VEN-VL: A Visual Ensemble MoE Framework for Effective and Efficient Multi-Modal Understanding

Yinghao Wu[1,2*], **Zhuoyan Luo**[2*],
Yiyao Yu[2], **Zhaojian Yu**[1,2], **Yujiu Yang**[2†], **Xiao-Ping Zhang**[1,2†]
[1]Shenzhen Key Laboratory of Ubiquitous Data Enabling, SIGS, Tsinghua University
[2]Tsinghua Shenzhen International Graduate School (SIGS), Tsinghua University
yang.yujiu@sz.tsinghua.edu.cn, xpzhang@ieee.org
Code: https://github.com/RobertLuo1/VEN-VL

Abstract

Despite the remarkable progress achieved by recent efficient methods in accelerating multimodal understanding, they still suffer from noticeable performance degradation. Their emphasis on the high compression ratio of a single visual clue and reliance on the heuristic pruning strategy with coarse attention alignment incurs a bottleneck on the information *capacity* and *density* of visual tokens. Addressing this limitation, we propose VEN-VL, a visual ensemble MoE framework for effective and efficient perception following the *enrich then compact* principle. Specifically, we first enrich the information capacity by unifying the visual representations of different perspectives, and then progressively compact it with adaptive routers in specialized visual experts to enhance the information density. Furthermore, we incorporate the reconstruction ability of vanilla structure via explicit visual supervision, facilitating crucial information preservation. Experimental results demonstrate our superiority in complex visual tasks with few information-condensed tokens, which effectively bridges the gap between performance and efficiency.

1 Introduction

Recent success of Large Language Models (LLM) (OpenAI, 2023; Grattafiori et al., 2024; Yang et al., 2025a) has facilitated a significant advancement in Large Visual-Language Model (LVLM) (Liu et al., 2023; Dai et al., 2023; Wang et al., 2024b; Zhu et al., 2023). The remarkable scalability of LVLMs, especially with Mixture of Experts (MoE) architectures (Lin et al., 2024; Wu et al., 2024; Team et al., 2025; Deitke et al., 2025; Comanici et al., 2025), demonstrates strong potential for complex visual understanding in real-world interactions. Due to the substantial computational overhead of LVLM deployment, efficiency and effectiveness are key considerations in model design.

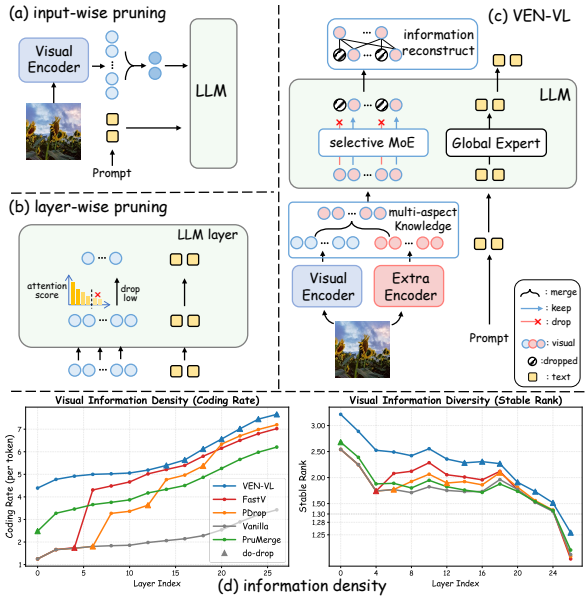


Figure 1: **Comparison with different paradigms.** Existing methods typically rely on either input-wise compression of single-aspect features (a) or heuristic layer-wise pruning based on coarse attention alignment (b), both of which suffer from the limitation of information capacity and density. In contrast, we propose **VEN-VL** (c), which first unifies multi-perspective visual representations via MKE to boost capacity, and then performs progressive refinement with dynamic routers of HTE in the LLM. As visualized in (d), **VEN-VL** demonstrates superior information density of each layer compared to previous efficient methods.

Considering the quartic complexity is highly associated with the token sequences, a straightforward solution is visual token pruning. Following this, a variety of approaches are presented to explore effective mechanisms for visual redundancy reduction while maintaining valuable tokens. It can be generally divided into two aspects in terms of their design principles. On the one hand, the Figure 1(a) depicts the methods (Li et al., 2023b; Ye et al., 2025b; Zhang et al., 2025; Shang et al., 2025; Yang et al., 2025b) that compress the single

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 22906–22920
October 24-29, 2026 ©2026 Association for Computational Linguistics

vision features into a compact representation before feeding them into the LLM. On the other hand, as shown in Figure 1(b), the counterparts (Xing et al., 2024; Chen et al., 2024a; Zhang et al., 2024; Guo et al., 2025) prune visual tokens at the layer of LLM based on the attention map.

Despite these progresses in model acceleration, they both suffer from the inevitable performance degradation. We speculate it stems from two key limitations. In terms of the *input-wise pruning*, the emphasis of the exaggerated compression rate on the single-stream vision features fails to establish the completed semantic modeling. This decreases the information capacity and diversity significantly (as indicated in Figure 1(d)), thereby becoming a bottleneck for subsequent multimodal understanding in LVLMs. For *Layer-wise pruning*, we identify that their information density for complex visual understanding still falls below the anticipation. It suggests that attention-map-based selection mistakenly drops the tokens with condensed information. It is associated with the naive information accumulation on coarse visual-text alignment, which potentially incurs shifting emphasis beyond the valuable visual tokens, *i.e.*, fine-grained semantics required for later layers to be fully comprehended. Hence, a central question arises: ***Can we enable LVLMs to realize superior performance with fewer but highly information-condensed visual tokens?***

To explore this and break through the limitations above, the *capacity* and *density* of visual information are supposed to be jointly considered. Intuitively, the former determines the upper bound of the model's multimodal understanding performance, whereas the latter plays a pivotal role in the precise and efficient visual perception. Therefore, we conduct a pilot study to quantify them via stable rank (Rudelson and Vershynin, 2007) and coding rate (Yu et al., 2020) in Figure 1(d) respectively, and present **VEN-VL**, a **V**isual **EN**semble MoE framework for effective and efficient **V**isual-**L**anguage understanding. To achieve improvement both on the information capacity and density, we formulate the design of **VEN-VL** following the *enrich then compact* principle. Firstly, as described in Figure 1(c), instead of simply compressing on the single stream visual feature, the Multi-Aspect Knowledge Ensemble (MKE) is leveraged for *enriching* semantic representations by different perspectives of visual clues with structure-aware compression. It not only enhances the comprehensive representations of the given visual cue but also suppresses its redundancy, where the *capacity* and diversity of valuable visual information can be largely boosted. Secondly, as for the information density, we propose Hierarchical Token Ensemble (HTE), which seamlessly integrates the inherent specificity of MoE to *compact* the visual distribution progressively. With the design of the adaptive router to dynamically decouple intricate semantics into different experts for fine-grained processing and selection, **VEN-VL** demonstrates superiority in information *density*, as presented in Figure 1(d). Moreover, to further provide indicative guidance on the token selection process, we introduce the Structure Information Preservation (SIP). SIP embodies the model with reconstruction ability on the pruned token via semantic propagation then supervision, thus ensuring the retained tokens restore vanilla condensed information to a great extent.

