Cat-MoD: Accelerating Multimodal Alignment via Caption Token Guided Asymmetric Mixture-of-Depths

Yijie Huang[1], **Xiaocui Yang**[1*], **Shi Feng**[1*], **Wen Zhang**[1], **Kaisong Song**[2], **Yifei Zhang**[1], **Daling Wang**[1]
[1]School of Computer Science and Engineering, Northeastern University Shenyang 110819, China
[2]Alibaba Group, Hangzhou, China
{2401837, 2401967}@stu.neu.edu.cn, kaisong.sks@alibaba-inc.com
{fengshi, yangxiaocui, wangdaling, zhangyifei}@cse.neu.edu.cn

Abstract

Efficiently aligning visual features with Large Language Models (LLMs) remains a critical bottleneck in Multimodal LLMs. Existing query-based alignment modules (e.g., Q-Former) rely on randomly initialized queries, resulting in an inefficient cold start exploration process. Furthermore, they enforce uniform cross-attention across all layers, leading to computational redundancy. Our empirical analysis reveals that query tokens initialized with language priors can rapidly capture global semantics, leading to early representation convergence after only a few layers. In this paper, we propose **Cat-MoD**, a **Ca**ption **t**oken Guided Asymmetric **M**ixture-**o**f-**D**epths framework. It incorporates a **Hybrid Query Construction** module where Guide Tokens initialized from coarse-grained linguistic priors rapidly anchor global semantic context, and randomly initialized Explorer Tokens remain active to capture fine-grained visual details. Leveraging this early convergence, we introduce an **Asymmetric Mixture-of-Depths** mechanism, where a similarity-aware router dynamically prunes redundant tokens from expensive cross-attention layers while preserving their context in self-attention. Experiments on multiple benchmarks demonstrate that Cat-MoD matches or surpasses baseline performance, while substantially reducing alignment FLOPs by approximately 37% during both training and inference, offering a highly efficient solution for multimodal alignment. Code: https://github.com/JasonOrange0726/Cat-MoD.

1 Introduction

Multimodal Large Language Models (MLLMs) have emerged as a dominant paradigm for vision-language understanding (Wang et al., 2025; Wu et al., 2024; Chen et al., 2024b; Liu et al., 2023).

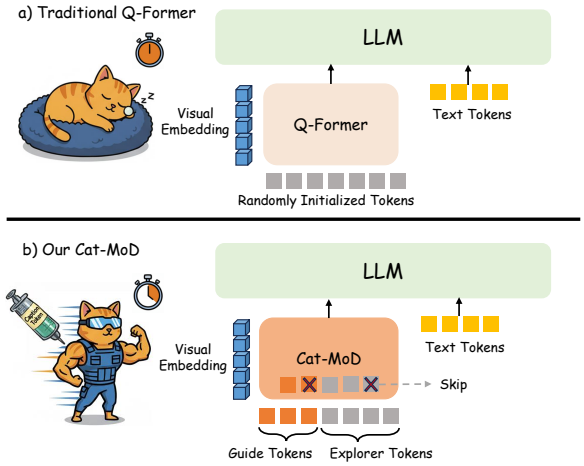


Figure 1: (a) Traditional Q-Former: Random initialization leads to inefficient cold-start exploration and redundant dense computation. (b) Cat-MoD (Ours): By injecting linguistic priors, Guide Tokens (Orange) rapidly anchor semantics, enabling the pruning of redundant interactions (Red Crosses) while Explorer Tokens (Grey) focus on fine-grained reasoning.

A central challenge in these systems lies in multimodal alignment, i.e., efficiently compressing high-dimensional visual representations for LLMs without sacrificing fine-grained semantic information (Yoon et al., 2025; Vasu et al., 2025; Cha et al., 2024). Among existing solutions, the Querying Transformer (Q-Former) adopted in BLIP-2 (Li et al., 2023) has proven particularly effective, leveraging a small number of learnable queries to attend over dense visual features and thereby dramatically reducing inference cost.

However, while the Q-Former is efficient in token compression, its internal computation exhibits inherent redundancies that have received little systematic analysis. First, queries of Q-Former are typically initialized as agnostic learnable parameters or inherited from pre-training, resulting in an inefficient exploration process during adaptation. These queries must gradually discover meaning-

* Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26354–26376
July 2-7, 2026

ful visual correspondences, leading to slow convergence and redundant computation. Second, Q-Former enforces uniform and deep cross-attention computation across all query tokens and all odd layers. In practice, not all queries continue to acquire novel visual information at deeper layers, raising concerns about substantial computational redundancy (Chen et al., 2024a).

To identify the root of this redundancy, we investigate layer-wise token dynamics by measuring the cosine similarity of consecutive representations (Figure 2). In the standard Q-Former (gray dashed line), randomly initialized tokens exhibit slow, homogeneous convergence, necessitating full computation at every depth. To probe whether semantic context can accelerate alignment, we conduct a pilot study by initializing a subset of queries using linguistic priors derived from image captions (termed *Guide Tokens*), while leaving the rest as random parameters (termed *Explorer Tokens*). This setup reveals a distinct behavioral divergence: Guide Tokens (orange line) achieve rapid saturation (similarity $\rightarrow$ 1.0) within just 3 layers, implying that their deep-layer Cross-Attention is redundant. Conversely, Explorer Tokens (blue line) maintain high variance and remain dynamically active.

Motivated by this, applying Mixture-of-Depths (MoD) (Raposo et al., 2024) to skip transformer blocks appears to be a natural solution. However, directly applying such symmetric block-skipping to the Q-Former is detrimental. Since the Q-Former relies on Self-Attention to broadcast visual information among queries, skipping entire transformer blocks severs the semantic connectivity, causing performance collapse (Kim et al., 2024; Xu et al., 2024). Based on these insights, we propose **Cat-MoD** (Figure 1), a framework that decouples computational pruning from context preservation. We introduce **Hybrid Query Construction**, initializing Guide Tokens with linguistic priors to rapidly anchor global semantics, while retaining Explorer Tokens for fine-grained details. We devise an **Asymmetric Mixture-of-Depths** mechanism, which employs a similarity-aware router to prune saturated tokens from expensive Cross-Attention while keeping them active in Self-Attention. This guarantees context integrity, maintaining or improving performance while reducing alignment FLOPs by 36–39% during both training and inference. Our main contributions are summarized as follows:

- We propose Cat-MoD, a framework that incorporates **Hybrid Query Construction** to enforce functional specialization. By initializing Guide Tokens with linguistic priors, we facilitate rapid semantic anchoring and enable significant computational savings.
- We propose a novel **Asymmetric Mixture-of-Depths** mechanism that decouples Cross-Attention pruning from Self-Attention to prevent context collapse. By supervising a similarity-aware router, our approach enables precise, saturation-driven token pruning.
- Experiments on 7 benchmarks demonstrate that Cat-MoD achieves a superior trade-off between efficiency and performance. It consistently maintains or improves accuracy over baselines while reducing FLOPs by 36–39%, verifying its effectiveness as a generalizable plug-and-play module.

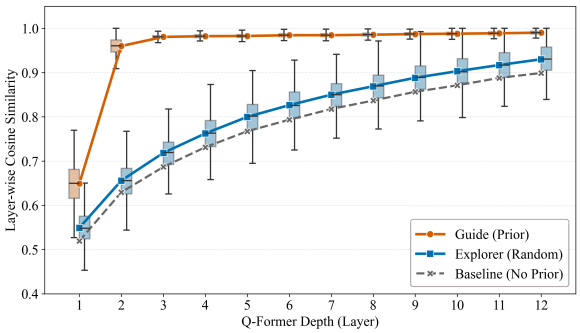


Figure 2: Impact of Language Priors on Token Dynamics. Layer-wise cosine similarity of query tokens using InstructBLIP (Dai et al., 2023) on the COCO dataset (Lin et al., 2014). The baseline (grey) converges slowly and uniformly, while Cat-MoD shows functional specialization: Guide Tokens (Orange) saturate within 3 layers, whereas Explorer Tokens (Blue) remain active.

