.7	436.1	74.5	51.4	88.0	50.0	
FUSION-X 3B	144	80.4	74.7	65.9	88.8	61.3	1600.4	450.2	74.6	51.5	85.2	49.5	
FUSION-X 3B	256	80.0	74.2	65.4	88.9	59.0	1595.3	440.5	74.4	51.5	85.2	48.1	
Cambrian-1 3B	-	74.6	64.7	58.5	87.7	44.1	1497.0	355.8	72.6	46.6	68.1	47.3	
Cambrian-1 8B	-	75.9	67.9	60.2	87.4	48.0	1547.1	372.6	74.7	48.7	71.0	50.0	
Florence-VL 3B	-	71.6	60.8	59.1	88.3	51.0	1498.7	403.9	70.6	58.1	71.1	44.9	
Florence-VL 8B	-	76.2	69.5	59.1	89.9	56.3	1560.0	381.1	74.9	57.3	74.2	50.0	

(b) Results on Vision centric, Knowledge based, and OCR & Chart benchmarks.

Model	# Latent Vis tok.	Vision Centric				Knowledge Based				OCR & Chart			
		MME-RW	RealWorldQA	CVBench	MMVP	A12D	MathVista	MMMU	SQA	TextVQA	OCRBench	ChartQA	DocVQA
Method	# Latent Vis tok.												
FUSION-X 3B	0	41.2	61.3	77.6	78.0	78.4	52.6	44.7	86.1	68.9	58.4	69.2	62.5
FUSION-X 3B	4	41.4	62.1	78.3	78.7	78.6	53.3	43.7	87.1	72.2	61.8	73.3	66.1
FUSION-X 3B	16	41.5	62.5	78.2	78.7	79.0	53.2	43.9	86.8	73.2	61.6	74.8	69.4
FUSION-X 3B	64	41.4	62.8	78.2	79.3	79.0	53.8	43.7	87.0	74.6	64.0	76.3	73.6
FUSION-X 3B	144	41.8	63.4	78.4	80.7	78.9	54.6	43.3	86.8	75.0	64.7	76.5	75.9
FUSION-X 3B	256	41.6	63.8	78.2	80.0	78.4	54.2	44.2	86.8	75.1	64.5	77.1	77.4
Cambrian-1 3B	-	38.4	59.0	68.6	40.0	73.9	48.4	42.7	79.2	68.7	54.6	66.8	66.1
Cambrian-1 8B	-	41.2	64.2	72.2	51.3	73.0	49.0	42.7	80.4	71.7	62.4	73.3	77.8
Florence-VL 3B	-	40.6	60.4	70.2	64.7	73.8	52.2	41.8	84.6	69.1	63.0	70.7	70.8
Florence-VL 8B	-	43.2	64.2	73.4	73.3	74.2	55.5	43.7	85.9	74.2	63.4	74.7	75.1

lier findings. Conversely, reducing the number of latent tokens yields slight performance gains on general and vision-centric benchmarks. Furthermore, when extending the sampling range to {16, 64, 144, 256, 576}, the model does not exhibit significant performance improvements. We hypothesize that, with 576 latent tokens, the use of localized attention requires interpolating global information to a much higher resolution, which may reduce information density and consequently limit performance gains.

B.5. Effectiveness of Training Strategy

In our main experiments, we unfreeze all parameters in every training stage. To assess the significance of this strategy, we conducted ablation studies by selectively freezing certain components of the model during different training

phases. To reduce computational cost, we adopt the Phi-3.5-mini-instruct+SigLip2 configuration and train each variant for 8000 steps in each stage. As shown in Table 8, fully unfreezing all parameters in each stage consistently results in superior performance. We attribute this to the fact that our framework departs from the conventional usage patterns of both the vision encoder and the language model. As a result, the model must continually adapt to a new processing paradigm, which requires end-to-end optimization across all components.

B.6. Supplementary Results

We provide detailed metrics and corresponding evaluation results for the models presented in Figure 6a and Figure 6b of the main paper. For the models shown in Figure 1,

Table 11. Comprehensive Benchmark Results Across All Models. We present a detailed comparison of benchmark metrics for all evaluated models, enabling a thorough assessment of their performance across various evaluation tasks.

	SeedBench_IMG	MMBench_TEST_EN	MMBench_TEST_CN	MME	MMVet	MMMU_VAL	MathVista	HallusionBench	LLaVA_Bench	AID	ScienceQA_TEST	MMstar	OCRBench	RealWorldQA	TextVQA_VAL	ChartQA_TEST	DOCVQA_TEST	POPE
VITA-1.5-8B	74.1	79.8	78.6	2328	52.7	52.6	66.2	44.6	72.8	79.2	95.8	60.2	74.1	66.9	70.3	84.1	92.2	88.9
Molmo-1B	67.9	64.6	46.6	1470	37.8	33.9	37.6	38.0	66.7	86.4	87.5	43.1	54.7	60.4	78.8	78.0	77.1	86.6
Molmo-7B	74.6	73.6	75.2	1853	41.5	49.1	48.7	46.4	72.4	91.0	58.1	56.1	65.6	71.1	81.7	84.1	92.2	89.0
Eagle-X5-7B	73.6	70.1	62.5	1866	33.3	37.6	38.8	35.4	59.5	73.6	71.2	41.7	55.1	63.8	70.5	66.4	80.0	89.3
Eagle-X5-8B	76.3	75.9	70.6	1922	53.3	43.8	52.7	50.0	72.0	76.1	84.3	53.0	62.6	66.5	77.1	80.1	86.6	89.8
Eagle-X5-13B	74.8	75.7	70.4	1947	52.6	41.6	54.4	47.8	70.2	74.0	82.0	53.7	61.9	64.8	75.5	77.6	85.9	89.8
InternVL2.5-1B	59.0	70.7	66.3	1950	47.2	41.2	47.1	39.4	60.2	69.0	93.1	51.3	77.4	57.5	72.0	75.9	84.8	89.9
InternVL2.5-2B	60.9	74.7	71.9	2138	62.6	43.2	51.1	42.3	62.7	74.9	96.0	54.3	80.2	60.1	74.3	79.2	88.7	90.6
InternVL2.5-4B	66.9	81.1	79.3	2338	61.5	51.8	60.8	46.6	84.3	81.4	97.6	58.7	82.0	64.3	76.8	84.0	91.6	90.9
InternVL2.5-8B	69.7	84.6	82.6	2344	62.8	56.2	64.5	49.0	87.3	84.6	98.4	63.2	82.1	70.1	79.1	84.8	93.0	90.6
LLaVA-OneVision(SI)-0.5B	62.1	55.8	50.2	1392	32.2	36.2	32.7	31.7	61.7	53.5	67.5	37.5	56.5	46.1	65.2	60.5	70.0	86.8
LLaVA-OneVision(SI)-7B	75.4	80.5	78.0	2146	50.6	46.8	58.6	47.5	77.6	82.8	86.6	56.7	69.7	69.5	78.5	80.9	87.5	87.2
VILA1.5-3B	68.0	64.5	56.2	1647	38.8	34.2	31.8	31.2	65.5	57.9	69.2	40.6	43.7	53.2	67.2	50.7	39.2	86.8
VILA1.5-8B	65.0	62.1	60.4	1698	41.9	37.4	37.4	35.3	71.7	58.8	73.2	39.7	43.8	53.2	68.5	52.7	40.6	83.3
VILA1.5-13B	72.7	74.4	67.7	1718	45.0	41.1	42.3	39.3	73.5	69.9	79.1	44.2	46.0	53.3	65.0	59.5	58.6	85.0
Cambrian-8B	73.3	74.6	67.9	1802	48.0	41.8	48.1	30.6	71.0	74.6	81.0	50.7	61.4	60.0	72.6	72.6	77.8	86.4
Cambrian-13B	73.2	73.2	65.0	1876	48.9	41.6	47.7	39.4	76.1	73.6	79.3	47.1	61.0	58.6	72.8	73.8	76.8	86.8
Florence-VL-3B	70.6	71.6	60.8	1902	51.0	41.8	52.2	58.1	71.1	73.8	84.6	44.9	63.0	60.4	69.1	70.7	75.9	88.3
Florence-VL-8B	74.9	76.2	69.5	1941	56.3	43.7	55.5	57.3	74.2	74.2	85.9	50.0	63.4	64.2	74.2	74.7	78.8	89.9
Qwen2-VL-2B	72.4	74.6	73.8	1899	51.5	42.2	48.0	42.4	50.5	74.7	78.7	47.5	79.7	60.7	79.7	73.5	90.1	87.3
Qwen2-VL-7B	76.0	82.8	80.6	2276	61.8	53.7	61.6	50.4	70.1	83.0	85.5	60.7	84.3	68.5	84.3	83.0	94.5	88.4
Qwen2.5-VL-3B	74.0	79.1	78.1	2157	61.8	53.1	62.3	46.3	77.0	81.4	81.4	56.3	82.8	65.5	79.3	84.0	93.9	87.3
Qwen2.5-VL-7B	77.0	83.2	82.8	2312	69.7	58.0	68.1	51.9	91.0	84.3	89.0	64.1	88.8	68.4	84.9	84.0	95.7	85.9
InternVL2-1B	65.2	65.2	61.0	1808	31.5	36.3	40.4	34.3	52.3	87.9	87.9	45.6	75.5	50.2	70.9	67.8	81.7	84.9
InternVL2-2B	70.9	73.4	71.2	1864	39.7	36.3	47.0	38.0	60.0	94.1	94.1	49.8	78.1	57.4	73.4	71.7	86.9	85.2
InternVL2-4B	73.2	78.5	73.9	2064	50.9	48.3	58.5	42.4	68.6	96.3	96.3	53.9	78.4	60.5	74.7	80.7	89.2	84.6
InternVL2-8B	75.4	82.0	80.9	2215	54.3	51.2	58.3	45.0	73.3	97.1	97.1	61.5	79.4	64.2	77.4	84.3	91.6	84.2
Phi-3.5-Vision-4B	69.7	76.0	66.1	1838	43.2	44.6	43.3	40.5	63.7	88.9	88.9	47.5	59.9	53.6	61.1	72.0	75.9	82.8
Phi-4.0-5B	73.2	83.4	77.7	1961	51.9	56.0	65.8	40.5	66.5	97.3	97.3	58.9	84.4	64.1	76.9	83.0	93.2	85.9
LLaMA3.2-11B	72.7	68.0	65.8	1822	57.6	48.0	47.7	40.3	85.8	77.3	83.9	49.8	75.3	63.3	75.3	80.0	87.5	88.1
LLaVA-v1.5-7B	65.8	66.5	46.4	1808	32.9	35.7	25.5	27.6	61.8	55.5	69.2	33.1	31.8	54.8	45.5	17.8	40.6	86.1
LLaVA-v1.5-13B	68.2	69.2	58.2	1780	35.6	37.0	27.7	24.5	66.1	61.1	72.6	34.3	33.7	55.3	48.9	18.2	50.1	88.4
LLaVA-Next-Vicuna-7B	69.6	69.2	62.3	1769	40.2	37.6	32.5	27.6	72.7	67.0	70.3	37.6	53.2	57.8	63.8	54.3	57.7	87.5
LLaVA-Next-Vicuna-13B	71.4	70.0	68.5	1745	44.9	37.3	35.1	31.8	73.9	72.2	73.7	40.4	53.7	57.6	66.9	61.4	65.1	87.8
MM1.5-1B	70.2	65.2	63.1	1611	37.4	35.8	37.2	30.2	71.6	82.1	82.1	30.2	60.5	53.3	72.5	67.2	81.0	84.2
MM1.5-3B	72.4	68.1	65.4	1798	41.0	37.1	44.4	35.1	73.0	85.8	85.8	35.6	65.7	56.9	75.6	74.2	87.7	85.6
MM1.5-7B	73.4	72.4	69.5	1861	42.2	41.8	47.2	40.6	74.2	89.6	89.6	42.7	63.5	62.5	76.5	78.6	88.1	86.3
FUSION-X 3B(Ours)	75.3	80.3	74.8	2011	60.3	44.2	54.9	51.9	85.2	79.2	87.3	50.9	63.7	63.7	73.9	75.8	71.1	88.7
FUSION-X 8B(Ours)	78.2	82.0	76.2	1955	60.0	42.2	59.4	51.4	88.9	81.4	90.3	52.7	66.6	66.1	74.7	79.8	77.8	88.8
FUSION-L 3B(Ours)	74.1	77.6	70.8	1980	56.7	43.4	48.6	48.7	77.6	77.3	85.6	47.7	54.9	61.8	71.4	67.6	63.5	88.3
FUSION-L 8B(Ours)	75.9	80.0	73.6	1938	57.3	41.8	55.2	46.7	82.1	79.2	88.3	49.3	59.5	65.1	72.8	73.0	66.0	88.5