We conduct extensive experiments on prevalent multi-modal benchmarks, *e.g.*, MMBench (Liu et al., 2024b), TextVQA (Singh et al., 2019) and MME (Fu et al., 2024), etc. Experimental results demonstrate that **VEN-VL** excels at complex visual understanding tasks compared to previous efficient and MoE-based LVLMs with merely **7.5**% tokens remained, which effectively bridges the gap between performance and efficiency. Moreover, we further validate the strong compatibility of our design with the current MoE framework (Team et al., 2025), demonstrating its considerable potential for scaling to larger models. Overall, our contributions can be summarized as follows:

- Based on the observation that the information *capacity* and *density* play an essential role in bridging the gap between performance and efficiency, we propose the *enrich then compact* principle that guides the design of the visual ensemble MoE framework for visual-language understanding. **VEN-VL** first formulates an enriched and compact visual space (MKE), and then progressively refines the visual distribution by adaptive routers (HTE).

- We introduce the Structure Information Preservation (SIP) to provide guidance on the token selection process via assigning the reconstruction ability of the original visual cues.

- **VEN-VL** achieves superior performance in complex multi-modal understanding tasks, *e.g.*, $+8.1\%$ in TextVQA and $+5.2\%$ in MMB

with fewer but more information-condensed visual tokens ($\sim 10\%$).

2 Related Work

Efficient LVLM. Efficiency is one of the key topics in LVLM, where the design can be divided into two main categories. 1) *Input visual representations compression*: early methods (Dai et al., 2023; Du et al., 2022; Zhu et al., 2023; Li et al., 2023b) leverage Q-Former (Li et al., 2023b) to compress tokens into fixed-length 1D learnable queries. VoCo-LLaMA (Ye et al., 2025b) utilizes LLM for distillation, which converts the visual features into one specific token. LLava-Mini (Zhang et al., 2025) introduces the modality pre-fusion module for visual-text interaction, achieving minimal tokens fed into the LLM backbone. LLaVA-PruMerge (Shang et al., 2025) and VisionZip (Yang et al., 2025b) directly perform redundancy reduction on the input visual features. Obviously, they overemphasize the exaggerated compression ratio, which faces the challenges in modeling the valuable spatial properties of the original visual representations and limited semantic information. 2) *Inner visual token reduction*: These methods (Chen et al., 2024a; Xing et al., 2024; Ye et al., 2025a; Guo et al., 2025) typically perform similar token pruning mechanisms relying on the attention map of each layer in LVLM. However, we observe that the pre-defined threshold-based token selection without guidance is coarse, which exhibits vulnerability in handling diverse visual scenarios, leading to inevitable information density loss.

Mixture of Experts in LVLM. To unleash the scaling potential of LVLM and simultaneously avoid extensive computational overhead, Mixture of Experts (MoE) (Shazeer et al., 2017; Eigen et al., 2013) is adopted to enable routers for specific parameter activations. Prior work LLaVA-MoE (Lin et al., 2024) advances LLava (Liu et al., 2023) with the proposed MoE-Tuning strategy, empowering the visual-text understanding capacity. Moreover, several approaches (Team et al., 2025; Wu et al., 2024; Deitke et al., 2025) demonstrate a crucial step forward in MoE-based LVLM training. Although the MoE largely boosts the performance with sparse activated parameters, it still incurs huge memory consumption considering the quartic costs of the visual tokens. Therefore, we endeavor to excavate the potential of fewer but information-dense visual tokens by incorporating the natural selective properties of MoE, which leaps towards bridging the performance and efficiency gap.

3 Method

We introduce **VEN-VL**, a MoE framework with comprehensive visual ensemble. As depicted in Figure 2, it is composed of three main components. Following the *enrich then compact* principle, we first present Multi-Aspect Knowledge Ensemble, termed as MKE in Section 3.2, which enrich semantic diversity while alleviating the redundancy of the visual representation. Subsequently, based on the knowledge-enhanced input, the hierarchical visual token ensemble strategy in Section 3.3 incorporates the adaptive router in MoE to improve the information density of visual tokens gradually. Finally, in Section 3.4, we elaborate on the proposed structural information preservation.

3.1 Architecture

The overall architecture of **VEN-VL** is presented in Figure 2. Given an image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the original height and width. The Multi-Aspect Knowledge Ensemble (MKE) with a dual-branch visual encoder is to extract different perspectives of vision features, *i.e.*, semantic-centric $\mathcal{F}_{main} \in \mathbb{R}^{H_1 \times W_1 \times D_1}$ and fine-grained details-amplified $\mathcal{F}_{extra} \in \mathbb{R}^{H_2 \times W_2 \times D_2}$. To establish a unified visual space while suppressing the redundancy, MKE then performs merging mechanisms to generate a joint embedding $\mathcal{V} \in \mathbb{R}^{L_v \times D}$, where D represents the hidden size of LLM. Subsequently, Hierarchical Token Ensemble (HTE) is applied to take $\mathcal{V}$ as the input and increase the information density of visual tokens progressively via incorporating adaptive routers of MoE. By hierarchical compacting the visual tokens, it ultimately derives pruned but semantic-condensed tokens $\mathcal{V}^c \in \mathbb{R}^{L_v^c \times D}$ where $L_v^c \ll L_v$, for multi-modal understanding. Further, to consolidate the ensemble process of the visual tokens, Structural Information Preservation (SIP) is proposed to reconstruct the original structure of the input feature $\mathcal{V}$ by semantic propagation with visual supervision.

3.2 Multi-Aspect Knowledge Ensemble

A family of existing efficient methods (Shang et al., 2025; Zhang et al., 2025; Ye et al., 2025b) focuses on the large compression ratio of the visual representation from a single stream, *e.g.*, CLIP. Despite the acceleration of the whole pipeline, it struggles with the complex multi-modal understanding task,

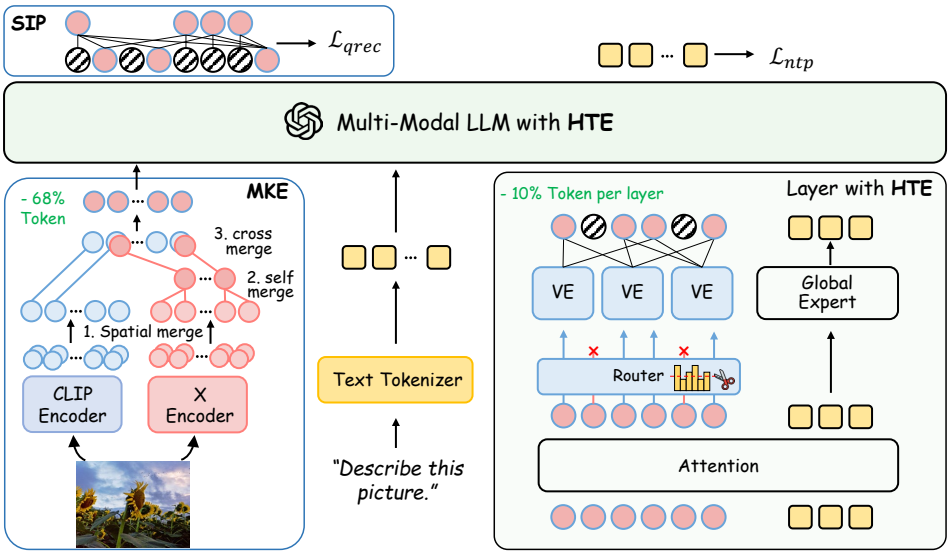


Figure 2: **Overview of VEN-VL.** The model incorporates three components: MKE, HTE, and SIP. The MKE first extracts visual features of different perspectives and ensembles them into a compact and unified representation through spatial- and cross-merging, which enhances semantic diversity and reduces redundancy. Subsequently, the HTE benefits from the inherent specialty of MoE to perform fine-grained token selection with dynamic routers. It progressively refines the visual distribution across layers, gradually increasing the information density. Finally, SIP is used to provide inductive guidance on the reduction process, which empowers the model with reconstruction ability of the structural information.