2 Related Work

2.1 Semantic Anchoring and Hierarchical Visual Alignment

Efficiently aligning visual features with LLMs remains challenging due to the cold start inefficiency inherent in randomly initialized queries (Dai et al., 2023). Early works like Li et al. (2020) utilized semantic anchors to facilitate alignment. Extending this understanding, recent studies reveal a hierarchical saturation pattern in MLLMs: global semantics are captured early, whereas fine-grained details require deep interaction (Yoon et al., 2025; Zhang et al., 2025; Kaduri et al., 2024). In this

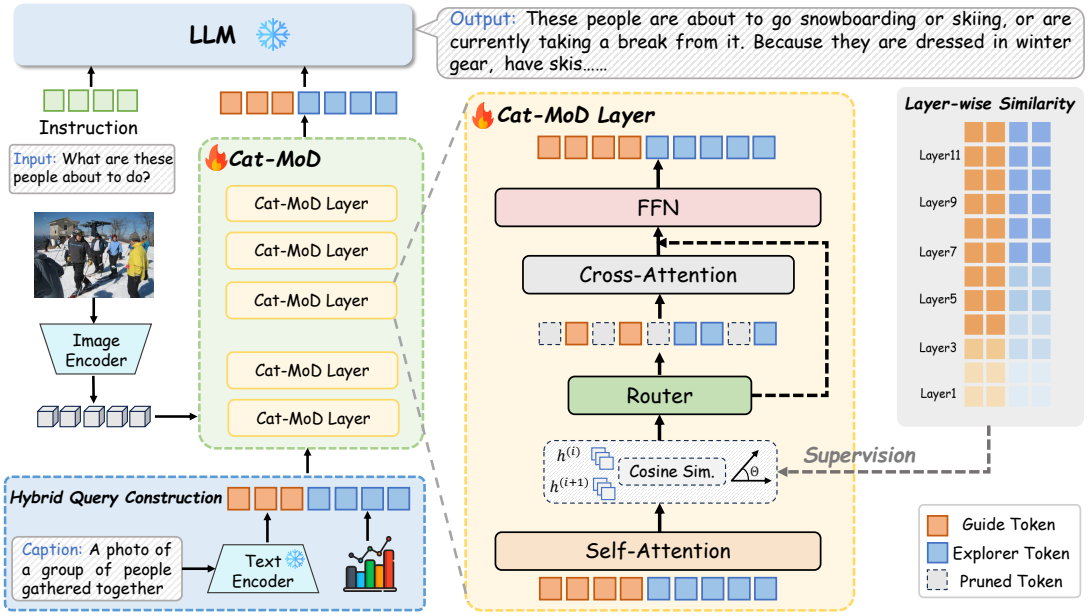


Figure 3: The Cat-MoD Architecture. The framework utilizes Hybrid Query Construction to anchor global semantics via caption-initialized Guide Tokens (Orange) alongside learnable Explorer Tokens (Blue). Within each layer, an Asymmetric Mixture-of-Depths mechanism employs a Similarity-Aware Router to assess token saturation. Saturated tokens (dashed boxes) bypass the expensive Cross-Attention module via a residual link, while active tokens continue visual interaction, significantly reducing computational costs.

work, we propose Cat-MoD, which leverages this hierarchical saturation phenomenon by employing a hybrid initialization: caption-derived Guide Tokens rapidly anchor the global context, allowing Explorer Tokens to specialize in deep-layer visual reasoning for fine-grained feature extraction, thereby achieving both fast convergence and high precision.

2.2 Token Pruning and Sparse Computation

Sparse computation methods like Mixture-of-Depths (MoD) (Luo et al., 2024; Raposo et al., 2024) and Token Pruning have become standard for efficiency. Recent works such as FastV (Chen et al., 2024a) and PyramidDrop (Xing et al., 2024) accelerate MLLMs by permanently discarding redundant visual tokens in deeper layers. However, applying such hard pruning to the Q-Former is detrimental. Since queries rely on Self-Attention to broadcast information, dropping tokens ruptures semantic context, causing performance collapse (Kim et al., 2024). To address this, Cat-MoD introduces an Asymmetric MoD strategy, where we selectively prune only the expensive Cross-Attention path for saturated tokens while maintaining their participation in the lightweight Self-Attention. This unique decoupling reduces computational cost without sacrificing semantic connectivity.

3 Methodology

We propose Cat-MoD to accelerate MLLMs by leveraging the distinct saturation patterns of semantic and random queries. As shown in Figure 3, Cat-MoD comprises two key components: (1) Hybrid Query Construction for functional specialization; and (2) Asymmetric Mixture-of-Depths mechanism that uses a similarity-aware router to dynamically decouple internal context modeling from external visual perception.

3.1 Hybrid Query Construction

Standard Q-Formers initialize queries as a homogeneous set of learnable parameters. Lacking initial semantic context, these queries are forced to localize visual objects from scratch through inefficient exploration. To overcome this limitation, we propose a Hybrid Query Construction strategy that enforces a distinct functional specialization among queries. We explicitly formulate the initial query, which serves as the input to the first transformer layer, as $\mathbf{X}^{(0)} \in \mathbb{R}^{N \times D}$, obtained by concatenating two functional groups:

$$\mathbf{X}^{(0)} = [\mathbf{Q}_{guide}\,;\,\mathbf{Q}_{explorer}], \tag{1}$$

where N is the total number of queries and D is the embedding dimension. By injecting language priors, we establish a clear dichotomy in their roles.

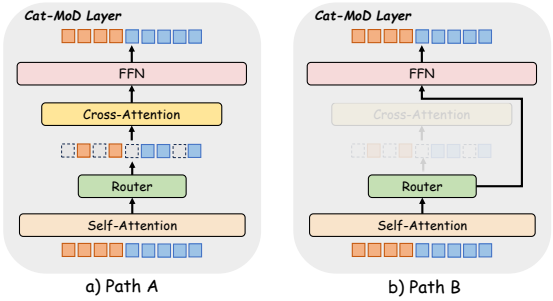


Figure 4: **Overview of Asymmetric Routing Pathways.** The router dynamically selects between Path A for active tokens and Path B for saturated tokens, skipping the expensive Cross-Attention block.

Guide Tokens ($\mathbf{Q}_{guide}$): The Semantic Anchors. These tokens are initialized by projecting coarse-grained image captions via a frozen text encoder. Specifically, we employ an off-the-shelf captioning model to generate the descriptive caption for each image, which serves as the linguistic prior for initialization (refer to Appendix A for details). We further discuss the end-to-end deployment trade-off of online caption generation, including a parallel execution setup and break-even analysis, in Appendix G. This initialization establishes Guide Tokens as semantic anchors, enabling rapid alignment with global visual features.

Explorer Tokens ($\mathbf{Q}_{explorer}$): The Contextual Reasoners. Remaining as randomly initialized learnable parameters, their role is to complement Guide Tokens by actively seeking background information, fine-grained textures, or implicit relationships absent from the caption. Unlike their counterparts, $\mathbf{Q}_{explorer}$ maintains high activity levels throughout intermediate layers and undergo natural saturation only in deeper layers, ensuring comprehensive visual understanding.

3.2 Asymmetric Mixture-of-Depths

The conventional Mixture-of-Depths (MoD) (Raposo et al., 2024) optimizes efficiency by conditionally skipping the entire computation block $\mathcal{D}(\cdot)$ for redundant tokens. It can be written as:

$$\mathbf{x}_i^{(l)} = \begin{cases} \mathbf{x}_i^{(l-1)} + \mathcal{D}(\mathbf{x}_i^{(l-1)}) & \text{if } \mathcal{R}(\mathbf{x}_i^{(l-1)}) > \tau, \\ \mathbf{x}_i^{(l-1)} & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{R}(\cdot)$ is the routing function, $\mathbf{x}_i^{(l-1)}$ denote the representation of the i-th token, τ is the activation threshold. However, such symmetric skipping is detrimental to the Q-Former, as it severs the Self-Attention links required for broadcasting context among queries. To address this, we propose **Cat-MoD**, which decouples routing into a fine-grained, component-level adaptation. As illustrated in Figure 4, a Similarity-Aware Router governs two distinct pathways: **Path A**: The Router activates the full computational path, executing both Self-Attention (SA) and Cross-Attention (CA); **Path B**: The Router bypasses the expensive CA, while keeping SA active to prevent context collapse common in hard routing. Formally, let $\mathbf{X}^{(l-1)} \in \mathbb{R}^{N \times D}$ denote the input query representations for the l-th layer. The computation proceeds in three stages:

Stage 1: Context Preservation (Path A). First, all tokens participate in Self-Attention to aggregate global context and maintain alignment with instruction tokens. This generates the context-aware intermediate representation $\mathbf{H}^{(l)}$:

$$\mathbf{H}^{(l)} = \text{SelfAttn}(\text{LN}(\mathbf{X}^{(l-1)})) + \mathbf{X}^{(l-1)}, \tag{3}$$

where $\text{LN}(\cdot)$ denotes Layer Normalization. This step is critical for preventing *context collapse*, ensuring that even tokens pruned from visual interaction remain accessible as semantic anchors.