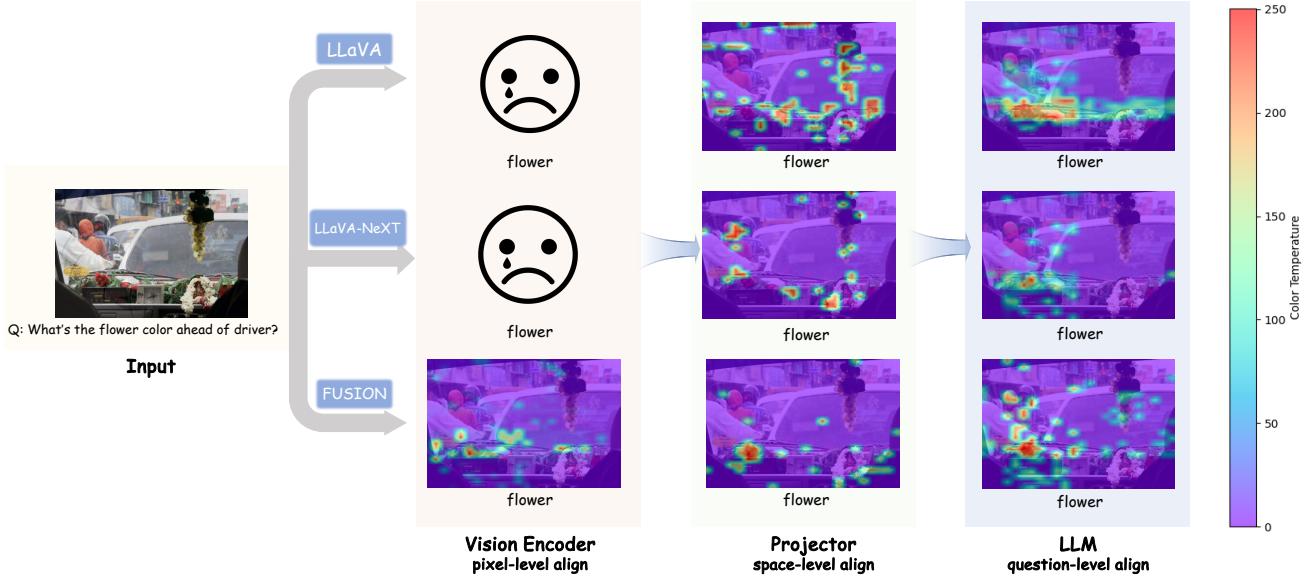


Figure 8. Comparison of Modality Alignment with traditional MLLMs. We conduct a comparative analysis of modality alignment across three different levels using FUSION, LLaVA, and LLaVA-NeXT. To ensure a fair evaluation, we adopt consistent visualization and augmentation strategies across all models.

we primarily relied on the open leaderboard¹ provided in [15], supplemented by results reported in individual papers as well as our own reproduction experiments. The results corresponding are presented in Table 11 and Table 9, Table 10, respectively.

C. Comparison and Deeper Exploration of Model Architectures

C.1. Compare with Traditional MLLMs

In this section, we provide a comparative analysis between our model and traditional MLLMs using additional heatmap visualizations. However, due to the unavailability of public weights and code for [19], we are unable to include pixel-level alignment visualizations for that model.

For the evaluation of space-level and question-level alignment, we adopt LLaVA[40] and LLaVA-OneVision[34] as our baselines. As shown in Figure 8, our model demonstrates significantly stronger alignment in both aspects. Specifically, the absence of Dual-Supervised Semantic Mapping Loss in LLaVA and LLaVA-OneVision prevents these models from achieving meaningful semantic alignment, resulting in poor performance in space-level alignment.

In terms of question-level alignment, our model consistently focuses more accurately on semantically relevant regions in response to the query. These results highlight

the effectiveness of our overall architecture and confirm the benefits introduced by our alignment strategies.

C.2. Beyond Single Questions

A key limitation of Text-Guided Unified Vision Encoding lies in its handling of multi-turn dialogue. In real-world scenarios, it is impossible to anticipate future user questions in advance. As a result, image encoding is typically performed based only on the initial question or the current question at hand, which can lead to representational bias in the visual features. Our proposed Context-Aware Recursive Alignment Decoding offers a suboptimal yet effective solution to this challenge.

As described in the main text, we initialize the question-aware latent tokens using conventional image encoding without any text guidance. This initialization helps mitigate the encoding bias introduced by question-specific image representations. Furthermore, we dynamically update the latent tokens based on each incoming question. This design acts as a compensation mechanism for any prior representational drift introduced by previous questions. To further evaluate our model’s ability to support multi-turn dialogue, we adopt ConvBench[42], a benchmark specifically designed to measure multi-turn dialogue capabilities in vision-language models. Using our pretrained model, we assess its performance in this setting. Results are reported in Table 12. While the use of Text-Guided Unified Vision Encoding alone may result in a slight drop in performance for multi-turn tasks compared to a simple encoder, the in-

¹Available at: https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Table 12. Comparison of Performance for LVLMs on ConvBench with Direct Grading method. S_1 , S_2 , S_3 , and S_0 correspond to the average scores of the first turn, second turn, third turn, and overall, respectively. \hat{S}_2 , \hat{S}_3 , and \hat{S}_0 represent the scores for the second turn, third turn, and overall when the first answer is given as a premise. \tilde{S} indicates the score when the first two answers are given as a premise. R_2 is defined as $(S_1 + S_2 + S_3)/3$, indicating the mean performance over three turns. Meanwhile, R_1 is computed as $(R_2 + S_0)/2$, representing the model’s overall score.

TUNE	CARD	Ablation										
		R_1	R_2	S_1	S_2	S_3	S_0	$\hat{S}_2(\hat{S}_2 - S_2)$	$\hat{S}_3(\hat{S}_3 - S_3)$	$\hat{S}_0(\hat{S}_0 - S_0)$	$\tilde{S}_3(\tilde{S}_3 - \hat{S}_3)$	
✓	✓	5.31	5.38	5.45	5.59	5.11	5.24	6.81(+1.22)	6.61(+1.50)	7.00(+1.76)	6.43(-0.18)	7.82(+0.82)
				5.23	5.35	5.36	5.63	5.07	5.11	6.87(+1.24)	6.55(+1.48)	6.98(+1.87)
✓		5.19	5.31	5.42	5.54	4.96	5.09	6.61(+1.07)	6.47(+1.51)	6.92(+1.83)	6.39(-0.08)	7.76(+0.84)
<i>General</i>												
Model		R_1	R_2	S_1	S_2	S_3	S_0	$\hat{S}_2(\hat{S}_2 - S_2)$	$\hat{S}_3(\hat{S}_3 - S_3)$	$\hat{S}_0(\hat{S}_0 - S_0)$	$\tilde{S}_3(\tilde{S}_3 - \hat{S}_3)$	$\tilde{S}_0(\tilde{S}_0 - \hat{S}_0)$
GPT-4V		7.09	7.30	7.30	7.48	7.12	6.88	8.23(+0.75)	8.00(+0.88)	8.25(+1.37)	7.34(-0.66)	8.18(-0.07)
Claude		6.54	6.75	6.53	7.04	6.68	6.32	7.48(+0.44)	7.06(+0.38)	7.55(+1.23)	7.18(+0.12)	8.13(+0.58)
FUSION-X-3B		6.03	6.10	6.35	6.28	5.66	5.97	7.28(+1.00)	6.81(+1.15)	7.46(+1.49)	6.49(-0.32)	8.15(+0.69)
ShareGPT4V-7B		5.83	5.99	6.02	6.14	5.80	5.67	7.19(+1.05)	6.77(+0.97)	7.31(+1.64)	6.93(+0.16)	8.19(+0.88)
XComposer2		5.82	5.98	5.98	6.17	5.78	5.66	7.35(+1.18)	7.04(+1.26)	7.66(+2.00)	7.00(-0.04)	8.20(+0.54)
FUSION-X-8B		5.70	5.82	6.12	6.08	5.27	5.58	7.09(+1.01)	6.45(+1.18)	7.22(+1.64)	6.64(+0.19)	8.01(+0.79)
InternVL-Chat-V1-5		5.60	5.76	6.11	5.93	5.25	5.43	6.34(+0.41)	5.60(+0.35)	6.50(+1.07)	6.32(+0.72)	7.79(+1.29)
Qwen-VL-Chat		5.54	5.65	5.96	5.78	5.22	5.43	7.04(+1.26)	6.53(+1.31)	7.26(+1.83)	6.57(+0.04)	8.00(+0.74)
InternVL-Chat-V1-2		5.49	5.69	5.80	5.88	5.39	5.29	6.66(+0.78)	6.12(+0.73)	6.75(+1.46)	6.31(+0.19)	7.70(+0.95)
LLaVA-V1.5-7B		5.16	5.29	4.95	5.59	5.34	5.03	7.28(+1.69)	6.68(+1.34)	7.28(+2.25)	6.72(+0.04)	7.97(+0.69)
mPLUG-Owl2		5.04	5.17	4.98	5.38	5.14	4.91	6.77(+1.39)	6.64(+1.50)	7.22(+2.31)	5.93(-0.71)	7.62(+0.40)
LLaVA-V1.5-13B		4.94	5.14	5.03	5.41	4.99	4.74	7.43(+2.02)	7.13(+2.14)	7.70(+2.95)	6.14(-0.99)	7.60(-0.10)
ShareGPT4V-13B		4.85	5.03	5.16	5.06	4.86	4.67	7.42(+2.36)	7.17(+2.31)	7.65(+2.98)	6.24(-0.93)	7.65(+0.00)

introduction of CARD leads to a significant improvement in overall dialogue capabilities. In fact, our model surpasses the performance of static encoding-based baselines.