especially the one that necessitates diverse interpretations of visual cues. The rationale behind falls on the significant degradation of the information capacity and diversity. Therefore, we propose the Multi-Aspect Knowledge Ensemble (MKE) with two key incentives to tackle the issues. First, considering that CLIP exhibits insufficient fine-grained details modeling due to the optimization on image-level visual-linguist alignment, we incorporate an additional visual encoder with heterogeneous representation to enable comprehensive semantics activation. Moreover, we observe that the aforementioned approaches employ a new set of learnable embeddings for information integration, disregarding the valuable structural information. Instead, as illustrated in Figure 2, we strive to maintain the vanilla visual relationship and suppress the redundancy with structural aware compression. With sequential merging on the visual representations of different perspectives, it well establishes the compact and unified visual space.

Specifically, given an input image $\mathcal{I}$, we leverage CLIP to derive the semantic-aligned features $\mathcal{F}_{main} \in \mathbb{R}^{N\times D}$. Simultaneously, we enrich the information capacity with the representation $\mathcal{F}_{extra} \in \mathbb{R}^{N\times D}$ at a fine-grained level, which is extracted from the auxiliary encoder. Although the dual-stream features primarily emphasize different aspects, the direct concatenation of them fails to formulate a complete holistic semantics due to the substantial noise accumulation and huge computational overhead. In order to ensemble the crucial information seamlessly with efficiency and avoid the overwhelming disruption by the redundancy, we adopt a structure-aware compression strategy with sequential merging. Firstly, we compress on the spatial axis of both visual representations, which aims at transforming the locality into channel dimension. It initially alleviates the redundancy in the *local-level* while preserving the original 2D grid topology, which can be formulated as:

$$\hat{\mathcal{F}} = \Psi(\mathcal{F}, r) \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times(C\cdot r^2)}, \tag{1}$$

where Ψ denotes the transformation and r is the downsampling ratio. Then, as we intend to retain the critical structure information and facilitate the holistic understanding of the visual cues, we keep $\mathcal{F}_{main}$ unchanged and employ self/cross-merge on the $\mathcal{F}_{extra}$ for redundancy suppression and semantic complementary in the *global-level*. More detailedly, inspired by (Bolya et al., 2022; Yin et al., 2022; Kong et al., 2022), we introduce a bipartite merge operator $\Phi(\mathbf{X}, \mathbf{Y}, k)$. It divides the input into a source set $\mathbf{Y}$ and a target set $\mathbf{X}$, then identifies the k edges with the top k cosine similarity $\cos(\mathbf{x}_i, \mathbf{y}_j)$, and aggregates the matched pairs via

weighted averaging:

$$\Phi(\mathbf{X}, \mathbf{Y}, k) = \alpha \mathbf{x}_i + (1-\alpha)\mathbf{y}_j \quad (2)$$

With the pre-defined operator, we decouple $\hat{\mathcal{F}}_{extra}$ into two disjoint features $\hat{\mathcal{F}}^A_{extra}$ and $\hat{\mathcal{F}}^B_{extra}$, and obtain the more compact representation:

$$\tilde{\mathcal{F}}_{extra} = \Phi(\hat{\mathcal{F}}^A_{extra}, \hat{\mathcal{F}}^B_{extra}, k_{self}) \quad (3)$$

Finally, it is further projected into $\mathcal{F}_{main}$ via cross-merge for semantic complementary, which amplifies the original information with fine-grained visual cues:

$$\mathcal{V} = \Phi(\hat{\mathcal{F}}_{main}, \tilde{\mathcal{F}}_{extra}, k_{cross}) \quad (4)$$

In this manner, we formulate the unified and compact visual space $\mathcal{V}$ by semantic ensemble with structure-aware compression, significantly enriching the information capacity and diversity.

3.3 Hierarchical Token Ensemble

While MKE establishes a semantic-***enriched*** visual space at the input level, ***compacting*** these visual tokens efficiently and effectively within the deep layers of the LLM afterwards remains critical. Existing layer-wise pruning methods typically rely on attention maps for token selection. However, this approach may mistakenly drop tokens with condensed information due to the naive visual-text alignment in the early layer, *e.g.*, those corresponding to small objects or subtle texture representations. To address this issue, we propose the Hierarchical Token Ensemble (HTE), which benefits from the inherent specificity of the MoE architecture to decouple intricate semantics and progressively refine the visual distribution by the selective mechanism. Specifically, as shown in Figure 2, we first adopt a specialized visual MoE design to process the fine-grained semantics separately without deteriorating the general reasoning capabilities of LLM. Then, we retain the original pre-trained FFN as the Global Expert (E_{global}), which operates on both visual and textual tokens to maintain general semantic knowledge. Simultaneously, we introduce a bank of lightweight experts $\mathcal{E} = \{E_k\}_{k=1}^N$, which are activated exclusively for visual tokens to capture specific visual semantics. Supposing that $\mathbf{x} \in \mathbb{R}^D$ denotes the input hidden state. The output $\mathbf{y}$ integrates the dense global processing with sparse visual expertise:

$$\mathbf{y} = \mathbf{x} + E_{global}(\mathbf{x}) + \mathbb{I}_{vis}(\mathbf{x}) \cdot \sum_{i \in \mathcal{K}} w_i(\mathbf{x}) E_i(\mathbf{x}) \quad (5)$$

where $\mathbb{I}_{vis}$ is the modality indicator, and $\mathcal{K}$ signifies the set of selected indices corresponding to the experts. The routing weights $w(\mathbf{x})$ and indices $\mathcal{K}$ are derived via a Top-k gating mechanism:

$$w(\mathbf{x}) = \text{Softmax}(\mathbf{x}W_r), \quad \mathcal{K} = \text{TopK}(w(\mathbf{x})), \quad (6)$$

where $W_r \in \mathbb{R}^{D \times N}$ represents the router projection. It dynamically guides visual tokens to the experts which are best suited for the particular semantic content.

It is worth noting that the utilization of routers is not only for computing the token distribution, but also serves as a judge for token saliency. We observe that the routing activation reflects the distinctiveness of the visual information within a token. On one hand, tokens that exhibit a high saliency score, *i.e.*, $w_i(\mathbf{x})$, for specific experts necessitate oriented processing to fully interpret the intricate, fine-grained semantics. In contrast, that is typically neglected by the previous methods that rely on the coarse attention map for token pruning. On the other hand, tokens with low saliency scores are likely to contain trivial or redundant information that has already been captured by earlier layers.