Stage 2: Similarity-Aware Routing Decisions. Based on the row vectors $\mathbf{h}_i^{(l)}$ from $\mathbf{H}^{(l)}$, we determine the *demand for visual update* for each token. We introduce a lightweight router $\mathcal{R}^{(l)}$ (parameterized by $\mathbf{W}_1, \mathbf{W}_2$) that predicts a scalar unsaturation score $s_i^{(l)} \in [0, 1]$:

$$s_i^{(l)} = \sigma(\mathbf{W}_2 \cdot \text{GeLU}(\mathbf{W}_1 \cdot \text{LN}(\mathbf{h}_i^{(l)}))). \tag{4}$$

A token is added to the active set $\mathcal{I}_{active}$ only if its score exceeds a learnable threshold τ: $\mathcal{I}_{active} = \{i \mid s_i^{(l)} > \tau\}$, corresponding to tokens that are still unsaturated and require Cross-Attention.

Stage 3: Dynamic Visual Interaction (Path B). Finally, tokens in $\mathcal{I}_{active}$ perform standard Cross-Attention to interact with visual features $\mathbf{V}$, while saturated tokens bypass the module via an identity connection. Let $\tilde{\mathbf{x}}_i^{(l)}$ denote the post-attention state:

$$\tilde{\mathbf{x}}_i^{(l)} = \begin{cases} \text{CrossAttn}(\mathbf{h}_i^{(l)}, \mathbf{V}) + \mathbf{h}_i^{(l)}, & \text{if } i \in \mathcal{I}_{active} \\ \mathbf{h}_i^{(l)}, & \text{otherwise.} \end{cases} \tag{5}$$

The resulting $\tilde{\mathbf{X}}^{(l)}$ is then processed by the FFN to yield the final layer output $\mathbf{X}^{(l)}$. This ensures that even tokens bypassing visual interaction undergo necessary semantic transformations.

Algorithm 1: Forward Pass of Cat-MoD

Input: Visual features $\mathbf{V}$, Caption C, Threshold τ, Total layers L
Output: Query representations $\mathbf{H}^{(L)}$

$\mathbf{Q}_{guide} \leftarrow \text{TextEncoder}(C)$
$\mathbf{Q}_{explorer} \leftarrow \text{Param}(\mathcal{N}(0, \mathbf{I}))$
$\mathbf{H}^{(0)} \leftarrow \text{Concat}([\mathbf{Q}_{guide}; \mathbf{Q}_{explorer}])$
for $l \leftarrow 1$ **to** L **do**
  $\mathbf{H}' \leftarrow \text{SelfAttn}(\text{LN}(\mathbf{H}^{(l-1)})) + \mathbf{H}^{(l-1)}$
  $s^{(l)} \leftarrow \sigma(\mathcal{R}^{(l)}(\mathbf{H}'))$
  $\mathcal{I}_{active} \leftarrow \{i \mid s_i^{(l)} > \tau\}$
  if *Layer l has Cross-Attention* **then**
    $\mathbf{H}''_{\mathcal{I}_{active}} \leftarrow \text{CrossAttn}(\mathbf{H}'_{\mathcal{I}_{active}}, \mathbf{V})$
    $\mathbf{H}''_{i \notin \mathcal{I}_{active}} \leftarrow \mathbf{H}'_{i \notin \mathcal{I}_{active}}$
  else
    $\mathbf{H}'' \leftarrow \mathbf{H}'$
  end
  $\mathbf{H}^{(l)} \leftarrow \text{FFN}(\text{LN}(\mathbf{H}'')) + \mathbf{H}''$
end
return $\mathbf{H}^{(L)}$

Explicit Supervision via Feature Evolution. To enable interpretable decisions, we supervise the router using the feature evolution of tokens. We define the ground-truth unsaturation signal $y_{i,GT}^{(l)}$ based on the cosine similarity between the input $\mathbf{h}_i^{(l)}$ and the potential dense output:

$$y_{i,GT}^{(l)} = 1 - \text{CosineSim}(\mathbf{h}_i^{(l)}, \text{CrossAttn}(\mathbf{h}_i^{(l)}, \mathbf{V})). \quad (6)$$

Here, a high similarity implies saturation ($y_{GT} \approx 0$). The router is trained jointly with the LLM using a Mean Squared Error (MSE) loss:

$$\mathcal{L}_{router} = \frac{1}{N} \sum_{i=1}^{N} \left\| s_i^{(l)} - \text{sg}(y_{i,GT}^{(l)}) \right\|^2. \quad (7)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator. The overall training objective is defined as a weighted sum of the generative language modeling loss and the router supervision loss:

$$\mathcal{L}_{total} = \mathcal{L}_{LM} + \lambda \mathcal{L}_{router}, \quad (8)$$

where $\mathcal{L}_{LM}$ denotes the standard causal language modeling loss, and λ is a hyperparameter balancing the trade-off between semantic generation quality and routing accuracy.

The overall forward pass of Cat-MoD is detailed in Algorithm 1. For clarity, we explicitly define the intermediate variables used in the algorithm. Specifically, H' denotes the intermediate representation after the Self-Attention block, where tokens have aggregated global semantic context. H'' denotes the representation after the conditional Cross-Attention block, which integrates visual information for active tokens ($\mathcal{I}_{\text{active}}$) while preserving the Self-Attention representations for pruned tokens.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. Following the established protocols of baseline models (Dai et al., 2023), we conduct experiments using the standard held-in multimodal instruction tuning datasets for training, and evaluate zero-shot performance on 7 held-out benchmarks. Detailed dataset statistics and splits are provided in Appendix B. The evaluation benchmarks are categorized into two groups based on task characteristics: General & Reasoning Tasks: GQA (Hudson and Manning, 2019), ScienceQA (Singh et al., 2019), IconQA (Lu et al., 2021) and VizWiz (Bigham et al., 2010) (requiring global semantic grounding). Text-Rich & Perception Tasks: TextVQA (Lu et al., 2022), OCR-VQA (Mishra et al., 2019), and ChartQA (Masry et al., 2022) (requiring fine-grained detail extraction).

Baselines. We benchmark Cat-MoD against representative MLLMs that natively integrate the Q-Former architecture, including InstructBLIP (Dai et al., 2023), MiniGPT-4 (Zhu et al., 2023), and BLIVA (Hu et al., 2024). Additionally, to validate the universality and scalability of our method, we extend our evaluation to the modern Qwen-2.5 family (Yang et al., 2024) across different model sizes by equipping them with the Cat-MoD alignment module. All baselines are evaluated under a unified training setting to ensure fair comparison (see Appendix B for detailed configurations).

Implementation Details. We construct our Cat-MoD using ViT-G/14 from EVA-CLIP (Sun et al., 2023) as the visual encoder, the alignment module is initialized from Q-Former ($D = 768, N = 32$) and reconfigured with our proposed Cat-MoD. We freeze the vision encoder and LLM, fine-tuning only the alignment module under this setting. We use the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.05. All experiments are conducted on 4 NVIDIA A800 (80GB) GPUs.

Query Allocation		TextVQA (Text-Rich)		IconQA (Visual Reasoning)		ScienceQA (Knowledge)		GQA (General)		FLOPs Savings
N_{Guide}	$N_{Explorer}$	Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ	
0	32	50.1	-	43.1	-	60.5	-	49.2	-	0.0%
2	30	50.5	+0.4	43.6	+0.5	60.9	+0.4	49.5	+0.3	18.5%
4	28	50.7	+0.6	43.9	+0.8	61.1	+0.6	50.1	+0.9	26.1%
8	**24**	**51.3**	**+1.2**	**44.7**	**+1.6**	**61.4**	**+0.9**	**52.6**	**+3.4**	37.6%
16	16	49.5	-0.6	42.5	-0.6	60.2	-0.3	48.1	-1.1	49.8%
32	0	44.8	-5.3	38.2	-4.9	56.2	-4.3	44.6	-4.6	54.6%

Table 1: **Ablation on Guide Token Allocation.** We investigate the impact of caption-initialized tokens (N_{Guide}). **Finding:** Performance peaks at $N_{Guide} = 8$, striking an optimal balance where sufficient semantic anchors stabilize alignment without crowding out Explorers, achieving superior accuracy across all benchmarks.