Additionally, Figure 9 showcases qualitative examples of multi-turn interactions conducted by FUSION-X 3B, demonstrating its ability to engage in complex, multi-turn tasks such as captioning, reasoning, and creative generation. This illustrates that our model moves beyond functioning as a simple “answer machine”, and instead exhibits robust, context-aware dialogue capabilities.

D. Synthesized Language-Driven Framework Process

D.1. Prompt Design and Data Generation

We present the data generation prompts used in both the Pretraining and Finetuning stages. For the Pretraining data, we generated five distinct types of data: SynthColor, Synth-

Count, SynthSpatial, SynthText, and SynthScene. Each of these datasets was generated using a single prompt, and the details of these prompts are provided in Table 14, Table 15, Table 16, Table 17, Table 18.

In the Finetuning phase, specifically during Stage 1.5, we generated seven categories of data, including SynthMultiChoice, SynthConvLong, SynthConvShort, SynthReasoning, SynthTextQA, SynthContrastLong, and SynthContrastShort. Among these, SynthMultiChoice, SynthConvLong, SynthConvShort, and SynthReasoning were first created based on Captions to generate descriptions. These descriptions were then used to generate the corresponding QA pairs. The details of these prompts are provided in Table 24, Table 20, Table 22, Table 19. A more complex procedure was employed for SynthTextQA. Given the complexity of the OCR task and the limitations posed by caption elements, we first generated SynthText from the caption. Then, SynthText was used to generate Descriptions, and finally, the

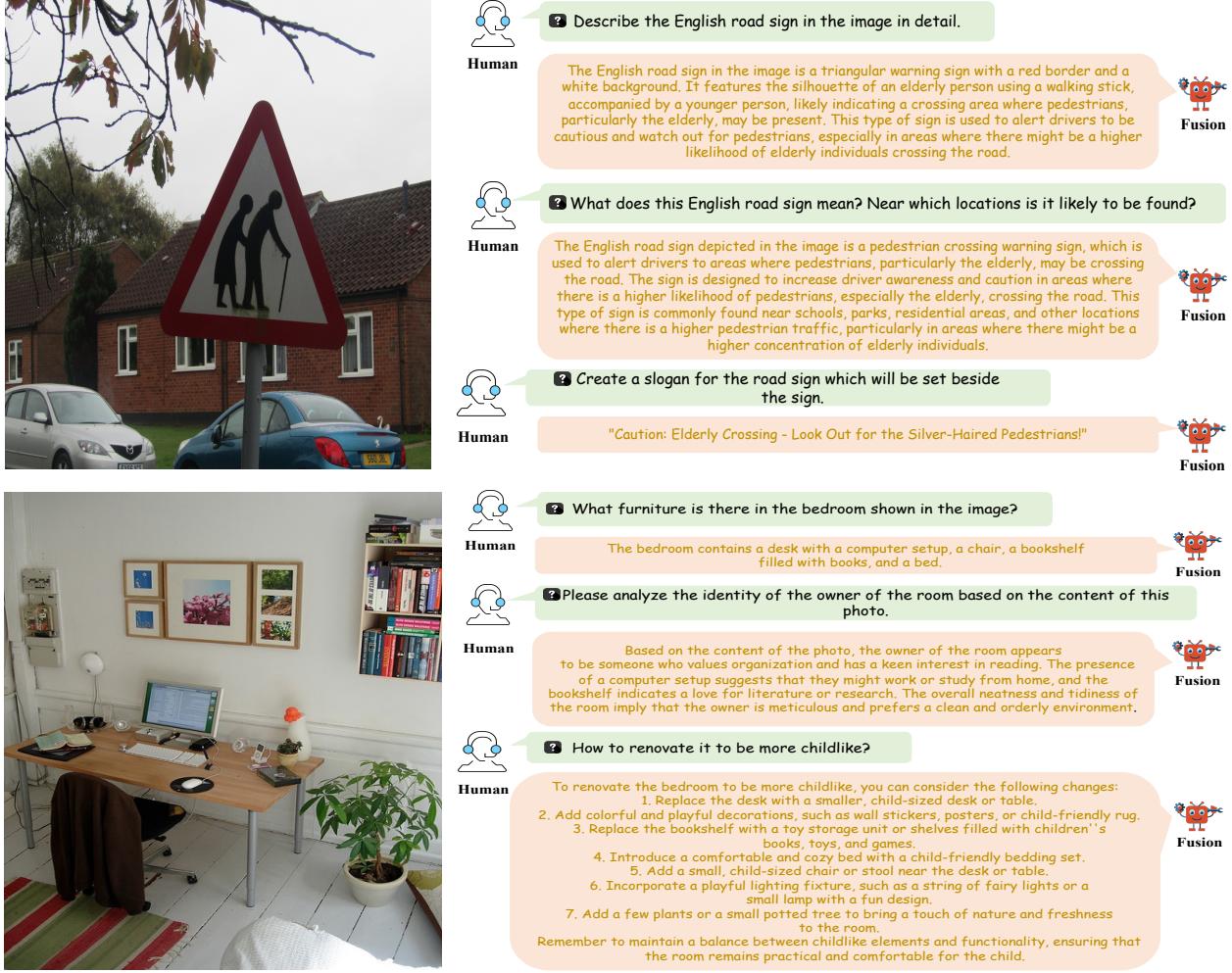


Figure 9. Illustration of our model’s multturn dialogue

QA pairs were produced. This approach ensures that the generated descriptions contain text elements, resulting in higher-quality QA pairs. The prompts are provided in Table 28, 29. For the SynthContrastLong and SynthContrastShort categories, the challenge was to highlight the detailed contrast between similar images. To address this, we carefully designed a prompt capable of generating two descriptions, each corresponding to an image, along with QA pairs that emphasize the differences between these two descriptions. The design of these prompts are presented in Table 27, 26. The primary focus of the prompts in this phase was to prioritize variability in QA pairs, thereby enhancing the model’s generalization across diverse multimodal contexts.

In Stage 2, we generated SynthMultiChoice, SynthConvLong, and SynthConvShort, using a similar process to Stage 1.5. The details of these prompts are presented in Table 25, 21, 23. The design of the prompts in this stage focused on

enhancing the model’s ability to follow specific instructions for various tasks.

We present specific examples and forms of the different data types in Figures 10, 11, 12, 13, 14, 15, 16, 17, 18, 19.

D.2. QA Pairs Data Filtering

Generative models inherently introduce challenges, including ambiguity, missing information, and inconsistencies in generated content. To address these issues, we implement a rigorous, multi-stage filtering process.

The pretraining data filtering process we adopt follows the approach outlined in [45]. We primary focus on the selection of finetuning data. Our filtering procedure consists of four main stages, each designed to ensure the quality and relevance of the data used in subsequent training steps.

Table 13. Metric and Prompt used for Caption Filtering

Caption Filtering
<pre>## Rule-Based Metrics</pre> <ul style="list-style-type: none"> • Alphanumeric Filter: Tokenization: false, Min ratio: 0.60 • Character Repetition Filter: Rep length: 10, Max ratio: 0.09373663 • Flagged Words Filter: Language: en, Tokenization: false, Max ratio: 0.0 • Perplexity Filter: Language: en, Max perplexity: 5500.0 • Special Characters Filter: Min ratio: 0.16534802, Max ratio: 0.42023757 • Word Repetition Filter: Language: en, Tokenization: false, Rep length: 10, Max ratio: 0.03085751 • Image-Text Matching Filter: HF BLIP: Salesforce/blip-itm-base-coco, Min score: 0.8, Max score: 1.0, Horizontal flip: false, Vertical flip: false, Reduce mode: avg, Any or all: any, Mem required: 1500MB • Image-Text Similarity Filter: HF CLIP: openai/clip-vit-base-patch32, Min score: 0.28 <pre>## Prompt</pre> <p>Assume you are an expert in the field of AI image generation. Your goal is to select high-descriptive prompts that will enable the successful generation of images. I will provide you with a specific descriptive prompt, and your task is to evaluate it thoroughly. Consider the prompt's level of detail, its logical coherence, and the clarity with which it describes the desired image. It is essential to assess whether the prompt contains sufficient information to guide the diffusion model effectively, ensuring that it can produce an image that meets expectations. You should only respond with Yes or No.</p>

D.2.1 Caption Filtering

The first step focuses on selecting high-quality captions for generating the QA pairs. The captions are sourced from the Datacomp[18] dataset, which contains a significant amount of low-quality or irrelevant data. Prior to generating descriptions, we conduct a filtering process to remove low-quality captions, such as those containing watermarks, advertisements, or non-English text. Initially, rule-based filtering methods are applied, including:

- Removal of character-level and word-level repetitions,
- Removal of captions with high perplexity values,
- Ensuring strong image-text alignment between the caption and its corresponding original image.

This step ensures that the selected captions are better aligned with real-world images and more easily extensible. Subsequently, we use LLM to score these captions, ensuring their semantic correctness and syntactic richness. The scoring process emphasizes the diversity of the elements in the captions. The prompt, along with the specific rule-based methods, is outlined in Table 13. For the filtering process, we use Data-Juicer for rule-based filtering and LLaMA3.1 70B for scoring. After this step, approximately 20 million captions are retained.

D.2.2 Description Filtering

The second step focuses on selecting high-quality descriptions for generating QA pairs and images. The 10 million captions selected in Step 1 are input into LLaMA3.1 70B to generate corresponding descriptions. A similar filtering approach is employed for descriptions as in Step 1, with the key difference being the exclusion of image-text matching calculations. After filtering, approximately 12 million descriptions are retained.

D.2.3 Image Generation Filtering

This step aims to ensure the quality of the generated images and their alignment with the corresponding descriptions.

Let I represent the generated image, and D represent the corresponding description. We first calculate the ClipScore between the description D and the generated image I to measure the alignment.

To assess the image quality, we utilize the Structural Similarity Index (SSIM) metric. As described in the main paper, I will be resized and processed by Siglip. Additionally, I will be divided into four non-overlapping sub-images, each processed individually through Siglip. The SSIM is computed both between the resized image and its original version, as well as between each sub-image and its corresponding resized sub-image.

Let the height and width of the image be denoted as H

and W , respectively, and let the crop size used in Siglip be S . The first step is to calculate the SSIM between the original image I and the resized image:

$$\text{SSIM}_w = \text{SSIM}(I, \text{Resize}(\text{Resize}(I, (S, S)), (H, W)))$$

For the four sub-images, we first partition the original image into four non-overlapping regions. Specifically, each sub-image $I_{\text{sub}(i,j)}$ corresponds to a section of the image, where i and $j \in \{1, 2\}$ indicate the row and column of the sub-image. The extraction of the (i, j) -th sub-image is defined as:

$$I_{\text{sub}(i,j)} = I \left[\frac{H}{2}(i-1) : \frac{H}{2}i, \frac{W}{2}(j-1) : \frac{W}{2}j \right]$$

The SSIM between each sub-image and its resized version is computed as:

$$\text{SSIM}_s(i, j) = \text{SSIM}(I_{\text{sub}(i,j)},$$

$$\text{Resize} \left(\text{Resize}(I_{\text{sub}(i,j)}, (S, S)), \left(\frac{H}{2}, \frac{W}{2} \right) \right)$$

Next, we compute the overall SSIM score for the entire image by combining the SSIM of the full image (SSIM_w) and the SSIM values for each of the four sub-images ($\text{SSIM}_s(i, j)$):

$$\text{SSIM}_a = w_1 \times \text{SSIM}_w + \sum_{i=1}^2 \sum_{j=1}^2 w_2 \times \text{SSIM}_s(i, j)$$

where w_1 is the weight for the SSIM of the full image and w_2 is the weight for each sub-image's SSIM. These weights are chosen to appropriately balance the contribution of the full image and sub-images to the overall image quality score.