Therefore, A reduction-gated mechanism to refine the visual distribution is leveraged. We define the importance score $s(\mathbf{v})$ of a visual token $\mathbf{v}$ as its maximum routing activation:

$$s(\mathbf{v}) = \max_{k \in \mathcal{K}} g_k(\mathbf{v}) \quad (7)$$

At specified layers, we apply a threshold τ for pruning. The set of retained visual tokens $\mathcal{V}_{next}$ for the subsequent layer is denoted as:

$$\mathcal{V}_{next} = \{\mathbf{v} \in \mathcal{V}_{curr} \mid s(\mathbf{v}) > \tau\} \quad (8)$$

By filtering tokens dynamically with their saliency instead of coarse attention alignment, HTE progressively increases the information density of the visual sequence across layers. This ensures that the deep-layer computation is concentrated solely on the highly information-condensed tokens required for precise multi-modal understanding.

3.4 Structure Information Preservation

As the visual modality contains a wide spectrum of features including spatial relationships, textures, and complex compositions, the visual token selection strategy should maximize the preservation of information capacity and density. However, we notice that existing efficient methods (Xing et al.,

2024; Yang et al., 2025b) tend to perform layer-wise token pruning in a heuristic manner. That is simply relying on the attention map as the only criterion for reduction. Yet, it exhibits vulnerability in information maintenance and even exacerbates the loss due to the lack of explicit visual supervision.

Therefore, we propose Structure Information Preservation (SIP) to provide inductive guidance on the token selection process. It endows the model with reconstruction ability of the original structure $\mathcal{V}$ via semantic propagation, which ensures the remaining visual tokens largely preserve vanilla condensed information. Inspired by (Xiao et al., 2024; Huang et al., 2022; Nikolentzos and Vazirgiannis, 2020), we adopt a parameter-free approach to avoid the model developing excessive reliance on the learnable parameters *e.g.*, cross attention or self-attention mechanism, for reconstruction adaptation instead of concentrating on the information-condensed visual tokens. Specifically, we first reassign the pruned token embedding $\mathcal{V}^p$, with RoPE for structural property initialization. Then, the remaining information-condensed tokens $\mathcal{V}^c$ iteratively propagate the holistic semantics to $\mathcal{V}^p$ in a shared space until convergence. Formally, at 0-th iteration, the affinity $\mathcal{Z}$ is formulated by scaled dot product similarity: $\mathcal{Z} = \lambda \mathcal{V}^c_{(0)} (\mathcal{V}^p_{(0)})^\top$, where λ denotes the scaling factor. To mitigate the noise accumulation during interaction and prevent causing redundancy, the top-k activations with high saliency scores are selected:

$$\tilde{\mathcal{Z}} = \begin{cases} \mathcal{Z}, & i \in \mathrm{TopK}(\mathcal{Z}), \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

While at t iterations, $\mathcal{V}^c_{(t)}$ is updated jointly by the original representation $\mathcal{V}^t_{(0)}$ and structure-aware token embedding $\mathcal{V}^p_{(t-1)}$ from previous iteration:

$$\mathcal{V}^c_{(t)} = \omega\, \mathrm{Norm}(\tilde{\mathcal{Z}})^\top \mathcal{V}^p_{(t-1)} + (1-\omega)\, \mathcal{V}^c_{(0)}, \tag{10}$$

Subsequently, it is projected into $\mathcal{V}^p_{(t)}$:

$$\mathcal{V}^p_{(t)} = \omega \tilde{\mathcal{Z}} \mathcal{V}^c_{(t)} + (1-\omega) \mathcal{V}^p_{(0)}, \tag{11}$$

where ω signifies the level of semantic propagation. After t iterations, we obtain the reconstructed visual representation $\mathcal{V}^p$ that retains the original structural information. To further provide guidance on refining the visual distribution Section 3.3 for precise information preservation, we incorporate visual supervision between the $\mathcal{V}^p$ and $\mathcal{V}$ via fine-grained alignment:

$$\mathcal{L}_{qrec} = \sum || \left(\mathrm{Quant}(\mathcal{V}^p) - \mathrm{Quant}(\mathcal{V})\right) ||_2, \tag{12}$$

where $\mathrm{Quant}\,()$ indicates quantization (Mentzer et al., 2023), which is utilized to prevent overwhelmingly regressing the exact value (Wang et al., 2024a). With guidance, the model is encouraged to accumulate structural information more effectively and to alleviate visual redundancy, thereby improving both information capacity and density.

4 Experiment

4.1 Experiment Setup

Detailed descriptions of the implementation, training strategies, along with the datasets and benchmarks, are provided in the Appendix Section A.1.

4.2 Main Result

Comparison with the SoTA LVLMs. To validate the effectiveness of **VEN-VL**, we conduct a comprehensive evaluation across diverse benchmarks. As shown in Tab. 1, **VEN-VL** consistently outperforms state-of-the-art efficient and lightweight models with significantly fewer visual tokens. Specifically, compared to the strong baseline TinyLLaVA (implemented with Qwen2.5-1.5B), **VEN-VL** achieves substantial gains of +8.1% in TextVQA, +5.2% in MMB, and +6.8% in MMB^{CN}, despite retaining only ∼10% of the visual tokens. This highlights a significant leap towards bridging the gap between performance and efficiency. Moreover, the results across varying LLM backbones demonstrate the robust scalability of our framework. With Qwen3-1.7B, **VEN-VL** achieves state-of-the-art results among previous methods with comparable model scale.

Comparison with previous token pruning methods. We make a comprehensive comparison with previous efficient methods of two representative categories under fair settings, *i.e.*, implementation with Qwen3-0.6B. We adopt FLOPs to quantify the computational overhead, which is measured by a single forward pass with 2880 visual tokens and 100 text tokens. As shown in Tab. 2, previous methods sacrifice performance for the large compression ratio, where reducing tokens to 1/8 leads to significant performance drops. In contrast, **VEN-VL** breaks through the dilemma by jointly considering the information *capacity* and *density*. Although

Table 1: **Benchmarked results with SoTA MLLMs**. Compared with previous efficient LVLM methods, *e.g.*, token-pruning based and lightweight design, our **VEN-VL** demonstrates superiority in versatile visual understanding tasks with fewer but information-condensed tokens across different scales of LLM backbone. Optimal and sub-optimal performance is signified by **Bold** and underline. Note that, †: denotes the reproduced results using the official code. P/Sum indicates perception scores and the total of perception and cognition metrics.