4.2 Determination of Optimal Query Configuration

We analyze the trade-off between semantic grounding and visual reasoning by varying the allocation of Guide Tokens (N_{Guide}) on the InstructBLIP (Vicuna-7b) architecture. To evaluate how this balance impacts different multimodal capabilities, we selected four representative benchmarks: TextVQA, IconQA, ScienceQA and GQA. This selection allows us to stress-test the distinct roles of Explorer Tokens in capturing fine-grained details and Guide Tokens in preserving global semantics. See Appendix D for more details.

Table 1 reveals a distinct inverted U-shaped trend. **Benefit of Semantic Anchoring ($N \leq 8$):** Introducing a moderate number of Guide Tokens yields consistent gains across all tasks. This suggests that injecting caption priors successfully mitigates the cold-start problem, providing a stable semantic anchor that accelerates alignment. **The Explorer Scarcity Phenomenon ($N > 8$):** Excessive allocation to Guide Tokens leads to a sharp performance decline. Crucially, this degradation is non-uniform: fine-grained tasks like TextVQA suffer the most severe drop (-5.3% at $N = 32$), whereas knowledge reasoning tasks like ScienceQA are less affected (-4.3%). This confirms our hypothesis: since Guide Tokens saturate and are pruned early, an over-allocation to this group leaves insufficient active queries (Explorer Tokens) in deeper layers to attend to fine-grained visual details (e.g., small text) that are absent in the global caption.

Consequently, the configuration of $N_{Guide} = 8$ emerges as the optimal balance point, maximizing accuracy while reducing alignment FLOPs by 37.6%. We therefore adopt this setting ($N_{Guide} = 8, N_{Explorer} = 24$) for all subsequent experiments. Notably, this optimum is not sharply tuned. A finer-grained sweep in Appendix D shows that configurations with $N_{\text{Guide}} \in [4, 12]$ consistently remain above the dense baseline, with performance degrading only gradually once the number of Guide Tokens becomes too large. This broad plateau indicates that Cat-MoD is not overly sensitive to precise query allocation, while the sharper drop beyond $N_{\text{Guide}} > 12$ further supports our Explorer Scarcity hypothesis.

4.3 Main Results

Efficiency and Performance Gains. As detailed in Table 2, Cat-MoD achieves a superior trade-off compared to the dense baseline. On the primary InstructBLIP (Vicuna-7B) benchmark, our method improves average accuracy by +1.5% while reducing alignment FLOPs by 37.2%. Notably, the method yields consistent gains across all categories, validated by our statistical significance analysis (Appendix F). The improvement is particularly substantial on GQA (+3.4%, $p < 0.001$), demonstrating the benefit of semantic anchoring for general reasoning. Furthermore, fine-grained tasks see statistically significant improvements (e.g., +1.2% on TextVQA, +1.6% on IconQA). This confirms that pruning saturated Guide Tokens not only enhances efficiency but also reduces noise, allowing Explorer Tokens to focus on critical visual details.

Universality as a Plug-and-Play Module. Our method demonstrates robust generalization across diverse architectures. As shown in Table 2, Cat-MoD seamlessly adapts to established Q-Former-based MLLMs (InstructBLIP, MiniGPT-4, and BLIVA). Crucially, we extend this success to the Qwen-2.5 family. Although these models do not natively integrate a Q-Former, equipping them with Cat-MoD endows them with competitive visual capabilities (e.g., +4.2% on GQA with Qwen2.5-7B). In all tested scenarios, our method instantly reduces inference cost by ∼36–39% while main-

Model	Backbone	Method	Reasoning & General				Text-Rich & Perception			Efficiency(Avg)	
			GQA	IconQA	VizWiz	SciQA	TextVQA	ChartQA	OCRVQA	Acc	FLOPs Sav
InstructBLIP	Vicuna-7B	Q-Former	49.2	43.1	34.5	60.5	50.1	5.5	47.6	41.5	0.0%
		Cat-MoD	**52.6**	**44.7**	**35.7**	**61.4**	**51.3**	**6.1**	**49.2**	**43.0**	**37.2%**
	FlanT5-XXL	Q-Former	47.9	51.2	30.9	70.6	46.6	8.1	55.0	44.3	0.0%
		Cat-MoD	**50.4**	**51.8**	**31.4**	**71.3**	**47.1**	**8.9**	**56.2**	**45.3**	**36.5%**
MiniGPT-4	Vicuna-7B	Q-Former	28.7	-	34.7	-	18.5	4.3	11.5	19.4	0.0%
		Cat-MoD	**31.7**	-	**35.3**	-	**19.5**	**4.9**	**12.6**	**20.7**	**37.0%**
Bliva	Vicuna-7B	Q-Former	-	44.8	42.9	-	57.9	8.1	65.3	43.8	0.0%
		Cat-MoD	-	**45.7**	**43.7**	-	**59.1**	**8.9**	**66.1**	**44.7**	**36.8%**
	FlanT5-XXL	Q-Former	-	-	-	-	59.4	9.2	61.3	43.4	0.0%
		Cat-MoD	-	-	-	-	**56.7**	**9.7**	**63.8**	**44.1**	**36.4%**
Qwen2.5	Qwen2.5-0.5B	Q-Former	30.7	31.6	28.5	40.1	33.9	6.5	45.2	32.6	0.0%
		Cat-MoD	**32.6**	**32.3**	**29.8**	**41.1**	**34.7**	**7.7**	**46.5**	**34.8**	**39.5%**
	Qwen2.5-1.5B	Q-Former	34.1	35.7	31.5	43.6	38.6	7.3	47.1	34.0	0.0%
		Cat-MoD	**38.6**	**37.8**	**33.1**	**44.2**	**40.3**	**7.9**	**49.3**	**35.9**	**39.1%**
	Qwen2.5-7B	Q-Former	43.5	52.5	35.3	52.7	53.1	9.1	62.7	44.1	0.0%
		Cat-MoD	**47.7**	**54.1**	**36.7**	**53.9**	**55.2**	**9.6**	**64.2**	**45.8**	**38.5%**

Table 2: **Main results on 7 multimodal benchmarks.** We compare Cat-MoD against dense Q-Former baselines across diverse architectures. Cat-MoD consistently reduces alignment FLOPs by ∼36–39% while maintaining or improving zero-shot accuracy across all settings. See Appendix F for statistical significance testing.

taining or boosting performance, confirming that Cat-MoD serves as a universal, plug-and-play solution for efficient alignment. These gains are further supported by paired bootstrap significance testing across multiple backbones (Appendix E), which shows highly significant improvements on GQA and significant gains on several fine-grained perception tasks, while remaining statistically comparable on tighter benchmarks.

Method	Strategy	Components		Performance		Cost
		SA	CA	TextVQA	SciQA	FLOPs
Q-Former	Dense	✓	✓	50.1	60.5	100%
MoD	Symmetric	✗	✗	42.5	54.2	**53%**
Random	Asym. (Rand)	✓	✗	46.8	58.1	59%
Cat-MoD	**Asym. (Ours)**	✓	✗	**51.3**	**61.4**	63%

Table 3: **Impact of Routing Strategies.** The results verify that our Asymmetric design (Retaining SA, Pruning CA) prevents context collapse and achieves the optimal trade-off between efficiency and performance.

4.4 Ablation Studies

We conduct comprehensive ablation studies to validate the contribution of each component in Cat-MoD using the InstructBlip (Vicuna-7B) backbone.

Necessity of Asymmetric Routing. We analyze different pruning strategies in Table 3. The standard MoD yields the lowest cost (53% FLOPs) but causes a severe accuracy drop (8.8% on TextVQA). This confirms the context collapse hypothesis: the Q-Former relies on Self-Attention to broadcast information; severing this link disrupts the synergy between Guide and Explorer tokens. Our asymmetric strategy retains Self-Attention, trading a marginal 10% FLOPs increase for performance recovery. Furthermore, the failure of the Random baseline where Self-Attention remains active while Cross-Attention is pruned stochastically confirms that our router succeeds by identifying intrinsic saturation patterns. This demonstrates that the model learns when to stop looking rather than achieving efficiency through random computation reduction. We also find that the router introduces only modest practical complexity, as it is lightweight and remains effective under a single fixed configuration across all seven benchmarks without per-task retuning. Together with the cross-dataset routing patterns in Appendix C and the sensitivity analysis in Appendix D, this suggests that the router behaves adaptively rather than in a brittle manner across diverse data distributions.