Finally, the total score for the image is computed by combining the ClipScore with the overall SSIM score SSIM_a , using a weighting factor λ :

$$\text{Total Score} = \text{ClipScore}(I, D) + \lambda \times \text{SSIM}_a$$

In our implementation, w_1 is set to 1 and w_2 is set to 0.25. λ is set to 0.5 to balance the importance of the SSIM score and ClipScore. After this filtering step, we retain approximately 6 million images.

D.2.4 QA Pair Filtering

The final step involves ensuring the quality and relevance of the generated QA pairs. We begin by using MLLM to verify the correctness of the answers within each QA pair, filtering out all incorrect QA pairs. Subsequently, we convert each QA pair into a declarative sentence and compute

the ClipScore between the statement and the corresponding image. The final score for a QA pair is the average of the ClipScores for each declarative sentence within the QA pair:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^N \text{ClipScore}(\text{QA}_i, I)$$

where N is the number of questions in the QA pair. The QA correctness is evaluated using the LLaVA-OneVision 72B, while the transformation to declarative sentences is carried out with LLaMA3.1-70B. After filtering, we retain approximately 4 million QA pairs.

Table 14. Prompt used for SynthColor

SynthColor
<pre> ## Role Assume you are an AI visual assistant. ## Goal I will provide you with a brief description of an image. Your primary task is to accurately construct this image in your mind. Based on your understanding and the basic image, you need to adjust the colors of objects to enhance the expressiveness and visual information of the image. If the original description is vague or insufficient, creatively adjust or add color elements to improve clarity and impact. ## Rule To maintain harmony: <ul style="list-style-type: none"> The number of color types should not exceed three The overall description should be concise Colors should be diverse, avoiding repetition After modifying, you need to provide a new description that specifies the color of each significant object. Ensure that the new description is: <ul style="list-style-type: none"> Logically consistent Overall harmonious Focused on visual information Minimizing irrelevant content No more than 12 words ## Output Format Present the description content directly, without any extraneous information.</pre>

Table 15. Prompt used for SynthCount

SynthCount
<pre> ## Role Assume you are an AI visual assistant. ## Goal I will provide you with a brief description of an image. Your primary task is to accurately construct this image in your mind. Based on your understanding and the basic image, you need to modify the quantity of objects to enhance the expressiveness and the visual information of the image. If the original description is vague or insufficient, creatively adjust or add elements to improve clarity and impact. ## Rule To maintain harmony: <ul style="list-style-type: none"> The quantity of each object type should not exceed three The total number of objects should not exceed six After modifying, you need to provide a new description that quantifies each significant object type. Ensure that the new description is: <ul style="list-style-type: none"> Logically consistent Overall harmonious Focused on visual information Minimizing irrelevant content No more than 12 words ## Output Format Present the description content directly. Here are some possible formats: <ol style="list-style-type: none"> Three balloons float above a table where two children play board games. Two dark mountains loom above the water with three people at the shoreline. </pre>

Table 16. Prompt used for SynthSpaital

SynthSpaital
Role
Assume you are an AI visual assistant.
Goal
I will provide you with a brief description of an image. Your primary task is to accurately construct this image in your mind, focusing particularly on the spatial arrangement and relative positions of objects within it. Based on your understanding, you need to modify the positions of objects to highlight their interactions and improve spatial coherence. If the original description lacks detail on positioning, creatively adjust or add elements to clarify spatial relationships and enhance visual impact.
Rule
Ensure that the arrangement is logical and maintains aesthetic harmony. After modifying, you need to provide a new description that quantifies each significant object type. Ensure that the new description is:
<ul style="list-style-type: none"> • Logically consistent • Overall harmonious • Focused on visual information • Minimizing irrelevant content • No more than 12 words
Output Format
Present the description content directly. Here are some possible formats:
1. Bicycle leaned against the left side of a park bench under a tree.
2. Bed against the right wall, with luggage beside it, near curtained windows.
3. Woman sits inside dumpster, back against the left wall, surrounded by trash.

Table 17. Prompt used for SynthText

SynthText
Role
Assume you are an AI visual assistant.
Goal
I will provide you with a brief description of an image. Your primary task is to accurately construct this image in your mind. Based on your understanding and the basic image, you need to add textual content to enhance the expressiveness and the visual information of the image, and these texts should be closely related to the theme and visual elements of the image.
Rule
The textual information can include:
<ul style="list-style-type: none"> • Words and numbers on signs • Billboards • Posters • Digital displays • Graffiti • Slogans • Traffic directions • Road signs • Historical markers • Famous quotes
Additionally, you can add other types of text as needed to ensure the diversity and richness of the text. You can also create visual content appropriately if the given description is not suitable for adding text.
After adding text, you need to provide a new description that includes the text you added. Remember:
<ul style="list-style-type: none"> • Your response must contain text • You must specify the type for each text • The description should be logically consistent • Focus on visual information • No more than 12 words
Output Format
Present the description content directly. Here are some possible formats:
1. A vibrant mural displays 'Dream Big!' with colorful butterflies and stars.
2. A street sign reads 'Main St' with a red arrow pointing right.

Table 18. Prompt used for SynthScene

SynthScene
Role
Assume you are an AI visual assistant.
Goal
I will provide you with a brief description of an image. Your primary task is to accurately construct this image in your mind. Based on your understanding and the basic image, you need to adjust or enhance the scene elements to improve the expressiveness and visual information of the image.
Rule
If the original description is vague or insufficient, creatively adjust or add scene elements such as:
<ul style="list-style-type: none"> • Landmarks • Iconic buildings • Natural landscapes • Urban settings • Specific environments • Historical sites • Famous monuments • Picturesque vistas • Distinctive locations
Remember:
<ul style="list-style-type: none"> • Each caption must contain only one scene element • Scene elements should be diverse (avoid repetition like multiple Eiffel Towers) • Description should be logically consistent and harmonious • Focus on visual information, minimize irrelevant content • Response must be no more than 12 words
Output Format
Present the description content directly, without any extraneous information.

Table 19. Prompt used for SynthReasoning

Reasoning
Role
Suppose you are an AI visual assistant, viewing a detailed description of an image.
Goal
The task is to use the provided description of an image, create some plausible questions about the image, and provide the answers in detail. You can create complex questions beyond describing the scene. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first. To answer such questions, you should require first understanding the visual content, then based on the background knowledge, think through the solution step by step. Either explain why the things are happening that way, or provide guides and help to user's request.
Rule
Please give the Q&A content directly without any extraneous words and separate questions and answers with Q and A. Ensure that the answer to each question is no less than 100 words. Additionally, you should control the QA pairs not exceeding 6 pairs to preserve visual clarity.

Table 20. Prompt used for Stage1.5 SynthConvLong

Stage1.5-SynthConvLong

Role

Suppose you are an AI visual assistant, viewing a detailed description of an image.

Goal

Design a conversation where a user asks a wide range of questions about an image, and the AI visual assistant responds based on what it sees. The questions should cover a broad spectrum of aspects related to the image, ensuring variety and inclusivity of different types of inquiry. The aim is to explore the image from multiple angles and include different kinds of questions to create a diverse, dynamic conversation.

Rule

The questions should explore the following aspects:

- **Object types:** What objects are visible in the image? What are their characteristics?
- **Counting objects:** How many of a certain type of object or entity can be seen in the image?
- **Object actions:** Are objects being used or manipulated in any way?
- **Object locations:** Where are specific objects located relative to one another in the image?
- **Relative positions between objects:** How are objects positioned in relation to one another in the scene?
- **Background Elements:** What is the background environment like? What elements or features are visible in the backdrop?
- **Lighting & Color:** What kind of lighting or color schemes dominate the image? How does the lighting affect the mood or feel of the scene?
- **Textures and Details:** Are there any significant textures or small details in the image that stand out?

Ensure that only questions with clear, confident answers are included:

1. The content referenced in the question is clearly visible in the image, and the answer can be confidently provided.
2. It is clear from the image that the content is not present.

What's more, the conversation should span these different dimensions, including variations in the types of questions to ensure the dialogue explores a rich variety of observations about the image. Ensure that the conversation is logically consistent, focuses on describing the visual information, and avoids unnecessary distractions. Keep the number of QA pairs to no more than 8 to maintain clarity.

Example Output Format

Q: How many people are visible in the image?

A: there are two people visible, one sitting at a desk and the other standing near the window.

Q: What are the people doing in the image?

A: The person sitting at the desk appears to be typing on a laptop, while the other person is holding a cup, likely standing near the coffee machine.

Q: What is in the background of the image?

A: In the background, there's a large window with some greenery visible outside, suggesting it's a sunny day.

Q: How is the room lit?

A: The room is brightly lit by natural sunlight streaming in through the window, with soft shadows from the furniture.

Q: Is there any art or decoration on the walls?

A: Yes, there is a large modern painting hanging on the wall above the desk.

Table 21. Prompt used for Stage2 SynthConvLong

Stage2-SynthConvLong
<pre>## Role Suppose you are an AI visual assistant, viewing a detailed description of an image. ## Goal Design a conversation where a user asks questions about an image, and the AI visual assistant responds with high precision and specificity, based solely on what is clearly visible in the image. Each answer should be directly aligned with the visual content, ensuring that it is 100% accurate and reflects the scene as it appears. The questions should be carefully framed to avoid any speculative or unclear answers. ## Rule Key aspects to focus on: <ul style="list-style-type: none"> • Object types: Identify objects and details based on their distinct visual attributes, ensuring the description matches what is observable in the image. • Counting objects: Provide exact numbers of objects or entities visible, with no guesswork. • Object actions: Only describe actions that can be clearly observed in the image, with no assumptions. • Object locations: Describe precise spatial relationships between objects based on their positioning in the image. • Relative positions between objects: Indicate how objects are arranged in relation to one another, using clear visual evidence from the image. Ensure that only questions with clear, confident answers are included: 1. The content referenced in the question is clearly visible in the image, and the answer can be confidently provided. 2. It is clear from the image that the content is not present. What's more, avoid asking questions that cannot be confidently answered from the image. The conversation can include more complex questions, such as asking about the background knowledge of the objects in the image or discussing events happening in the image. However, make sure these questions are based on clear visual evidence from the image and not uncertain speculation. Ensure that the conversation is logically consistent, focuses on describing the visual information, and avoids unnecessary distractions. Keep the number of QA pairs to no more than 8 to maintain clarity. ## Example Output Format Q: How many lamps are visible in the image? A: here are two lamps, both on either side of the room, near the walls. Q: What color is the couch? A: The couch is a deep brown color, with a smooth leather texture. Q: Where is the desk located in the image? A: The desk is located against the far wall of the room, with a chair placed in front of it. Q: What is the shape of the rug? A: The rug is rectangular, positioned underneath the coffee table. Q: Is there a TV in the room? A: No, there is no television visible in the room.</pre>