Method	LLM	Retain Ratio (%)	Image Question Answering				Benchmarks					
			GQA	SciQA	TextVQA	AI2D	$SEED^{I}$	MME(P/Sum)	MMB	MMB^{CN}	POPE	MMMU
LLaVA-1.5	Vicuna-7B	100%	62.0	66.8	58.2	56.4	66.1	1510/1862	64.3	58.3	85.9	34.4
Qwen-VL-Chat	Qwen-7B	100%	57.5	68.2	61.5	57.7	65.4	1488/1848	60.6	56.7	–	35.9
MiniCPM-V2	MiniCPM-2.4B	100%	–	80.4	74.3	64.8	–	–/1788	69.4	65.9	86.6	39.6
InternVL2-2B	InternLM2-1.8B	100%	–	94.1	73.5	74.4	71.6	–/1863	73.4	70.9	85.3	34.9
Qwen2-VL-2B	Qwen2-1.5B	100%	–	78.1	**79.9**	74.1	71.5	–/1890	74.9	–	78.8	**41.7**
Large Visual-Language Model With Token Pruning												
PruMerge	Vicuna-7B	∼ 5%	–	68.5	56.0	–	–	1350/–	60.9	–	76.3	–
PruMerge+	Vicuna-7B	∼ 25%	–	68.3	57.1	–	–	1462/–	64.9	–	84.0	–
LLaVA-Mini	Vicuna-7B	∼ 10%	61.8	69.7	59.1	–	60.2	1477/–	67.5	–	85.3	–
FastV	Vicuna-7B	∼ 10%	46.1	51.1	47.8	–	51.9	–/1256	48.0	–	48.0	34.0
SparseVLM	Vicuna-7B	∼ 10%	52.7	62.2	51.8	–	51.1	–/1589	56.2	–	75.1	32.7
PyramidDrop	Vicuna-7B	∼ 10%	47.5	69.0	50.6	–	64.3	–/1561	58.8	58.5	86.0	–
VisionZip	Vicuna-7B	∼ 10%	55.1	69.0	55.5	–	53.4	–/1676	60.1	–	77.0	36.2
Large Visual-Language Model with Lightweight Design												
TinyLLaVA†	Qwen1.5-0.5B	100%	57.4	60.9	47.4	–	–	1196/–	55.0	52.4	83.7	31.6
TinyLLaVA†	Qwen2.5-0.5B	100%	58.3	59.1	49.2	–	–	1230/–	58.9	54.2	86.1	33.6
TinyLLaVA†	Qwen1.5-1.8B	100%	55.5	65.3	47.7	–	–	1224/–	57.1	55.5	83.4	34.1
Bunny-2B	Qwen1.5-1.8B	100%	59.6	64.6	53.2	–	–	1300/–	59.1	58.5	85.8	–
Imp-2B	Qwen1.5-1.8B	100%	61.9	66.1	54.5	–	–	1304/–	63.8	61.3	86.7	–
Mini-Gemini-2B	Gemma-2B	100%	60.7	63.1	56.2	–	–	1341/1653	59.8	51.3	85.6	31.7
MoE-LLaVA-2B	Qwen-1.5-1.8B	100%	61.5	63.1	48.0	–	–	1292/–	59.7	57.3	87.0	–
LLaVA-MOD	Qwen1.5-1.8B	100%	58.7	68.0	58.5	–	–	1352/–	66.3	61.9	87.0	–
LLaVA-MOD	Qwen2-1.5B	100%	58.8	69.2	59.9	–	–	1384/–	68.9	64.4	87.2	–
TinyLLaVA†	Qwen2.5-1.5B	100%	**62.0**	72.0	57.4	–	–	1450/–	68.6	63.0	85.5	37.0
VEN-VL	Qwen3-0.6B	∼ 7.5%	58.8	84.2	61.0	68.3	69.0	1320/1614	66.5	64.5	**87.5**	35.6
VEN-VL	Qwen2.5-1.5B	∼ 10%	60.4	90.5	65.5	74.6	70.3	1401/1736	74.1	69.8	85.5	37.9
VEN-VL	Qwen3-1.7B	∼ 7.5%	60.6	**94.5**	73.7	**76.9**	**72.5**	**1504/1984**	**76.6**	**76.5**	86.0	39.7

Table 2: **Comparison with previous token pruning methods across LVLMs.** Comprehensive evaluation with previous token pruning based LVLMs under fair settings, *i.e.*, all adopt Qwen3-0.6B as the LLM backbone with the identical training data. Benefiting from the enriched and dense visual representation, **VEN-VL** achieves superior performance with significantly fewer tokens. Vanilla refers to the standard LLaVA (Liu et al., 2023) implementation. Tokens denote the ratio of retained visual tokens.

Method	Token↓	FLOPs(G)↓	MMMU	MME	MMB	AI2D	TextVQA	GQA	$SEED^{I}$	Avg.↑
Vanilla	100%	2330.9	32.1	1318	53.4	48.8	**44.0**	**55.6**	62.7	49.1
PruMerge	12.5%	226.9	28.1	1292	41.2	40.5	33.5	47.9	49.1	40.9
PruMerge+	20.0%	350.1	27.8	1302	43.2	42.3	34.0	47.6	50.7	41.7
FastV	12.5%	527.4	25.3	1276	45.5	43.7	35.8	42.8	50.3	41.3
PyramidDrop	12.5%	953.2	29.6	1260	46.2	45.0	37.9	45.0	52.2	43.0
VEN-VL	7.5%	558.0	**34.3**	**1398**	**55.8**	**50.3**	42.1	54.8	**63.5**	**50.1**

with the increase of FLOPs, **VEN-VL** significantly outperforms the previous methods while retaining only 7.5% of visual tokens.

4.3 Ablation Study

In this section, we conduct analysis experiments on the widely used benchmarks using Qwen3-0.6B as the LLM backbone. The LLaVA-1.5 dataset (Liu et al., 2023) is for model training, and the other training hyperparameters remain consistent by default. More results are deferred to the appendix.

Component Analysis. We investigate the effectiveness of each component in Tab. 3. The integration of MKE brings the improvement of +0.8% on Avg. with the compression ratio of 68%. It validates the rationality of our design where enriched semantic modeling with heterogeneous priors and sequential merging for redundancy alleviation benefits the information capacity and diversity. HTE further progressively refines the visual distribution by pruning the token sequence to merely 7.5% but with superior information density, *e.g.*, +1.1% on AI2D, enabling effective and efficient visual perception. It is worth noting that the ablation of SIP leads to challenge in maintenance the valuable structure information, *e.g.*, −1.0% in TextVQA.

Variant of MKE design. We quantify the impact of different merging strategies within MKE in Tab. 5. It can be seen that although directly concatenating the multi-aspect features without merging yields the highest performance, the quartic complexity brought by 200% is unacceptable. If operating both features with the individual self-merging

Table 3: Effectiveness of the main modules.

MKE	HTE	SIP	Token	MME	MMB	AI2D	TextVQA	Avg.
			100%	1318	53.4	48.8	44.0	48.3
✓			32%	1400	55.4	49.2	42.1	49.1
✓	✓		7.5%	1413	54.9	50.3	41.1	49.2
✓	✓	✓	**7.5%**	1398	55.8	50.3	42.1	**49.5**

Table 4: Impacts of designs of HTE.

HTE	MME	MMB	AI2D	TextVQA	Avg.
First-Half	1276	54.6	48.6	34.6	45.8
Sparse	1383	54.6	49.5	37.6	47.8
Dense	1291	54.3	48.5	36.6	46.4
Second-Half	1398	55.8	50.3	42.1	**49.5**

Table 5: Effect of different strategies of merging.

Method	Token	MME	MMB	AI2D	TextVQA	Avg.
MKE w/o. merge	200%	1440	56.9	51.0	44.4	50.9
MKE w. self-merge	25%	1317	55.8	48.7	36.5	47.0
MKE	32%	1398	55.8	50.3	42.1	49.5

Table 6: Impacts of different architectures of SIP.

Method	MME	MMB	AI2D	TextVQA	Avg.
w/o. SIP	1413	54.9	50.3	41.1	49.2
SIP + $\mathcal{L}_{rec}$	1328	54.5	48.4	40.3	47.7
SIP + $\mathcal{L}_{qrec}$	1398	55.8	50.3	42.1	**49.5**

Table 7: The compatibility of HTE on MoE frameworks.