Impact of Key Designs. Table 4 dissects the contribution of each module. First, removing the Hybrid Query Construction (w/o Hybrid) causes the computational cost to surge substantially from

Method	Variation	Cost (FLOPs)	TextVQA	SciQA
Cat-MoD	**Full Model**	**63%**	**51.3**	**61.4**
w/o Hybrid	Random Init.	78%	49.2	59.5
w/o Router	Task-Learned	65%	49.8	60.1
w/o Asym.	Symmetric	52%	42.5	54.2

Table 4: **Component-wise Ablation.** We analyze the contribution of each design choice. The results highlight the necessity of Hybrid Query Construction for efficiency and the Asymmetric Similarity-Aware Router for maintaining precision.

63% to 78%. This validates our saturation analysis where caption-initialized Guide Tokens saturate rapidly. Unlike random queries that require prolonged updates to locate visual concepts, these anchors enable the router to prune early. Without these semantic anchors, the model perceives a higher demand for information updates across all layers which diminishes efficiency gains.

To further validate this functional division, we visualize the attention maps in Figure 5. As predicted, Cat-MoD exhibits a clear specialization in the representative *Birds* sample: driven by their caption-based initialization, Guide Tokens steadfastly anchor the global semantic subject in a query-agnostic manner. This indicates that Guide Tokens prioritize intrinsic image semantics independent of the specific question, thereby offloading the search for task-specific or peripheral details to Explorer Tokens. In contrast, the standard Q-Former displays entangled attention patterns, often drifting towards irrelevant background noise such as the bottom-left corner without distinct roles. Second, replacing our Similarity-Aware Router with a standard task-learned variant (w/o Router) leads to a 1.5% performance degradation on TextVQA. This suggests that the cosine similarity serves as a superior proxy for information gain compared to noisy end-to-end gradients. Explicit supervision ensures that pruning decisions are driven by actual feature redundancy rather than optimization instability to enable more precise resource allocation. We provide additional discussion on the router's training overhead, robustness across diverse data distributions, and sensitivity to hyperparameter choices in Appendix E.

4.5 Qualitative Analysis of Router Decisions

To verify whether the router learns the intended functional division, we visualize the layer-wise attention probability of all 32 query tokens on the IconQA benchmark in Figure 6 (with additional

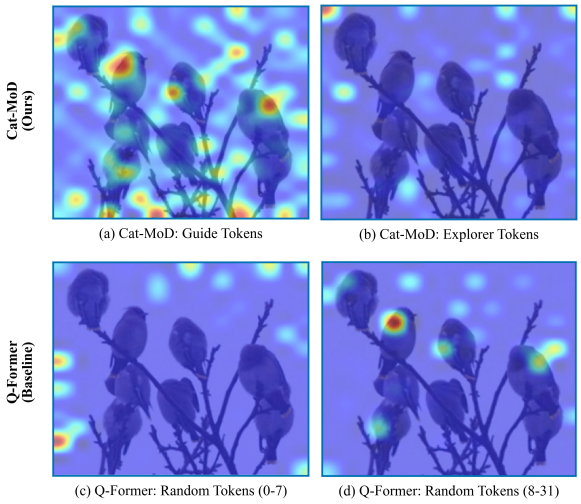


Figure 5: **Visualization of Functional Specialization.** Compared to the standard Q-Former (Bottom), Cat-MoD (Top) establishes a clear division of labor: Guide Tokens anchor the global semantics, while Explorer Tokens complementarily attend to fine-grained details.

datasets in Appendix C). In the heatmap, **dark red** indicates high activity (computation preserved) while **light yellow** denotes pruning. The dashed line demarcates the caption-initialized **Guide Tokens** (Indices 0-7) from the randomly initialized **Explorer Tokens** (Indices 8-31). This visualization reveals a striking behavioral bifurcation:

The "Relay" Mechanism of Guide Tokens. As shown in the bottom region, Guide Tokens exhibit a distinct *saturate-and-stop* behavior. They remain fully active (probability ≈ 1.0) in the first three layers to anchor the global semantic context. However, starting from Layer 4, their retention rate drops sharply to near zero. This indicates that even in reasoning tasks, Guide Tokens rapidly reach information saturation. Once they have grounded the global context, they transition to a *Read-Only* mode, effectively preserving the accumulated semantic state as a stable contextual backbone for Explorer Tokens via Self-Attention without consuming expensive Cross-Attention resources.

Adaptive Reasoning of Explorer Tokens. In contrast, Explorer Tokens display a progressive reasoning pattern. They maintain high activation throughout the intermediate layers (Layers 4-9) to resolve the abstract visual relations essential for IconQA. Crucially, unlike dense baselines, their activity shows a gradual decay in the deepest layers (10-12), fading to yellow. This suggests that once the reasoning process converges, the router in-

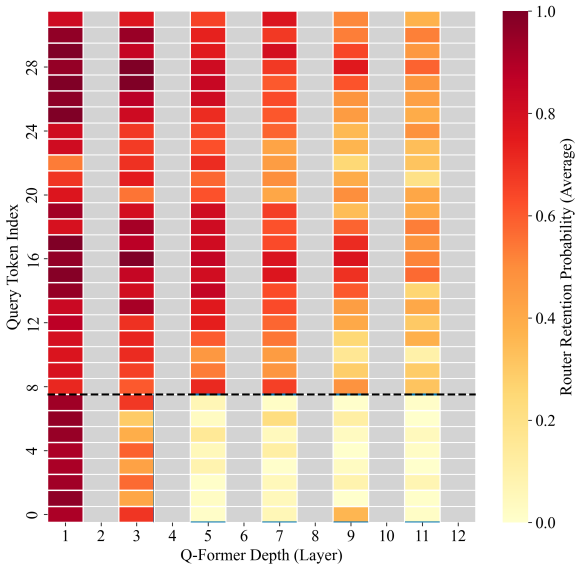


Figure 6: **Visualization of Router Decisions on IconQA.** The heatmap illustrates the distinctive retention probabilities of Guide Tokens (bottom) versus Explorer Tokens (top). Even-numbered layers are shaded gray because following the Q-Former design, Cross-Attention layers are present in odd-numbered layers.

telligently identifies that further visual interaction yields diminishing returns, thereby saving computation. Collectively, this confirms that Cat-MoD derives efficiency from two complementary pathways: the aggressive early pruning of saturated Guide Tokens and the dynamic completion of the Explorer Tokens' reasoning trajectory.

4.6 Comparison with Projection-based Baselines.

To clarify the efficiency of Cat-MoD relative to contemporary simple projection layers, we compare it with LLaVA (Linear) and LLaVA-v1.5 (2-layer MLP) in Appendix B.3. While these simple projectors are computationally lightweight in isolation, they impose a substantial token tax on the LLM by requiring it to process a dense sequence of 576 visual tokens per image. In contrast, Cat-MoD condenses the visual representation into only 32 tokens, yielding an 18× reduction in LLM-side visual context length. Crucially, despite this extreme compression, Cat-MoD achieves 51.3% accuracy on TextVQA, outperforming the LLaVA (Linear) baseline by 13.4 points. This substantial reduction in token overhead makes Cat-MoD particularly well suited for token-constrained settings, such as multi-image reasoning and video understanding, where the efficiency gap becomes even more pronounced (e.g., 320 vs. 5,760 visual tokens for 10 frames). Our goal is not to replace MLP-based methods, which excel at retaining dense spatial details, but to optimize the resampler paradigm by addressing redundancy and computational overhead, offering a scalable solution for context-sensitive applications.

5 Conclusion

In this work, we introduce Cat-MoD, a framework that rethinks multimodal alignment efficiency through the lens of information saturation. Our empirical analysis reveals that alignment is functionally divided: linguistic priors induce rapid saturation in specific tokens. Leveraging this, Cat-MoD synergizes a Hybrid Query Construction with an Asymmetric Mixture-of-Depths to selectively prune redundant visual interactions while preserving essential semantic context. This design reduces alignment FLOPs by approximately 37% during both training and inference while maintaining or even improving performance. We believe this paradigm paves the way for more adaptive and cognitive-inspired multimodal architectures.