Table 22. Prompt used for Stage1.5 SynthConvShort

Stage1.5-SynthConvShort
<p>## Role Suppose you are an AI visual assistant, viewing a detailed description of an image.</p> <p>## Goal Design a conversation where a user asks questions about an image, and the AI visual assistant responds with brief, concise, and to-the-point answers. The focus is on providing short responses that directly match what is visible in the image, with no elaboration. The AI should answer clearly and simply, avoiding long explanations or unnecessary details. The questions should cover a broad spectrum of aspects related to the image, ensuring variety and inclusivity of different types of inquiry. The aim is to explore the image from multiple angles and include different kinds of questions to create a diverse, dynamic conversation.</p> <p>## Rule The questions should explore the following:</p> <ul style="list-style-type: none">• Object types: What objects are visible in the image? What are their characteristics?• Counting objects: How many of a certain type of object or entity can be seen in the image?• Object actions: Are objects being used or manipulated in any way?• Object locations: Where are specific objects located relative to one another in the image?• Relative positions between objects: How are objects positioned in relation to one another in the scene?• Background Elements: What is the background environment like? What elements or features are visible in the backdrop?• Lighting & Color: What kind of lighting or color schemes dominate the image? How does the lighting affect the mood or feel of the scene?• Textures and Details: Are there any significant textures or small details in the image that stand out? <p>Ensure that only questions with clear, confident answers are included:</p> <ol style="list-style-type: none">1. The content referenced in the question is clearly visible in the image, and the answer can be confidently provided.2. It is clear from the image that the content is not present. <p>What's more, the conversation should span these different dimensions, including variations in the types of questions to ensure the dialogue explores a rich variety of observations about the image. Ensure that the conversation is logically consistent, focuses on describing the visual information, and avoids unnecessary distractions. Keep the number of QA pairs to no more than 8 to maintain clarity.</p> <p>## Example Output Format</p> <p>Q: What furniture is in the room? A: Sofa and table</p> <p>Q: What is placed on the table? A: Coffee cup</p> <p>Q: What is the color of the carpet? A: Grey</p> <p>Q: Where is the clock located? A: On the wall</p>

Table 23. Prompt used for Stage2 SynthConvShort

Stage2-SynthConvShort
<p>## Role Suppose you are an AI visual assistant, viewing a detailed description of an image.</p> <p>## Goal Design a conversation where a user asks questions about an image, and the AI visual assistant responds with high precision and specificity, based solely on what is clearly visible in the image. The answers should be as brief as possible, ensuring that they directly reflect the visible elements in the image. Each answer should be directly aligned with the visual content, ensuring that it is 100% accurate and reflects the scene as it appears. The questions should be carefully framed to avoid any speculative or unclear answers.</p> <p>## Rule Key aspects to focus on:</p> <ul style="list-style-type: none">• Object types: Identify objects and details based on their distinct visual attributes, ensuring the description matches what is observable in the image.• Counting objects: Provide exact numbers of objects or entities visible, with no guesswork.• Object actions: Only describe actions that can be clearly observed in the image, with no assumptions.• Object locations: Describe precise spatial relationships between objects based on their positioning in the image.• Relative positions between objects: Indicate how objects are arranged in relation to one another, using clear visual evidence from the image. <p>Ensure that only questions with clear, confident answers are included:</p> <ol style="list-style-type: none">1. The content referenced in the question is clearly visible in the image, and the answer can be confidently provided.2. It is clear from the image that the content is not present. <p>What's more, avoid asking questions that cannot be confidently answered from the image. The conversation can include more complex questions, such as asking about the background knowledge of the objects in the image or discussing events happening in the image. However, make sure these questions are based on clear visual evidence from the image and not uncertain speculation. Ensure that the conversation is logically consistent, focuses on describing the visual information, and avoids unnecessary distractions. Keep the number of QA pairs to no more than 8 to maintain clarity.</p> <p>## Example Output Format</p> <p>Q: Is the laptop open or closed? A: Open</p> <p>Q: How many chairs are around the table? A: Four</p> <p>Q: Is there a dog visible in the image? A: No</p> <p>Q: What is the person doing? A: Reading</p>

Table 24. Prompt used for Stage1.5 SynthMultiChoice

Stage1.5-SynthMultiChoice
<pre>## Role Imagine you are an AI visual assistant tasked with analyzing a detailed description of an image. ## Goal You are guiding a user through understanding the image by asking multiple-choice and yes/no questions. Each question should help the user explore the image from a variety of perspectives. The goal is to generate a dynamic conversation that covers as many aspects of the visual content as possible, ensuring diverse and varied inquiries. ## Rule The types of questions should involve a range of topics, including but not limited to: <ul style="list-style-type: none"> Object types: What types of objects can be seen in the image? What are their characteristics? Counting objects: How many instances of a particular object or entity are visible? Object actions: Are any objects being manipulated, used, or interacted with? Object locations: Where are specific objects or entities located in the image? Relative positions between objects: How are the objects positioned in relation to one another in the scene? Background Elements: What does the background consist of? Are there any environmental features or structures visible? Lighting & Color: What types of lighting are visible? How does the lighting affect the visual atmosphere of the scene? Textures and Details: Are there any notable textures, intricate details, or small visual elements in the image? Ensure that only questions with clear, confident answers are included: 1. The content referenced in the question is clearly visible in the image, and the answer can be confidently provided. 2. It is clear from the image that the content is not present. Ensure that each question brings out a different angle of the image, covering a wide range of observations and keeping the conversation engaging and informative. The questions should vary in scope, from general descriptions to more specific details, allowing the user to explore various facets of the image. Ensure that the conversation is logically consistent, focuses on describing the visual information, and avoids unnecessary distractions. Keep the number of QA pairs to no more than 8 to maintain clarity. ## Example Output Format Q What is the object located in the center of the image? A. A vase B. A sculpture C. A pillow D. A mirror A: B Q: Is the room illuminated by sunlight? A: Yes Q: What type of floor covering is visible? A. Carpet B. Wood C. Tiles D. Concrete A: A Q: What color is the chair in the corner of the room? A. Red B. Black C. Green D. Yellow A: D Q: Is the person holding something in their hand? A: NO Q: Where is the large plant in the room positioned? A. Next to the window B. Near the door C. Beside the desk D. In the corner of the room A: C</pre>

Table 25. Prompt used for Stage2 SynthMultiChoice

Stage2-SynthMultiChoice
<p>## Role Suppose you are an AI visual assistant, viewing a detailed description of an image.</p> <p>## Goal Design a conversation where a user asks questions about an image, and the AI visual assistant responds with high precision and specificity, based solely on what is clearly visible in the image. The focus is on high-quality matching where each answer must be directly tied to the observable content in the image. The questions should ensure that the answers are clear, precise, and grounded in visible evidence, and should avoid any speculative responses.</p> <p>## Rule Key aspects to focus on:</p> <ul style="list-style-type: none">• Object types: Identify objects and details based on their distinct visual attributes, ensuring the description matches what is observable in the image.• Counting objects: Provide exact numbers of objects or entities visible, with no guesswork.• Object actions: Only describe actions that can be clearly observed in the image, with no assumptions.• Object locations: Describe precise spatial relationships between objects based on their positioning in the image.• Relative positions between objects: Indicate how objects are arranged in relation to one another, using clear visual evidence from the image. <p>Ensure that only questions with clear, confident answers are included:</p> <ol style="list-style-type: none">1. The content referenced in the question is clearly visible in the image, and the answer can be confidently provided.2. It is clear from the image that the content is not present. <p>What's more, each question must avoid ambiguity and the answers are firmly grounded in visual evidence. The conversation can include more complex questions, such as asking about the background knowledge of the objects in the image or discussing events happening in the image. However, make sure these questions are based on clear visual evidence from the image and not uncertain speculation. Ensure that the conversation is logically consistent, focuses on describing the visual information, and avoids unnecessary distractions. Keep the number of QA pairs to no more than 8 to maintain clarity.</p> <p>## Example Output Format</p> <ul style="list-style-type: none">• Q: How many people are visible in the image? A. One B. Two C. Three D. None• A: B• Q: Is there a door visible in the image?• A: Yes• Q: What is the color of the table in the image? A. Black B. Brown C. White D. Blue• A: D• Q: What is the object on the desk? A. A phone B. A laptop C. A coffee cup D. A lamp• A: C <p>Q: What is the object beside the couch? A. A small table B. A lamp C. A plant D. A book A: C</p> <p>Q: Are there any curtains visible? A: No</p>

Table 26. Prompt used for SynthContrastShort

SynthContrastShort
<pre>## Role Suppose you are an AI visual assistant, viewing a detailed description of an image. ## Goal Your primary task is to accurately construct the basic visual of the image in your mind. Based on this basic visual, you need to create two descriptions of images. Both descriptions should stem from the description I provided but must differ distinctly in aspects such as color, quantity, type of objects, or their placement and so on. Then you need to create a set of questions and answers that highlight the differences between the two images. ## Rule Instructions for Image Descriptions: <ul style="list-style-type: none"> • Image1 Description: Provide a detailed depiction of image1, focusing on visual elements such as the types of objects, their quantities, actions, locations, and their relative positions to one another. • Image2 Description: Develop a contrasting depiction of image2 by altering key visual elements. Changes can include similar objects, a different color scheme, different locations, varying quantities or types of objects, or a new arrangement. Instructions for Questions and Answers: <ul style="list-style-type: none"> • Create a set of questions and answers that effectively highlight the differences between the two images. For clarity, in QA Section, refer to them as the 'left' (Image1) and 'right' (Image2) images. • The questions should inquire about specific differences in visual aspects such as color, quantity, or arrangement of objects. • Answer the questions using a single word or phrase. ## Example Output Format Output Format: <ul style="list-style-type: none"> • Image1: The image depicts [Detailed visual description of the image1], ensuring the description is no less than 120 words and no more than 180 words. • Image2: The image describes [Detailed visual description of the image2], ensuring it clearly differs from Image 1, the description is no less than 120 words and no more than 180 words. QA Section Example: (You should focus more on the visual differences between the two images) <ul style="list-style-type: none"> • Q: What objects are present in both the left and right images? • A: Trees and a house • Q: What objects are in the left image but not in the right image? • A: A sedan car • Q: What objects are in the right image but not in the left image? • A: A bicycle • Q: How do the colors of the objects differ between the two images? • A: Left image has green trees, right image has palm trees in blue • Q: Does the left image have more trees than the right image? • A: No • Q: Is the quantity of objects different between the two images? • A: Yes • Q: Do both images feature pedestrians? • A: Yes You can design other questions to explore additional aspects, such as the context of the objects, their arrangement, any textual elements, or the mood conveyed by the colors and lighting in each image. Additionally, you should control the QA pairs not exceeding 7 pairs to preserve visual clarity. Please present the descriptions and QAs directly, without any extraneous information.</pre>

Table 27. Prompt used for SynthContrastLong

SynthContrastLong
Role Suppose you are an AI visual assistant, viewing a detailed description of an image.
Goal Your primary task is to accurately construct the basic visual of the image in your mind. Based on this basic visual, you need to create two descriptions of images. Both descriptions should stem from the description I provided but must differ distinctly in aspects such as color, quantity, type of objects, or their placement and so on. Then you need to create a set of questions and answers that highlight the differences between the two images.
Rule Instructions for Image Descriptions: <ul style="list-style-type: none">• Image1 Description: Provide a detailed depiction of image1, focusing on visual elements such as the types of objects, their quantities, actions, locations, and their relative positions to one another.• Image2 Description: Develop a contrasting depiction of image2 by altering key visual elements. Changes can include similar objects, a different color scheme, different locations, varying quantities or types of objects, or a new arrangement. Instructions for Questions and Answers: <ul style="list-style-type: none">• Create a set of questions and answers that effectively highlight the differences between the two images. For clarity, in QA Section, refer to them as the 'left' (Image1) and 'right' (Image2) images.• The questions should inquire about specific differences in visual aspects such as color, quantity, or arrangement of objects.• The answers should provide clear and concise explanations, directly referencing the visual disparities between the left and right images.
Example Output Format Output Format: <ul style="list-style-type: none">• Image1: The image depicts [Detailed visual description of the image1], ensuring the description is no less than 120 words and no more than 180 words.• Image2: The image describes [Detailed visual description of the image2], ensuring it clearly differs from Image 1, the description is no less than 120 words and no more than 180 words. QA Section Example: <ul style="list-style-type: none">• Q: What are the differences in the objects found in the left and right images?• A: The left image contains [specific objects and features in Image1], while the right image includes [specific objects and features in Image2].• Q: What is the color difference between the objects in the left and right images?• A: The left image features [description of colors in Image1], whereas the right image has [description of colors in Image2].• Q: How does the quantity of objects differ between the two images?• A: In the left image, there are [number and types of objects in Image1], compared to [number and types of objects in Image2] in the right image. You can design other questions to explore additional aspects, such as the context of the objects, their arrangement, any textual elements, or the mood conveyed by the colors and lighting in each image. Additionally, you should control the QA pairs not exceeding 7 pairs to preserve visual clarity. Please present the descriptions and QAs directly, without any extraneous information.