Backbone	HTE	Token	FLOPs(G)	MME	MMB	AI2D	TextVQA	GQA	Avg.
Kimi-VL-16B-A3B		100%	7900	2195	80.8	82.6	91.9	60.0	78.7
Kimi-VL-16B-A3B	✓	23%	6019	2208	80.7	82.6	91.1	59.8	78.6
VEN-VL-0.6B		32%	716	1424	56.4	50.3	42.5	54.8	50.9
VEN-VL-0.6B	✓	7.5%	558	1413	54.9	50.3	41.1	54.4	50.2

strategy, it leads to drastic performance degradation despite the largest compression ratio. In contrast, our proposed MKE strategy strikes the optimal balance, *e.g.*, 49.5% on Avg. with 32% tokens. This verifies the effectiveness of the structure-aware compression, which preserves the vanilla spatial information and formulates the unified visual space for complete semantic modeling.

Analysis on the HTE. We make a comprehensive analysis of HTE in terms of its variants and compatibility in Tab. 4 and Tab. 7. 1) We notice that applying HTE in the shallow layers (First-Half) results in the lowest performance, particularly in fine-grained tasks like TextVQA. Similarly, applying HTE across all layers (Dense) or at sparse intervals yields suboptimal results, as the fine-grained semantics necessitate later layers for full comprehension. This motivates us to employ HTE on the deep layers (Second-Half), which indicates that with a well-aligned semantic space, the potential of compacting the visual tokens can be unleashed. 2) We further verify the compatibility of our design with the current MoE model. As shown in Tab. 7, directly applying HTE on Kimi-VL (Team et al., 2025) even without training, the performance gap is narrow while the inference speed can be largely boosted. It well depicts the considerable potential for scaling to larger models for more efficiency.

Analysis on the SIP. To verify the rationality of SIP, we showcase the impact of different supervision strategies involved in Tab. 6. As illustrated, the incompatibility incurs when directly regressing the exact value of original features via Mean Squared Error, *i.e.*, SIP + $\mathcal{L}_{rec}$ as the reconstruction objective. It leads to a notable performance degradation, especially on MME *i.e.*, dropped significantly to 1328. We consider that the tight per-value alignment constrains the model towards concentration on the trivial pattern rather than structural information, thus impeding the dynamic token pruning process. In contrast, we incorporate the channel quantization method ($\mathcal{L}_{qrec}$) to map the exact value into a low-level representation, which transforms the supervision process from the per-token alignment into a more topology-based guidance. Therefore, benefiting from the semantically alignment with the vanilla structure, **VEN-VL** yields the most robust performance (Avg. 49.5%) with the precise perception and improved reasoning ability.

5 Conclusion

This paper proposes **VEN-VL**, a visual ensemble MoE framework aiming to bridge the gap between performance and efficiency for multi-modal understanding. Addressing the critical limitations of information capacity and density, **VEN-VL** follows the *enrich then compact* principle to information-enriched visual space via Multi-Aspect Knowledge Ensemble (MKE). Subsequently, through the Hierarchical Token Ensemble (HTE), we leverage adaptive MoE routers to progressively refine the visual distribution, enabling dynamic and precise token selection within the LLM layers for compacting. To guarantee semantic integrity during this reduction process, we empower the framework with Structure Information Preservation (SIP) via explicit visual supervision. Experimental results demonstrate that **VEN-VL** achieves superior per-

formance on complex multimodal tasks with significantly fewer but information-condensed tokens and great compatibility to scaling to larger models.

Acknowledgment

This work is supported by Shenzhen Ubiquitous Data Enabling Key Lab under grant ZDSYS20220527171406015, and by Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Endowed Professorship Scheme of Shenzhen Pengrui Foundation. This work was partly supported by the National Key Research and Development Program of China (No. 2024YFB2808903), the Shenzhen Science and Technology Program (ZDCY20250901103533010).

Limitations

Despite the algorithmic optimization of VEN-VL, our empirical evaluations are currently constrained to small-scale models due to limited computational resources, and scaling this framework to massive state-of-the-art LLMs remains for future work. Additionally, while our approach significantly reduces theoretical FLOPs and visual token counts, translating these gains into actual end-to-end acceleration in real-world deployment further relies on system-level efforts, such as parallel deployment strategies and customized kernel optimizations.

References

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, pages 19–35. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pages 370–387. Springer.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36:49250–49267.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pages 91–104.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *ACL*, pages 320–335.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2024. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yichen Guo, Hanze Li, Zonghao Zhang, Jinhao You, Kai Tang, and Xiande Huang. 2025. Star: Stage-wise attention-guided token reduction for efficient large vision-language models inference. *arXiv preprint arXiv:2505.12359*.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. 2025. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *CVPR*, pages 22487–22497.

Linjiang Huang, Liang Wang, and Hongsheng Li. 2022. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *CVPR*, pages 3272–3281.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*, pages 235–251. Springer.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, and 1 others. 2022. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *ECCV*, pages 620–640. Springer.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, and 1 others. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pages 216–233.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*.

Giannis Nikolentzos and Michalis Vazirgiannis. 2020. Random walk graph neural networks. *NeurIPS*, 33:16211–16222.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Mark Rudelson and Roman Vershynin. 2007. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *CVPR*, pages 22857–22867.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, and 1 others. 2024. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others.

2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, pages 18709–18719.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and 1 others. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025b. Visionzip: Longer is better but not necessary in vision language models. In *CVPR*, pages 19792–19802.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025a. Atp-llava: Adaptive token pruning for large vision language models. In *CVPR*, pages 24972–24982.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. 2025b. Voco-llama: Towards vision compression with large language models. In *CVPR*, pages 29836–29846.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, pages 10809–10818.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. 2020. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *NeurIPS*, 33:9422–9434.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2024. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Experiment Setup

Implementation Details. We follow the similar architecture of LLaVA series (Liu et al., 2024a, 2023). In terms of the visual extractor, CLIP-ViT-L/14-336px(Radford et al., 2021) and RADIO-v2.5(Heinrich et al., 2025) are leveraged to obtain multiple-aspect visual features for semantic knowledge ensemble. To effectively process variable-resolution inputs, we employ the adaptive image tiling strategy. For the language backbone, we conduct experiments on Qwen2/3 family (Team et al., 2024; Yang et al., 2025a) of different parameter sizes with MoE design to verify our scalability.

Training Strategy. We follow a standard two-stage training pipeline. In the initial Vision-Language Alignment stage, we freeze the vision encoders and LLM, training only the projection layers with a learning rate of $2e^{-4}$ and a batch size of 1024. In the subsequent SFT stage, we make the entire model trainable with a learning rate of $4e^{-5}$. Notably, unlike prior MoE methods (Lin et al., 2024; Shu et al., 2024) that require a separate warm-up stage for experts, we integrate the MoE optimization directly into the SFT phase for simplicity. The overall training objective integrates the next-token-prediction loss $\mathcal{L}_{ntp}$ with the SIP guidance $\mathcal{L}_{qrec}$ and the MoE auxiliary load balancing loss. By default, the activated visual experts are configured with an intermediate dimension of half the original FFN size to maintain parameter efficiency.