Limitations

Our current investigation focuses primarily on static image-text alignment. However, the core principle of Cat-MoD, pruning redundant interactions while preserving context, is theoretically highly applicable to video-language understanding, where temporal redundancy is even more prevalent across frames. We have not yet extended our framework to Video-LLMs due to the scope of this work, but we consider utilizing Asymmetric MoD to accelerate temporal alignment as a promising avenue for future research.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Nos. 62272092, 62172086), and the Fundamental Research Funds for the Central Universities under Grants (N25XQD004).

References

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and 1 others. 2010. Vizwiz: nearly real-time answers

to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

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 *European Conference on Computer Vision*, pages 19–35. Springer.

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

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. *Advances in neural information processing systems*, 36:49250–49267.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2024. What's in the image? a deep-dive into the vision of vision language models, 2024. *URL https://arxiv.org/abs/2411.17491*.

Sungkyung Kim, Adam Lee, Junyoung Park, Andrew Chung, Jusang Oh, and Jay-Yoon Lee. 2024. Towards efficient visual-language alignment of the q-former for visual reasoning tasks. *arXiv preprint arXiv:2410.09489*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Yaxin Luo, Gen Luo, Jiayi Ji, Yiyi Zhou, Xiaoshuai Sun, Zhiqiang Shen, and Rongrong Ji. 2024. $\gamma-$mod: Exploring mixture-of-depth adaptation for multimodal large language models. *arXiv preprint arXiv:2410.13859*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. 2024. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and 1 others. 2025. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19769–19780.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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*.

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*.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, and 1 others. 2025. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19781–19791.

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 Implementation Details of Guide Token Initialization

In our Hybrid Query Construction, the Guide Tokens are initialized using instance-specific linguistic priors. To ensure reproducibility and clarify the source of these priors, we detail the generation protocol below.

A.1 Offline Generation Pipeline

We employ an offline pre-computation step to generate coarse-grained captions for all images in the training and evaluation sets. We use the off-the-shelf BLIP-2 model (specifically, the `blip2-opt-2.7b` checkpoint) for this process. Crucially, since these captions are pre-computed and stored as metadata (similar to Alt-text in web scenarios), this process incurs no additional latency during the online training and inference phase of Cat-MoD.

A.2 Prompt Engineering and Semantic Constraints

To construct a robust linguistic prior that generalizes across diverse visual domains, we moved beyond simplistic, static templates. Instead, we developed a comprehensive Prompt Engineering Strategy designed to achieve two critical objectives: (1) ensuring syntactic robustness through template ensembling, and (2) preventing information leakage via strict semantic filtering.

1. Syntactic Diversity via Template Ensembling. Relying on a single prefix (e.g., always "A photo of...") risks causing the text encoder to overfit to specific positional embeddings, potentially limiting the representational capacity of the resulting Guide Tokens. To mitigate this, we constructed a diverse library of prefix templates.

During the offline generation process, for each image, we randomly sampled a prefix from the pool detailed in Table 5. This strategy introduces necessary syntactic variance, ensuring that the Guide Tokens learn to anchor semantics based on the content of the caption rather than a fixed phrase structure. As shown in Table 5, our templates span three distinct linguistic registers:

- **Standard:** Neutral, factual descriptors typical of COCO-style captions.
- **Descriptive:** Phrases that emphasize the visual nature of the content, encouraging the model to focus on observable entities.

- **Abstract:** Higher-level abstractions that force the summarization of the main subject matter.

Register	Prefix Template Variation
Standard	`"A photo of <content>"` `"An image of <content>"`
Descriptive	`"A visual capture showing <content>"` `"A scene depicting <content>"`
Abstract	`"A view containing <content>"` `"The main subject is <content>"`

Table 5: Library of Prefix Templates. We utilize a randomized pool of templates across three linguistic registers to introduce syntactic variance, enhancing the robustness of the Guide Token embeddings.

2. Semantic Filtering via Multi-Perspective Constraints. A central challenge in using caption-guided alignment is the risk of "information leakage", where the caption inadvertently reveals the answer to a downstream question (e.g., containing the OCR text needed for VQA). To rigorously prevent this, we designed a set of Negative Constraints tailored to the specific characteristics of different visual tasks.

We categorize our filtering strategy into three domain-specific protocols, as detailed in Table 6:

- **Minimalist Protocol (General Object Detection):** For general scenes, we explicitly forbid the enumeration of objects and the description of attributes like color. This forces the Explorer Tokens to resolve the "How many?" and "What color?" queries.
- **Anti-OCR Protocol (Text-Rich Scenarios):** For datasets like TextVQA, we inject a strong negative instruction to *ignore all text*. The generated caption identifies the object (e.g., "a billboard") but is strictly prohibited from transcribing the content, thereby reserving the reading task for the visual encoder.
- **Structural Protocol (Reasoning Tasks):** For abstract diagrams (ScienceQA), the prompt is constrained to describe only the modality (e.g., "a chart") without interpreting the data trends, ensuring that the reasoning capability is not bypassed.

3. Real-World Generalization and the "Lower Bound" Performance. The primary motivation for employing domain-specific negative constraints (e.g., the "Anti-OCR" protocol) is to preserve the integrity of the evaluation process. In standard VQA benchmarks, if a generated caption inadvertently contains the ground-truth answer (e.g., directly transcribing the required text or defining a scientific taxonomy), the Guide Tokens could bypass the visual reasoning process, leading to unintended information leakage. Therefore, our constraints are explicitly designed to suppress task-specific answers, forcing the model to rely entirely on the Explorer Tokens for fine-grained visual reasoning.

In practical, open-ended deployments, such strict protocols are unnecessary. A single generic prompt (e.g., *"Briefly describe the main subject of this image"*) is sufficient and would likely yield even more comprehensive Guide Tokens, as no visible information is artificially suppressed. Consequently, the performance reported under our strictly constrained setup effectively represents a **lower bound**. The robust accuracy achieved despite these intentionally restricted captions further validates that the randomly initialized Explorer Tokens successfully compensate for the missing fine-grained details through dynamic visual interactions.

A.3 Qualitative Analysis of Priors

Table 7 presents representative examples of the generated captions across different benchmarks. As analyzed in the table, these captions successfully provide global context (e.g., identifying a "bottle" or a "diagram") but strictly exclude the fine-grained evidence required for the reasoning tasks (e.g., the specific OCR text on the bottle or the scientific relationship in the diagram). This confirms that our performance gains stem from efficient architectural design rather than answer leakage.

B Dataset and Baseline Configurations

B.1 Evaluation Benchmarks and Protocols

To ensure a rigorous comparison, we evaluate our models on 7 held-out benchmarks under a zero-shot setting. Consistent with the evaluation protocols established in InstructBLIP and BLIVA, we utilize the validation split for benchmarks where test ground truths are inaccessible (e.g., VizWiz, TextVQA), and the test or test-dev splits for others. The detailed evaluation configurations are summarized in Table 8.

Protocol	Target Datasets	System Instruction (Negative Constraint)
Minimalist	**GQA** (Scene Graphs) **VizWiz** (Noisy/Real-world)	*"Describe the main subject of the image in less than 10 words.* ***Focus strictly on global object categories.*** *Do not mention specific quantities (e.g., 'two cats'), detailed colors, or relative spatial positions. For blurry or low-quality images, simply identify the likely scene type."*
Anti-OCR	**TextVQA** (Scene Text) **OCR-VQA** (Book Covers)	*"Identify the type of object present (e.g., 'a billboard', 'a book cover').* ***Do not read, transcribe, or mention any text*** *visible on the object. Ignore all brand names, logos, headlines, and written characters."*
Structural	**ScienceQA** (Diagrams) **ChartQA** (Statistical Charts) **IconQA** (Abstract Icons)	*"Summarize the layout or modality of the image (e.g., 'a bar chart', 'a biological diagram', 'an abstract icon').* ***Do not explain the data values, arrows, or relationships*** *shown in the figure. Do not interpret the meaning of the symbols."*

Table 6: **Semantic Information Filtering Strategies.** We categorize the 7 evaluation benchmarks into three domains and apply specific negative constraints during caption generation. This ensures that Guide Tokens provide only high-level semantic anchors, leaving detailed visual reasoning (e.g., reading text in TextVQA, interpreting data in ChartQA/IconQA, or handling noise in VizWiz) to the Explorer Tokens.