Table 28. Prompt used for SynthTextQA-Step1

SynthTextQA-Step1

Role

Suppose you are an AI visual assistant, viewing a detailed description of an image.

Goal

Your primary task is to accurately construct the image in your mind. Based on your understanding and construction of the basic image, you need to further expand and enrich the content of the image.

Rule

You need to add textual content to enhance the expressiveness and the visual information of the image, ensuring that the text becomes the main subject of the image and is closely related to the theme and visual elements of the image. The textual information can include:

- Words and numbers on signs
- Billboards
- Posters
- Digital displays
- Graffiti
- Slogans
- Traffic directions
- Road signs
- Historical markers
- Famous quotes

To maintain harmony:

- The image contains no more than three textual elements.
- The total word count shuold not exceed twelve words.

After adding text, you need to provide a new image description that includes the text you added. Ensure that the new description is logically consistent, overall harmonious, and focuses on describing the visual information in the image, minimizing the interference of irrelevant content. Your description must be thorough and specific enough to serve as a visual guide for an image generation model, enabling it to accurately reproduce the imagined scene.

Example Output Format

Please ensure your response is no less than 110 words and no more than 150 words, and present the description content directly, without any extraneous information. Here are some possible formats:

1. A pink heart button displays "Love Yourself" in cursive, bold font, against a pastel-colored background. On a nearby wall, a motivational poster titled "Self-Love Journey" in a modern, sans-serif font, features a stylized illustration of a person embracing their reflection. Below the poster, a small sticker on a laptop reads "You Are Enough" in a playful, handwritten font, adding a touch of whimsy to the scene. The surrounding environment is minimalistic, with a few scattered flowers and a faint, gradient glow, emphasizing the uplifting and empowering message of the image.
2. A serene and vibrant painting titled "Sweet Dreams" dominates the wall, featuring an array of colorful flowers and puffy white clouds that evoke a sense of peacefulness. The title "Sweet Dreams" is written in elegant, cursive script at the top of the painting, while a small plaque at the bottom reads "Artist: Lily Rose" in simple, yet refined font. A tiny inscription on the frame states "Dream Big" in delicate letters, further emphasizing the whimsical and soothing ambiance of the artwork.
3. A beautiful, rustic wooden sign with the words "Welcome to New Hope Church" in elegant, golden letters hangs above the entrance of a quaint, countryside church. On the church door, a smaller sign reads "Sunday Service 10am" in a simple, yet inviting font, providing essential information to visitors. In the foreground, a stone path is lined with vibrant flowers and a directional signpost that states "Peace Garden" with an arrow pointing towards a serene garden area to the side, further emphasizing the church's peaceful atmosphere and inviting all to explore its tranquil grounds.

Table 29. Prompt used for SynthTextQA-Step2

SynthTextQA-Step2

```
## Role
Suppose you are an AI visual assistant provided with a description of an image containing textual elements.
## Goal
Your task is to design a conversation between yourself and a person asking about these textual elements.
## Rule
1. Conversation Format:

- Present the conversation directly, separating questions and answers with Q and A.
- Limit the conversation to no more than 5 QA pairs to maintain clarity.


2. Tone and Style:

- Answer in the tone of a visual AI assistant observing and describing the image.
- Ensure answers are based solely on the visible content of the image.


3. Content to Include:

- All textual elements mentioned in the image description must be covered in the conversation.


4. Question Design:

- Diversify the way questions are asked, especially when they are of the same type.
- Answer the question using a single word or phrase.
- For questions focusing on specific texts, vary the format:
  - Half of the time, answer only with the text.
  - The other half, describe the text.
  - Ensure you ask about each textual element in the description.
- Examples:
  - Q: What's written on the wooden sign?
  - A: Welcome Home
  - Q: Could you describe the text on the taxi's side panel?
  - A: NYC Taxi
  - Q: What inscription is on the traditional Japanese lantern?
  - A: Konnichiwa



5. Optional Questions:

- You may include additional questions exploring aspects like the location or font style of the text, but it's not mandatory.
- You can also design other questions to explore additional aspects.
- But also answer the question using a single word or phrase.


6. Avoiding Repetition:

- For similar questions, avoid repeating the same phrasing.
- Design questions that explore different aspects to keep the conversation engaging.