Datasets & Benchmarks. We train our model based on publicly multi-modal understanding datasets. The alignment stage leverages substantial image-text pairs, *e.g.*, LCS-558K (Liu et al., 2023) for concept learning. The SFT stage employs a diverse mix of 10M instruction-tuning data, spanning across general visual understanding, *e.g.*, ShareGPT4V (Chen et al., 2024b), OKVQA (Marino et al., 2019), OCR and documents, *e.g.*, DocVQA (Mathew et al., 2021), Docmatix (Laurençon et al., 2024), chart&diagram QA, *e.g.*, AI2D (Kembhavi et al., 2016), DVQA (Kafle et al., 2018) and reasoning tasks, *e.g.*, MathVista (Lu et al., 2023), CLEVER-MATH (Johnson et al., 2017), augmented with pure-text instruction data to preserve language capabilities. For evaluation, we conduct extensive evaluation across a wide range of benchmarks, including General Understanding (MMBench (Liu et al., 2024b), MME (Fu et al., 2024), SEED-Bench (Li et al., 2023a)), Fine-grained Perception (TextVQA (Singh et al., 2019), GQA (Hudson and Manning, 2019)), Hallucination (POPE (Li et al., 2023c)), and Knowledge and Reasoning (SciQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), MMMU (Yue et al., 2024)).

A.2 Additional Analyses and Ablations

Analysis on HTE's token selection method. In Tab. 9, we describe the reduction-gated mechanism in HTE by comparing different token scoring functions with Equation (7). Random selection strategy serves as a baseline, resulting in significant performance deterioration, *e.g.*, -107 on MME and -2.0% on MMB, which underscores that visual tokens in deep layers are not uniformly redundant and indicates that the content-aware pruning is essential. Strategies utilizing routing score consistently outperform the random baseline, with Sum achieving a competitive average of 49.3% and Max yielding the optimal performance of 49.5%. It validates our core premise that routing scores effectively serve as indicators of token importance: high-value tokens elicit strong, confident activation of specific experts, whereas redundant tokens exhibit flatter distributions. Furthermore, the superiority of Max suggests that identifying the peak expert demand rather than the accumulated total is the most precise metric for preserving valuable tokens that require specialized processing.

Real-world Inference Latency. We further evaluate the end-to-end inference latency on Qwen3-8B to provide a direct hardware measurement beyond FLOPs. Specifically, we report the time-to-first-token (TTFT) under the same evaluation setting in Figure 3. Compared with the Vanilla baseline, **VEN-VL** reduces the visual tokens to 7.5% and decreases TTFT from **1710ms** to **1146ms**, indicating that the compact visual representation can bring practical acceleration over the standard dense visual sequence. Meanwhile, due to the additional visual branch introduced by MKE, **VEN-VL** still has higher wall-clock latency than particular pruning methods such as PruMerge or FastV. However, we argue that simply sacrificing performance for acceleration is unsatisfactory. Although our design introduces additional computational overhead compared with these methods, it yields clear performance improvements. Furthermore, owing to

Table 8: FLOPs-matched comparison with efficient methods.

Method	Tokens	FLOPs(G)	MMMU	MME	MMB	AI2D	TextVQA	GQA	SEED	Avg.
Vanilla	100%	2330.9	32.1	1318	53.4	48.8	44.0	55.6	62.7	49.1
PruMerge+	32%	569.2	31.6	1277	50.8	46.2	36.1	52.9	58.2	45.9
FastV	15%	561.6	26.7	1189	45.8	44.2	36.4	42.2	51.4	41.3
PyramidDrop	2.7%	560.5	28.1	1264	43.4	42.6	32.3	45.0	50.1	40.9
VEN-VL	7.5%	558.0	34.3	1398	55.8	50.3	42.1	54.8	63.5	50.1

Table 9: Comparison of token selection methods.

Token Score	MME	MMB	AI2D	TextVQA	Avg.
Random	1291	53.8	47.8	40.6	47.1
Sum	1395	55.4	49.9	42.2	49.3
Max	1398	55.8	50.3	42.1	**49.5**

its strong scalability to larger MoE models, as validated in Tab. 7, our method offers a more promising solution for real-world deployment for effective and efficient multi-modal understanding.

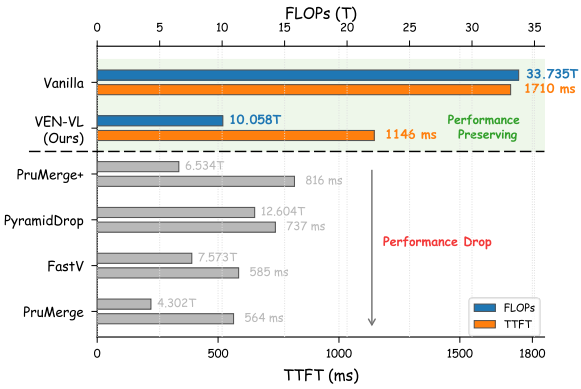


Figure 3: Real-world inference cost on Qwen3-8B. TTFT denotes time-to-first-token.

Comprehensive Comparison with Efficient Model. To further demonstrate the rationality of our design, we additionally incorporate a FLOPs-matched experiment with the previous efficient methods. As presented in 8, even with the increased number of visual tokens, the performance of these methods still falls considerably behind that of VEN-VL.

Visualization of Multi-aspect Visual Features. We visualize the feature responses from the two visual branches to further understand the complementarity of MKE. As shown in Figure 4, the CLIP branch tends to focus on salient objects and semantic regions, while the RADIO branch preserves more fine-grained textures and local patterns. This observation supports our motivation that multi-aspect visual features provide complementary visual clues, rather than repeated encoding from two homogeneous views. By unifying heterogeneous representations with MKE, **VEN-VL** obtains a visual representation with improved information capacity before the token refinement.

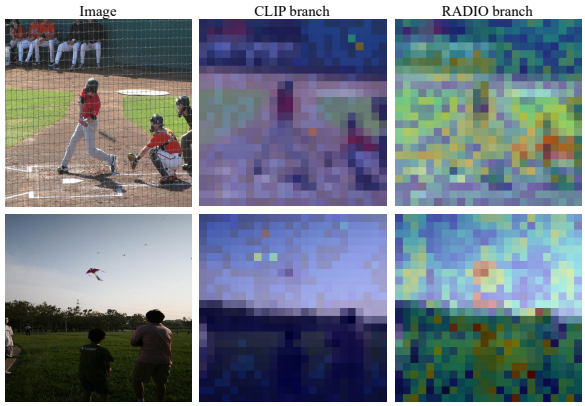


Figure 4: Visualization of multi-aspect visual features from CLIP and RADIO branches.

Individual effect of HTE and MKE. We conduct additional ablation to verify each of the effects of HTE and MKE in Tab. 10. With the single-encoder visual input, HTE only improves the average score from 48.3 to 50.9, which indicates that the MoE router itself can serve as an effective indicator to identify information-condensed tokens and improve information density. However, the FLOPs remain high because the input visual sequence still contains substantial redundancy. After incorporating MKE, the multi-aspect visual features are compressed into a more enriched and compact representation, reducing FLOPs from 1872G to 558G while maintaining comparable performance. Finally, the full **VEN-VL** further introduces SIP to preserve structure information, bringing the best average performance under the same token ratio and computational cost.

Table 10: Individual effect of HTE and MKE.