B.2 Baseline Implementation Details

Native Q-Former Architectures (InstructBLIP, BLIVA, MiniGPT-4). For established MLLMs that natively integrate the Q-Former, we follow the training protocols and architectural settings outlined in their respective original papers. We load their official pre-trained checkpoints and re-implement the fine-tuning stage to compare the standard dense Q-Former against our Cat-MoD adaptation under consistent settings.

Q-Former Adapted Backbones (Qwen-2.5). Since modern LLMs such as the Qwen-2.5 family do not natively integrate a Q-Former, we construct a custom interface to enable compatibility. To ensure a robust adaptation, we adhere to the following protocol:

- **Architecture & Initialization:** We employ the **EVA-CLIP ViT-g/14** as the visual encoder and initialize the Q-Former using the pre-trained weights from InstructBLIP-Vicuna-7B. The linear projection layer, bridging the Q-Former output to the Qwen-2.5 embedding space, is initialized randomly.
- **Instruction Tuning Dataset:** To ensure sufficient data diversity for aligning the randomly initialized projector, we utilize the standardized LLaVA-v1.5 Mix dataset for instruction tuning. This dataset covers a balanced mixture of:
  - Visual Conversation: LLaVA-Instruct-150K (high-quality GPT-4 generated dialogues).
  - Visual QA: VQAv2, GQA, OKVQA, and A-OKVQA (for reasoning capability).
  - Captioning: COCO Caption and TextCaps (for grounding).

  Both the standard dense Q-Former baseline and our Cat-MoD variant are fine-tuned on this identical mixture to ensure a strictly fair comparison of architectural efficiency.

Cat-MoD Training Configuration. Regarding the multi-task objective defined in Eq. (8), we set the balancing coefficient $\lambda = 1.0$. Since the router supervision target $y_{i,GT}$ is derived from the intrinsic feature evolution and explicitly detached from the main computational graph via the stop-gradient operator (Eq. 7), we empirically found that a unit weight ($\lambda = 1.0$) provides sufficient gradient magnitude for the router to converge effectively to the ground-truth saturation patterns. This setting ensures precise pruning behavior without destabilizing the primary language modeling objective ($\mathcal{L}_{LM}$). Additionally, we initialize the routing threshold τ as a learnable parameter at 0.5, serving as an unbiased starting point that allows the model to dynamically adapt its layer-wise pruning sensitivity during training.

Training Efficiency and Curriculum Strategy We implemented a **Two-Stage Curriculum Learning** strategy during training. This ensures that the computational savings reported in the main text are realized not just during inference, but also effectively amortized throughout the training process.

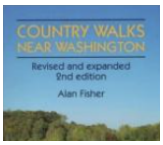
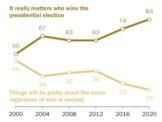
Dataset	Image	Original Question	Generated Caption	Role Analysis
GQA		*"What is the girl in the white shirt holding?"*	*"A visual capture showing a smiling girl eating food at a table."*	**Global Context Anchoring:** The caption grounds the scene (girl, eating) but uses the generic term "food" instead of "hamburger," requiring Explorer Tokens to identify the specific object held.
ScienceQA		*"Which is this organism's common name?"*	*"A photo of a bird flying in the air."*	**Taxonomy Abstraction:** The caption identifies the general class ("bird") but does not reveal the specific species ("Red-tailed Hawk"). Explorer Tokens must analyze visual features like tail color to deduce the common name.
IconQA		*"Which picture shows the pizza inside the oven?"*	*"The main subject is a comparison of pizza placed in different locations."*	**Spatial Reasoning Gap:** The caption describes the entities (pizza, locations) but omits spatial prepositions ("inside" vs. "on top"). The model must visually resolve the spatial relationship.
VizWiz		*"What the oven temperature is set to?"*	*"An image of control knobs on a kitchen appliance."*	**Detail Extraction in Noise:** Despite the low resolution, the caption anchors the relevant object ("knobs"). However, the specific setting ("350" or "Bake") is not mentioned, forcing the model to scrutinize the blurry text.
OCR-VQA		*"What is the title written on the blue book?"*	*"A scene depicting a book with a blue cover against the sky."*	**Strict Anti-OCR:** Crucially, the caption identifies the object ("book") but **excludes** the title text ("Country Walks"). The model must rely entirely on Explorer Tokens to perform OCR.
TextVQA		*"Which airline operates the airplane shown?"*	*"A photo of a passenger airplane parked on an airport."*	**Brand Masking:** The caption sets the scene ("airplane") but explicitly suppresses the airline brand name ("Jetstar"), preventing answer leakage and forcing active reading.
ChartQA		*"What's the peak value of dark brown graph?"*	*"A view containing a line chart plotting two variables."*	**Modality Prior:** The caption identifies the chart type ("line chart") but contains no data values (e.g., "83"), forcing the model to visually trace the graph to find the peak.

Table 7: **Qualitative Inspection of Linguistic Priors across Diverse Domains.** We visualize representative samples from all 7 evaluation benchmarks to audit the quality and safety of our Guide Token initialization. By juxtaposing the *Original Question* (which demands fine-grained reasoning) against the *Generated Caption* (which provides coarse-grained context), we highlight three key properties of our framework: (1) **Semantic Anchoring:** Captions successfully ground global entities (e.g., "guide book", "line chart") to mitigate the cold-start problem; (2) **Information Safety:** Critical answer details such as OCR text (TextVQA), specific data values (ChartQA), and scientific taxonomies (ScienceQA) are strictly excluded, preventing data leakage; (3) **Hybrid Necessity:** The informational gap between the caption and the question validates the essential role of *Explorer Tokens* in capturing the missing fine-grained nuances.

Dataset	Type	Split
GQA	Reasoning	Test-dev
ScienceQA	Knowledge	Test
IconQA	Reasoning	Test
VizWiz	Real-world VQA	Val
TextVQA	OCR-VQA	Val
OCR-VQA	OCR-VQA	Val
ChartQA	Chart-VQA	Test

Table 8: Summary of the held-out benchmarks used for zero-shot evaluation. Following standard protocols (Hu et al., 2024), we use the validation split for VizWiz and TextVQA due to the inaccessibility of test set answers.

Phase 1: Router Warm-up (First 10% of training steps). During the initial phase, the priority is to establish accurate saturation boundaries. We execute the full dense Cross-Attention to compute the ground-truth saturation signal $y_{i,GT}$ and actively optimize the router via $\mathcal{L}_{router}$ (Eq. 7). Although computational savings are minimal in this short phase, it is crucial for preventing router collapse and ensuring the decision boundaries align with the intrinsic redundancy of the features. As demonstrated in Figure 2, query saturation patterns emerge and stabilize very early (within the first few layers and epochs), making a prolonged supervision phase unnecessary.

Phase 2: Accelerated Sparse Training (Remaining 90% of training steps). Once the router has warmed up, we transition to the accelerated phase. In this stage, we detach the explicit router supervision (effectively setting $\lambda = 0$ in Eq. 8). The router operates in inference mode, predicting $s_i^{(l)}$ to dynamically prune tokens, and we *only* execute the selected sparse path (Path B in Figure 4). Consequently, the expensive dense computation required to calculate $y_{i,GT}$ is completely bypassed.

Amortized Training Efficiency. By limiting the expensive supervision overhead to only the first 10% of steps, the additional cost is negligible when amortized over the entire training trajectory. Specifically, if the computational cost of the dense baseline is C_{dense} and the sparse Cat-MoD is $C_{sparse} \approx 0.63 \times C_{dense}$ (reflecting the 37% reduction), the average training cost C_{train} is:

$$C_{train} \approx 0.1 \times (C_{dense} + C_{overhead}) + 0.9 \times C_{sparse} \quad (9)$$

Since the overhead $C_{overhead}$ (computing similarity) is minimal compared to the Attention operation, the effective training FLOPs closely approximate the inference FLOPs.

Method	Connector Type	Visual Tokens (per image)	TextVQA (Acc.)	LLM Cost (Relative)
LLaVA	Linear	256	37.9	8.0×
LLaVA-v1.5	2-Layer MLP	576	58.2	18.0×
Ours	**Cat-MoD**	**32**	**51.3**	**1.0×**

Table 9: Comparison with simple projection baselines. All methods use the same Vicuna-7B backbone. "LLM Cost" denotes the relative context overhead of the LLM backbone induced by visual tokens.