```

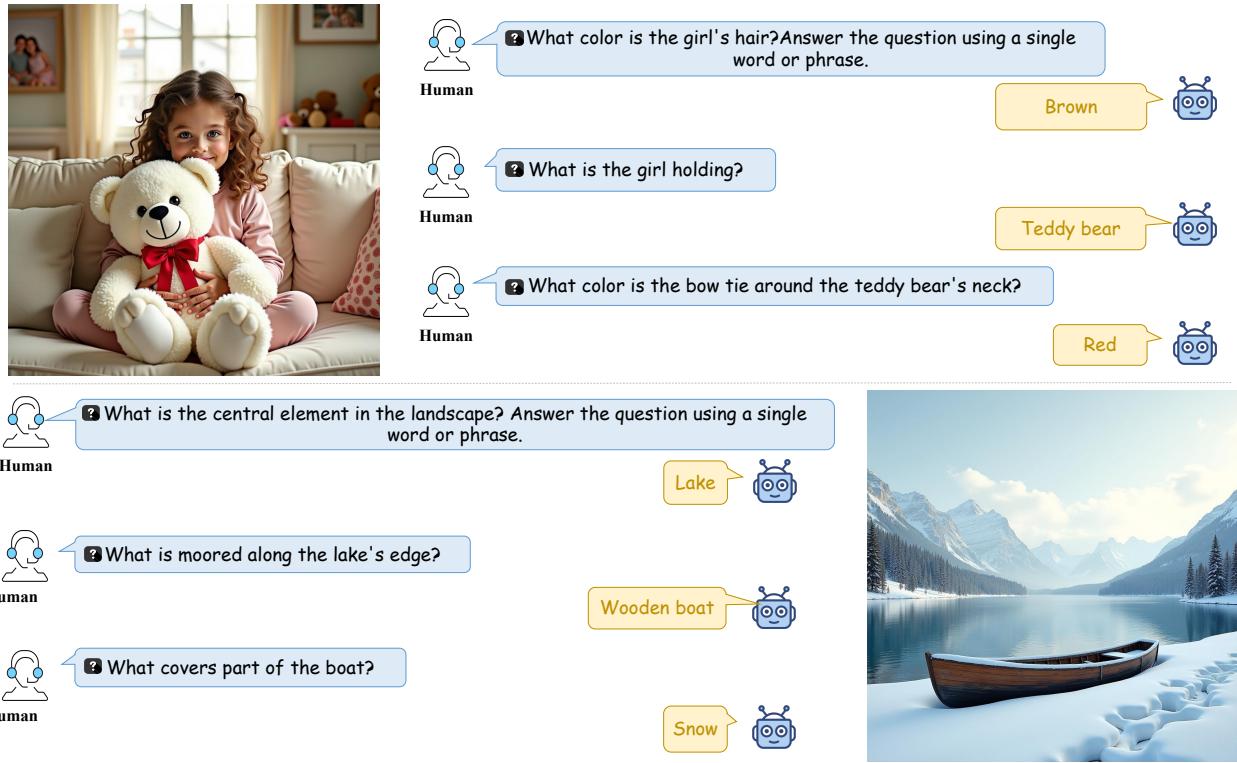


Figure 10. Illustration of our SynthConvShort

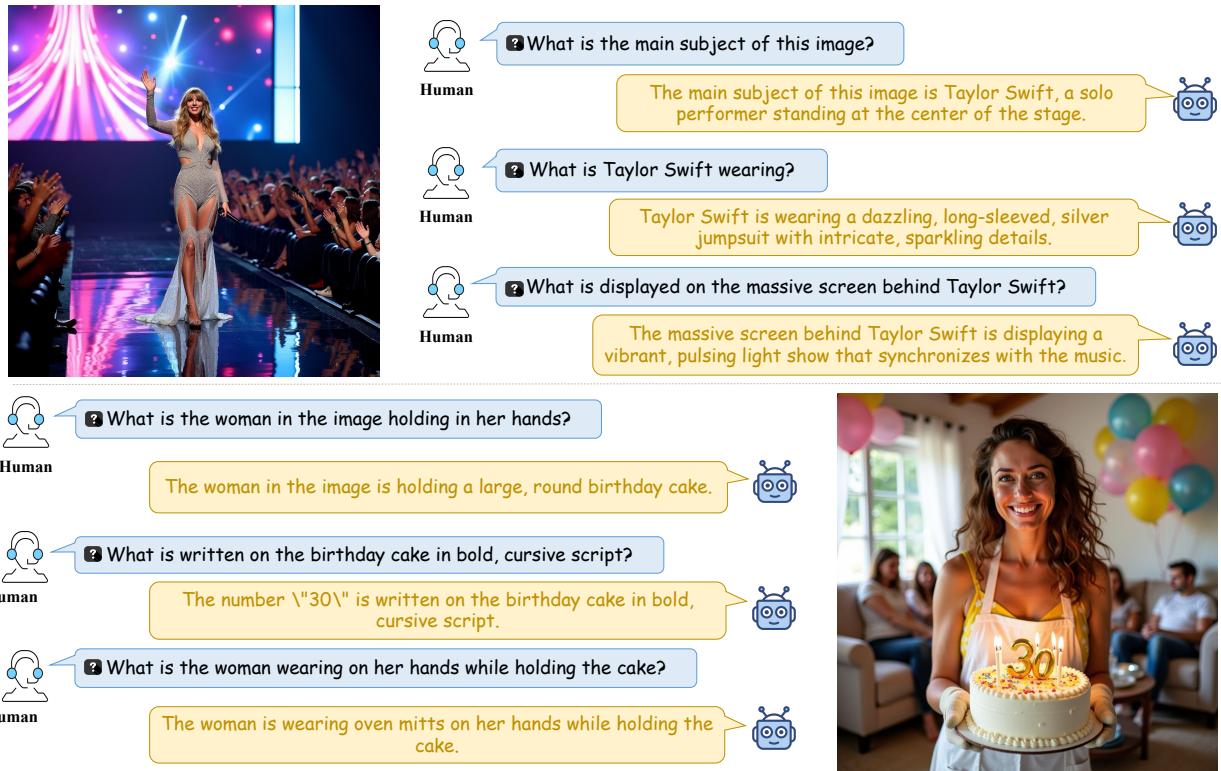


Figure 11. Illustration of our SynthConvLong

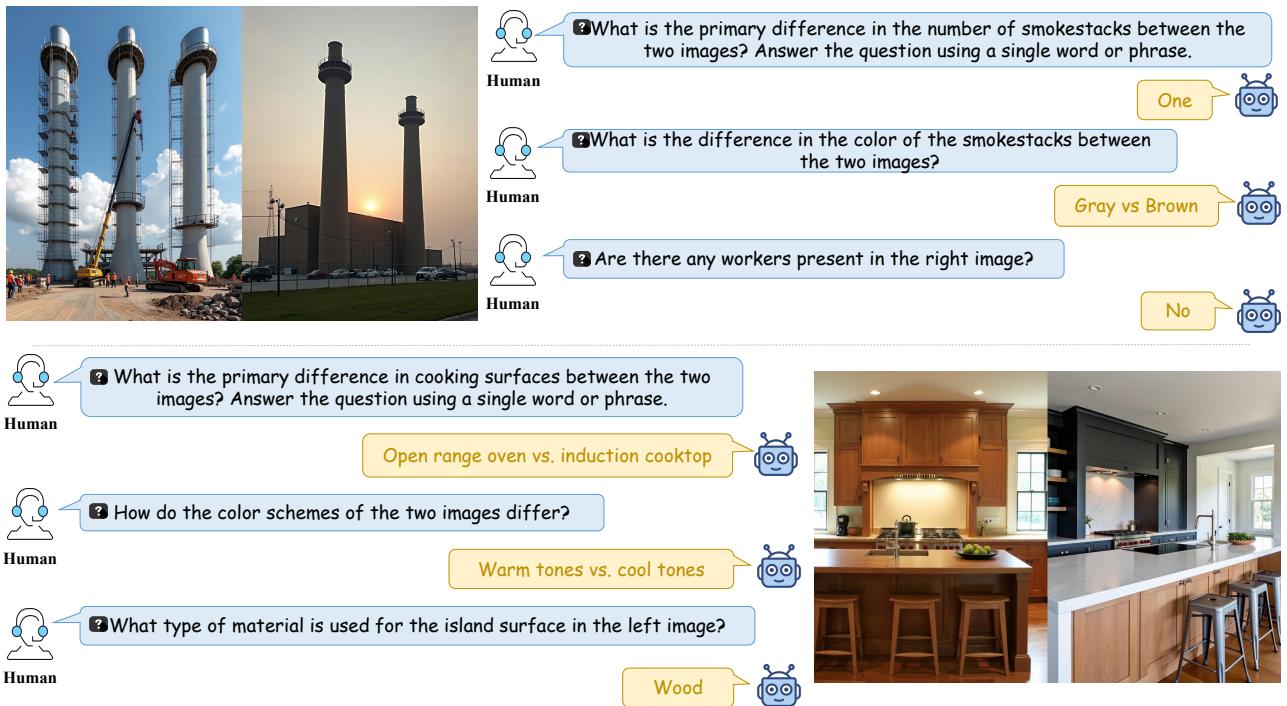


Figure 12. Illustration of our SynthContrastShort

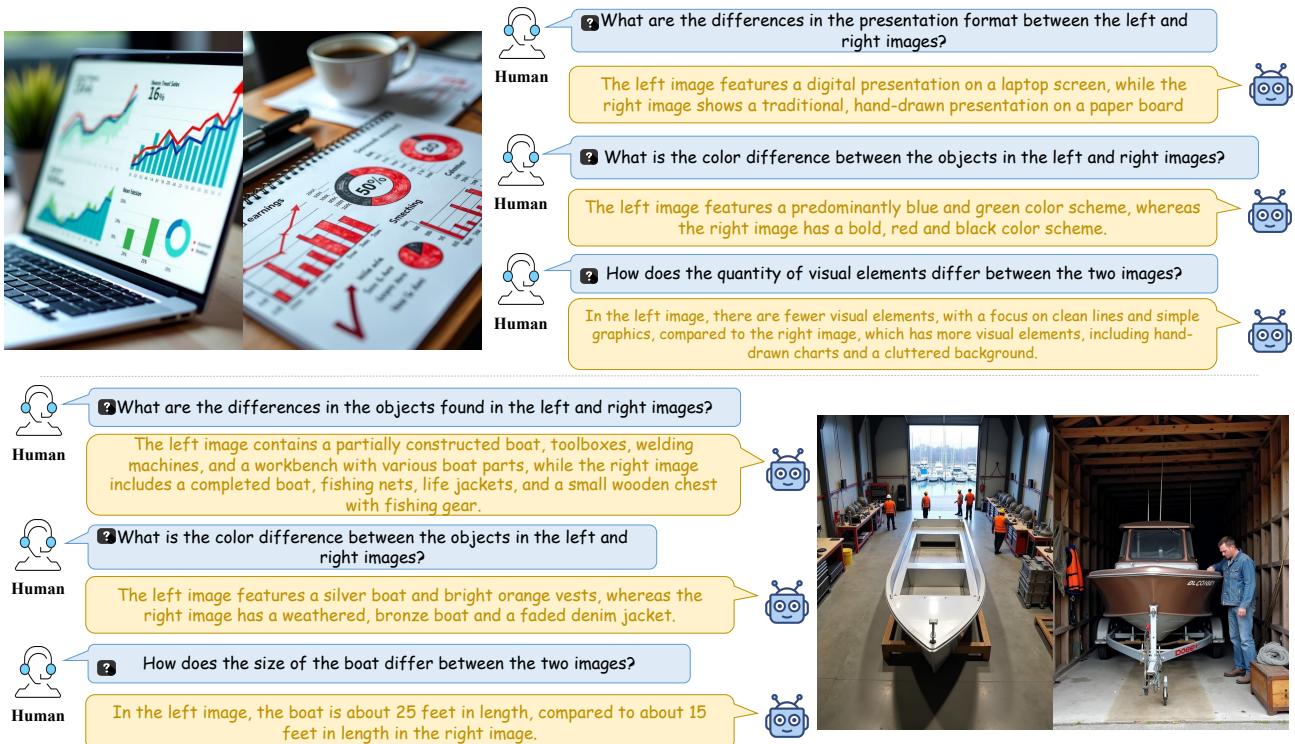


Figure 13. Illustration of our SynthContrastLong

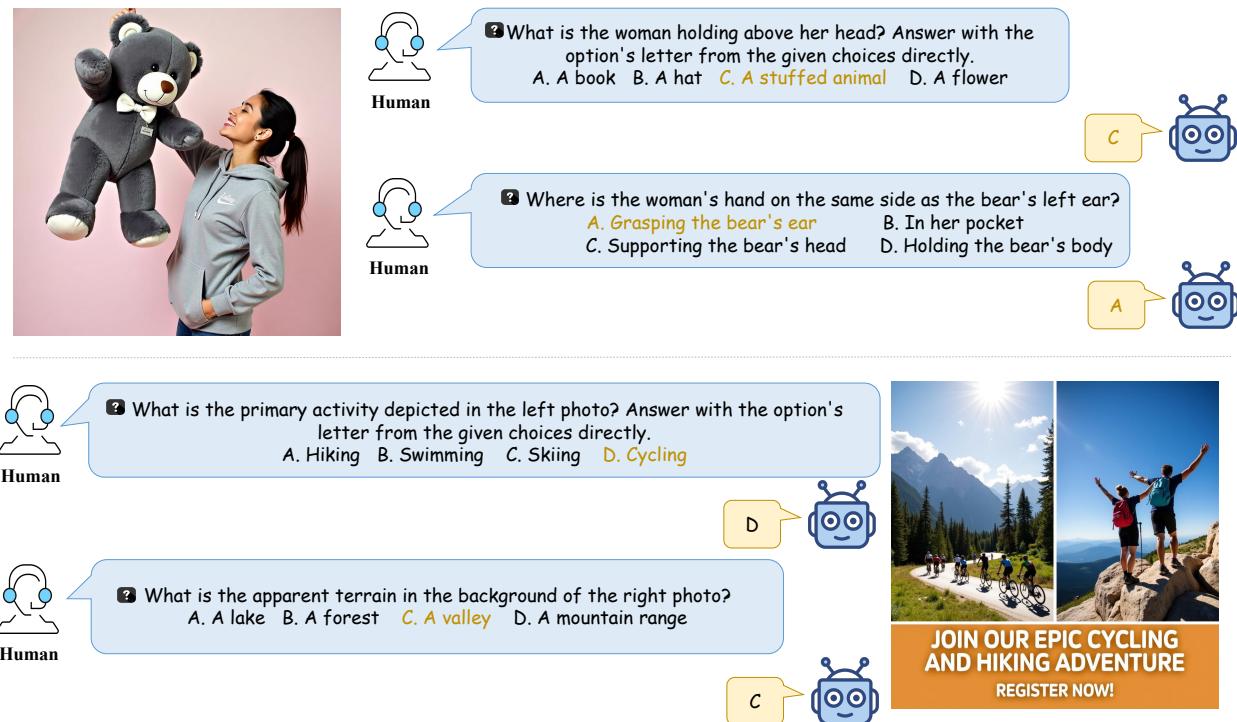