Method	Tokens	FLOPs(G)	MME	MMB	AI2D	TextVQA	Avg.
Vanilla	100%	2330	1318	53.4	48.8	44.0	48.3
HTE only	22.8%	1872	1409	56.7	50.9	45.8	50.9
MKE+HTE	7.5%	558	1413	54.9	50.3	41.1	49.2
VEN-VL	7.5%	558	1398	55.8	50.3	42.1	49.5

Parameter Analysis of DropRate. We analyze the sensitivity of DropRate in HTE across three representative benchmarks, including MMBench for natural images, OCRBench for OCR documents, and AI2D for chart and diagram understanding. As shown in Tab. 11, DropRate controls the effective compression strength of HTE. When DropRate is no larger than 0.1, the performance remains relatively stable while the number of visual tokens and FLOPs are significantly reduced. However, further increasing DropRate to 0.15 or 0.2 leads to performance degradation, especially on OCRBench and AI2D, where fine-grained visual details are critical. Therefore, we adopt DropRate $= 0.1$ as the default setting to balance performance and efficiency.

Table 11: Sensitivity analysis of DropRate in HTE.

DropRate	Final Tokens	FLOPs(G)	MMBench	OCRBench	AI2D
–	100%	2330	53.4	34.0	48.8
0.0	32%	716	56.4	33.8	50.3
0.05	16%	618	56.8	32.7	50.9
0.1	7.5%	558	55.8	33.0	50.3
0.15	3.3%	519	55.0	31.2	49.5
0.2	1.4%	492	53.7	30.4	47.9

Parameter Analysis of MKE Merge Ratios. We further study the merge ratios in MKE, where k_s denotes the self-merge ratio and k_c denotes the cross-merge ratio. As shown in Tab. 12, directly concatenating two visual branches improves the performance by enlarging information capacity, but it also doubles the input tokens and brings heavy computational overhead. In contrast, moderate merging ratios can effectively suppress redundancy while preserving multi-aspect information. When the merging becomes too aggressive, the performance on MMBench and OCRBench decreases, indicating that excessive merging may damage semantic regions and fine-grained text patterns. We set $k_s = 0.5$ and $k_c = 0.4$ by default, which provides a favorable balance between compactness and visual understanding.

Table 12: Sensitivity analysis of MKE merge ratios.

Setting	k_s	k_c	Input Tokens	FLOPs(G)	MMBench	OCRBench	AI2D
Vanilla	–	–	100%	2330	53.4	34.0	48.8
w/o. merge	–	–	200%	5297	56.9	34.2	51.0
MKE	0.0	0.0	50%	884	56.3	33.3	49.6
MKE	0.3	0.2	39%	674	56.0	32.7	50.1
MKE	0.3	0.4	35%	611	55.9	33.0	49.9
MKE	0.5	0.4	32%	558	55.8	33.0	50.3
MKE	0.8	0.4	28%	480	53.7	31.5	49.2
MKE	0.8	0.6	27%	463	53.3	30.2	48.7

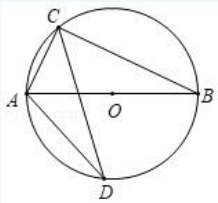
Parameter Analysis of Semantic Propagation Factor. We also evaluate the propagation factor ω in SIP. Different from DropRate and merge ratios, ω does not change the number of visual tokens or FLOPs, but controls the strength of structure information propagation. As shown in Tab. 13, $\omega = 0.8$ achieves the best 6-Avg. score, while smaller or larger values slightly degrade the overall performance. Therefore, we use $\omega = 0.8$ as the default setting in our experiments.

Table 13: Sensitivity analysis of the propagation factor.

ω	MMBench	OCRBench	AI2D	6-Avg.
0.7	56.0	33.0	48.8	47.6
0.8	55.8	33.0	50.3	47.9
0.9	55.6	32.6	50.7	47.4

A.3 Additional Qualitative Results

We include representative qualitative cases to showcase the visual reasoning and dense captioning capabilities of our method in Figure 5. By enriching the visual semantics and progressively refining the visual distribution, VEN-VL demonstrates superiority in fine-grained visual understanding tasks with few information-condensed tokens.



As shown in the figure, AB is the diameter of ⊙O, CD is the chord of ⊙O, ∠ADC = 26.0, then the degree of ∠CAB is __? Provide the reasoning process.

Qwen2-VL

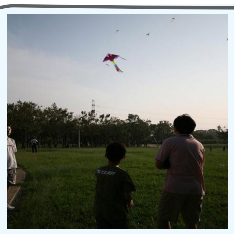
Given that AB is the diameter of circle O and ∠ADC = 26°, we can use the properties of circles and angles to find the measure of ∠CAB.
1. Since AB is the diameter of circle O, ∠ACB is a right angle (90°) because the diameter subtends a right angle to any point on the circle.
2. In any triangle, the sum of the angles is 180°. Therefore, ∠CAB + ∠CDB + ∠CAB = 180°.
3. Since CD is a chord of circle O, ∠CDB is also 90° because the angle subtended by a chord at the center of a circle is twice the angle subtended by the same chord at any point on the circumference.
4. Therefore, ∠CAB + ∠CAB = 180° - 90° = 90°.
Given that ∠ADC = 26°, we can find ∠CAB by subtracting this value from 90°:
∠CAB = 90° - 26° = 64°.
So, the degree of ∠CAB is 64°.

VEN-VL

Given that AB is the diameter of circle O, we know that any angle inscribed in a semicircle is a right angle. Therefore, ∠ACB = 90°.
Since ∠ADC = 26° and ∠ABC is an inscribed angle subtended by the same arc as ∠ADC, we can conclude that ∠ABC = ∠ADC = 26°.

However, since ∠ACB is a right angle (90°), we can find the measure of ∠CAB by subtracting the sum of the other two angles from 180°:
∠CAB = 180° - ∠ACB - ∠ABC
∠CAB = 180° - 90° - 26°
∠CAB = 64°

Therefore, the degree of ∠CAB is 64°



Describe this image in detail.

Qwen2-VL

The image depicts a serene outdoor scene set during what appears to be late afternoon or early evening, as indicated by the soft, warm lighting. The sky is mostly clear with a few scattered clouds, suggesting fair weather. The horizon is visible, with a line of trees and a power line pole on the left side of the image. The ground is covered with a lush green grassy field, which extends towards the horizon.
In the foreground, there are three individuals standing on the grass. The person on the left is partially visible, wearing a white shirt and dark pants. The person in the middle is wearing a green t-shirt and dark shorts. The person on the right is wearing a pink t-shirt and dark pants. All three individuals are facing towards the sky, watching a colorful kite flying high in the air.
In the background, there are more people standing and walking, suggesting that this is a popular spot for outdoor activities. The overall atmosphere is calm and relaxed, with people enjoying the outdoors and the pleasant weather.

VEN-VL

In the image, a tranquil scene unfolds in a park under a clear blue sky. Two individuals are engaged in the delightful activity of kite flying.

The person on the left, clad in a black shirt and shorts, is holding onto a string that leads up to a vibrant kite soaring high in the sky. This kite is a spectacle of colors with hues of purple, pink, and yellow. On the right, another person stands wearing a pink shirt paired with khaki shorts.

They too hold a string attached to the kite.
The park itself is lush with trees providing a serene backdrop to this joyful activity. In the distance, other people can be seen enjoying their day out in the park as well. The entire scene exudes an atmosphere of leisure and enjoyment.

Figure 5: **Output cases on math reasoning and dense caption.**