Figure 14. Illustration of our SynthMultiChoice

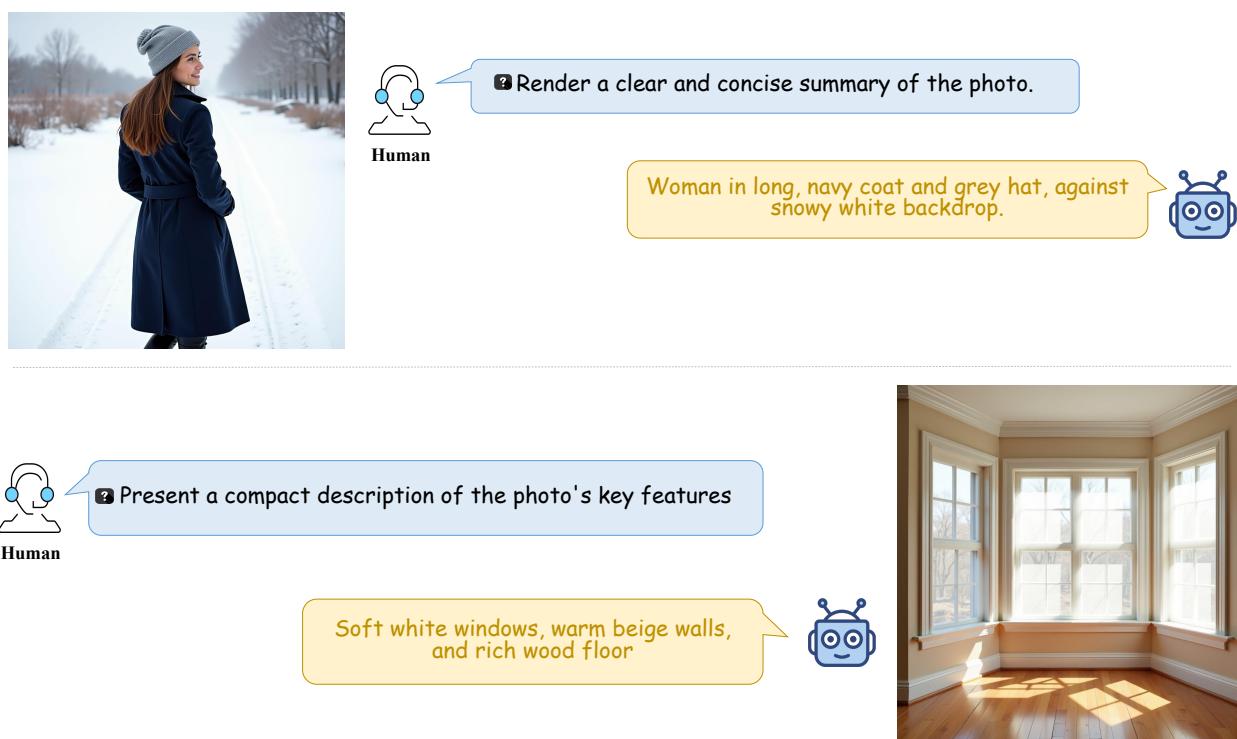


Figure 15. Illustration of our SynthColor

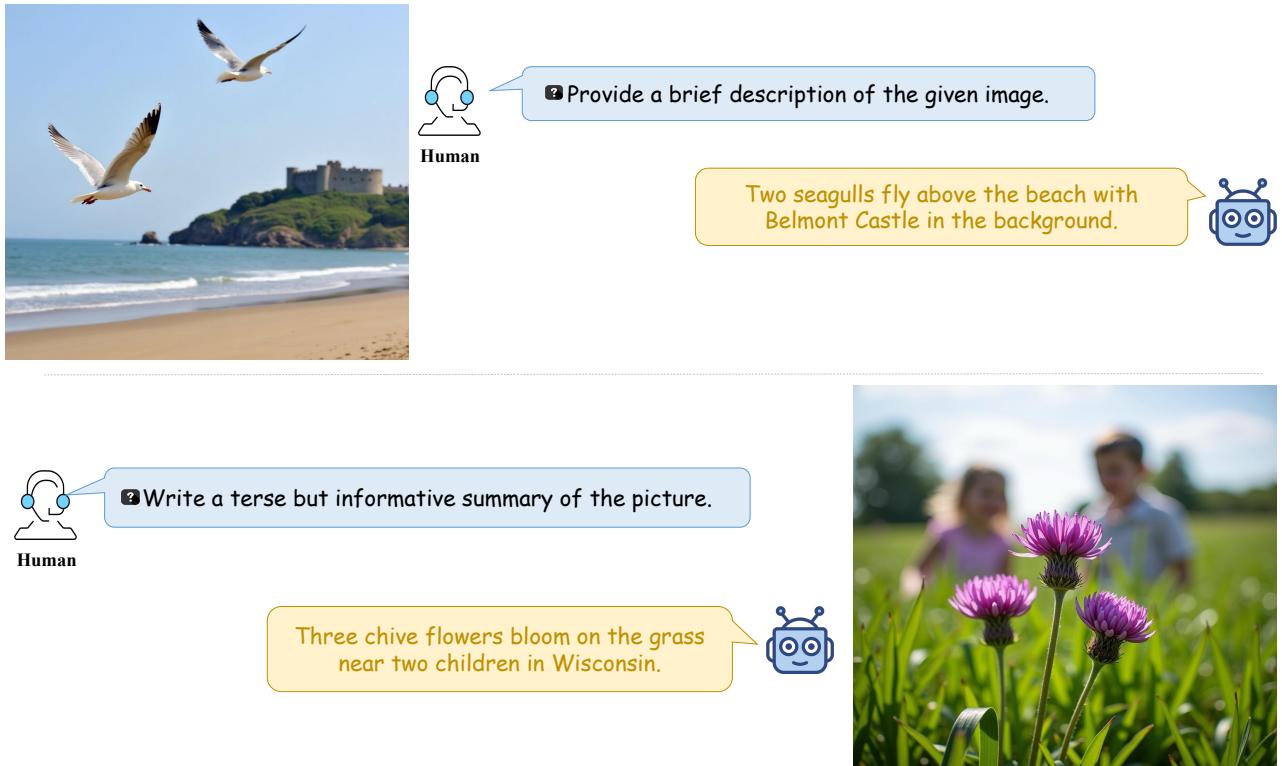


Figure 16. Illustration of our SynthCount



Figure 17. Illustration of our SynthScene

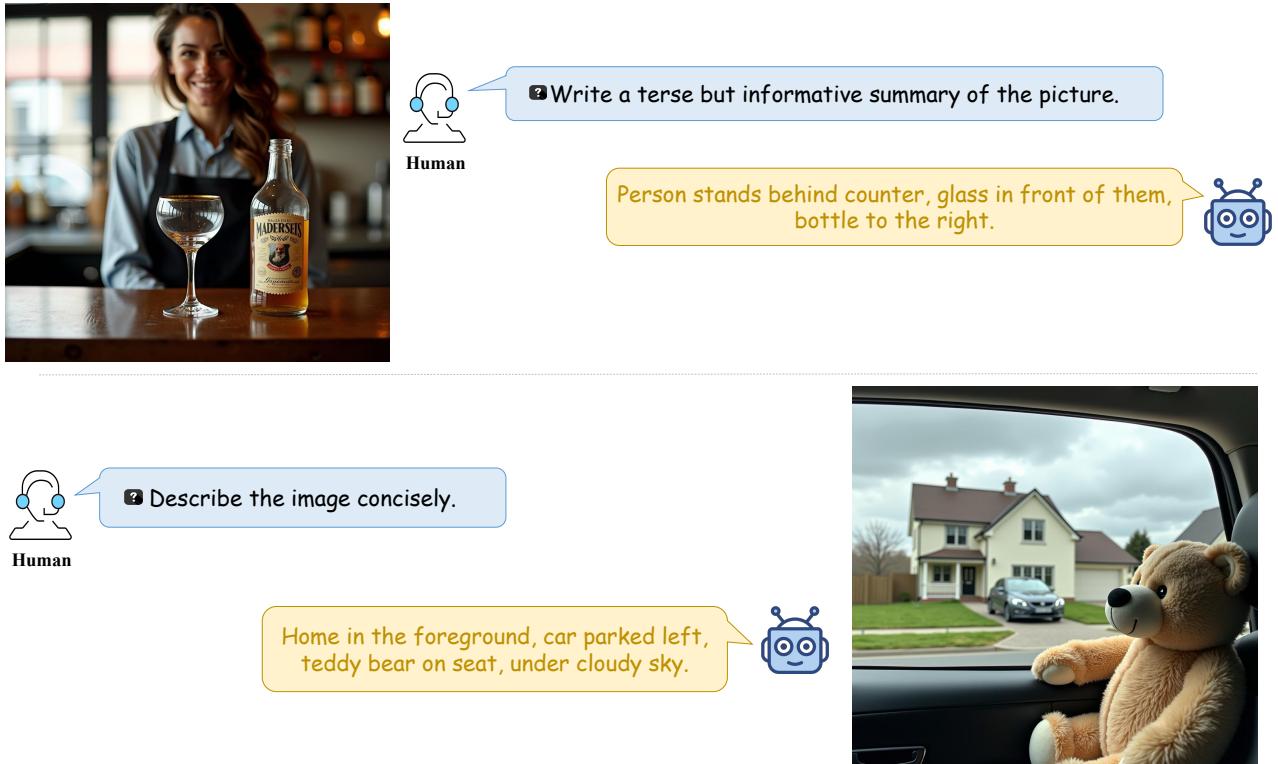


Figure 18. Illustration of our SynthSpatial

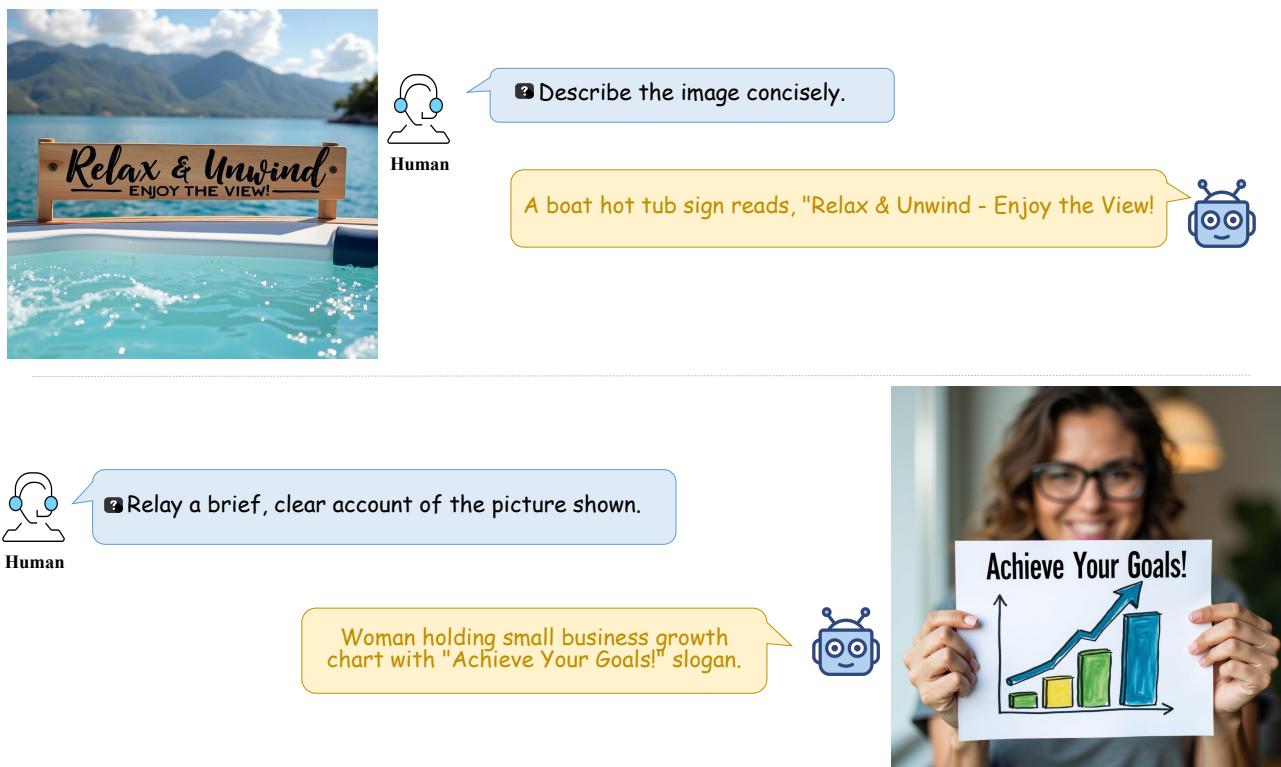


Figure 19. Illustration of our SynthText