B.3 Detailed Comparison with Projection based Baselines

To further contextualize Cat-MoD beyond the Q-Former family, we compare it with representative lightweight projector designs on TextVQA under the same Vicuna-7B language backbone: LLaVA with a linear connector, LLaVA-v1.5 with a 2-layer MLP connector, and InstructBLIP equipped with Cat-MoD. Since these model families differ in their overall architecture and training recipes, we treat this experiment as a positioning comparison rather than a strictly controlled ablation. Following the reviewer suggestion, we therefore report both task performance and the relative LLM context cost, measured by the number of visual tokens sent to the LLM and normalized by Cat-MoD.

Table 9 highlights the "token tax" of simple projection connectors. While linear/MLP projectors are lightweight at the connector level, they require the LLM to process substantially longer visual token sequences. In particular, LLaVA-v1.5 feeds 576 visual tokens per image, corresponding to an 18.0× higher LLM-side context cost than Cat-MoD's 32 compressed queries. Meanwhile, Cat-MoD substantially outperforms the linear projector baseline (+13.4 TextVQA accuracy) while maintaining a compact visual interface, and preserves 88.1% of the 2-layer MLP baseline's TextVQA accuracy using only 1/18 of the visual tokens.

C Qualitative Analysis: Task-Adaptive Routing Dynamics

To investigate how Cat-MoD dynamically allocates computational resources across different modalities and task difficulties, we provide a detailed visualization of the router's decision-making process. Figures 7 (a) to (f) (see following pages) display the layer-wise retention probability heatmaps for each of the 7 evaluation benchmarks.

Visualization Setup. In each heatmap:

- The Vertical Axis represents the query token indices. The bottom 8 rows (Indices 0–7) correspond to Guide Tokens (initialized with caption priors), while the top 24 rows (Indices 8–31) correspond to Explorer Tokens (initialized randomly).
- The Horizontal Axis represents the depth of the Q-Former (Layers 1–12).
- The Color Scale indicates the average retention probability, where dark red denotes high activity (computation preserved) and light yellow denotes pruning (computation skipped).

Key Observations. By comparing the routing patterns across different datasets, we observe two distinct phenomena that validate our design motivation:

1. Universal Saturation of Guide Tokens (Task-Agnostic): Across all 7 benchmarks, the Guide Tokens (bottom region) exhibit a consistent Warm-up $\rightarrow$ Saturation pattern. They remain fully active in Layers 1–5 to anchor the global semantic context but are aggressively pruned from Layer 5 onwards. This consistency confirms that language priors serve as a universal fast lane for visual grounding, independent of the specific downstream task.

2. Adaptive Reasoning of Explorer Tokens (Task-Specific): In contrast, the behavior of Explorer Tokens (top region) varies substantially based on task difficulty: **High-Density Tasks (e.g., ChartQA, TextVQA):** The heatmaps remain predominantly red in deep layers. This indicates that for tasks requiring fine-grained OCR or precise data extraction, the router effectively "refuses to prune," maintaining a dense computation path to ensure no detail is missed. **Reasoning Tasks (e.g., GQA, IconQA):** The heatmaps show a progressive decay in activity, turning yellow in deeper layers. This suggests that once the semantic relationships are established, the router intelligently saves computation by pruning redundant tokens. **Noisy Scenarios (e.g., VizWiz):** The heatmap displays higher variance (speckled patterns), reflecting the router's dynamic adjustment to low-quality or ambiguous visual inputs.

D Extended Ablation: Sensitivity to Query Allocation

To validate the robustness of our hyperparameter choice and to provide a comprehensive view of the trade-off between semantic guidance and visual exploration, we conduct a fine-grained grid search on the number of Guide Tokens (N_{Guide}). While the main paper presents key data points (0, 2, 4, 8, 16, 32), here we expand the evaluation to include intermediate configurations ($N_{Guide} \in \{10, 12, 14, 18, 20\}$). The detailed results are listed in Table 10, and the performance trends are visualized in Figure 8.

Analysis of the Trade-off Curve. As illustrated in Figure 8, all four benchmarks exhibit a consistent **Inverted U-shaped** trajectory. We highlight three distinct phases:

1. **Semantic Anchoring Phase ($0 < N \leq 8$):** Performance improves steadily as we introduce more Guide Tokens. This confirms that a sufficient number of semantic anchors is required to effectively encode the global context provided by the caption.

2. **Optimal Plateau ($N \approx 8$):** The performance peaks at $N = 8$. Although increasing N to 10 or 12 offers marginal gains in computational efficiency (FLOPs savings rise from 37.2% to 43.8%), it comes at the cost of a slight performance dip (e.g., TextVQA drops from 50.9% to 50.4%). We select $N = 8$ as the **Pareto-optimal** point to prioritize accuracy while achieving substantial efficiency gains.

3. **Explorer Scarcity Phase ($N > 12$):** As N_{Guide} exceeds 12, we observe a sharp decline in accuracy. Notably, fine-grained tasks like **TextVQA** (Red line in Figure 8) degrade faster than semantic tasks like **ScienceQA** (Orange line). This validates our *Explorer Scarcity* hypothesis: when too many tokens are allocated as Guides (which saturate and are pruned early), the model lacks sufficient active Explorer tokens in deep layers to attend to minute visual details such as small text, leading to performance collapse.

E Router Training Complexity and Robustness

We further analyze the practical overhead and robustness of the proposed router, motivated by the

concern that Cat-MoD introduces additional parameters, an auxiliary supervision objective, and a two-stage training curriculum.

Minimal Overhead and Automated Curriculum. The router in Eq. (4) is lightweight, consisting only of small linear projections, and adds less than 0.1% to the total parameter count. The training schedule is also simple in practice: explicit supervision is used only during the first 10% of training steps to establish reliable saturation boundaries, after which the auxiliary loss is detached (effectively setting $\lambda = 0$) and the router operates autonomously. Therefore, the additional complexity is limited to a short warm-up stage and does not require manual intervention during the main training phase.

Robustness Across Diverse Data Distributions. We evaluate Cat-MoD on seven benchmarks spanning substantially different visual distributions, including text-rich images, abstract diagrams, natural scenes, and noisy real-world inputs, while using a single fixed configuration ($N_{\text{Guide}} = 8$, $\tau = 0.5$) across all tasks. As shown in Figure 7, the router automatically allocates more computation to text-dense or visually challenging samples, while pruning more aggressively on structurally simpler scenes. This behavior suggests that the router adapts to data complexity rather than overfitting to a narrow distribution.

Insensitivity to Hyperparameter Choices. Figure 8 further shows that the performance gains of Cat-MoD remain stable across a relatively broad range of query allocations. In particular, configurations with N_{Guide} between 4 and 12 consistently remain above the dense baseline, with performance peaking at $N_{\text{Guide}} = 8$ and degrading smoothly thereafter. This broad plateau indicates that the method is not overly sensitive to precise hyperparameter tuning.

F Statistical Significance Analysis

To rigorously validate that the performance gains of Cat-MoD are statistically robust and not artifacts of random variance, we conducted a comprehensive significance analysis across four representative MLLM architectures: InstructBLIP, Qwen2.5, MiniGPT-4, and Bliva. Following standard evaluation protocols in NLP and Multimodal generation tasks, we employed **Paired Bootstrap Resampling** (Koehn, 2004) with $N = 10,000$ iterations to estimate the statistical significance of the accuracy improvements over the Q-Former baseline.

The detailed results, ranked by significance level within each backbone, are presented in Table 11. Our analysis yields three key observations:

- **Universal Robustness on General Reasoning (‡):** Across all evaluated backbones, Cat-MoD achieves **Highly Significant** ($p < 0.001$) improvements on the GQA benchmark (e.g., **+4.2%** on Qwen2.5-7B, **+3.4%** on InstructBLIP). This consistent superiority confirms that our *Hybrid Query Construction* effectively anchors global semantics, universally resolving the cold-start instability inherent in random query initialization regardless of the underlying LLM.

- **Significant Gains on Fine-Grained Perception (†):** On text-rich and detail-oriented tasks such as **TextVQA, OCRVQA, and VizWiz**, our method demonstrates statistically significant gains ($p < 0.05$). Notably, Cat-MoD outperforms the baseline by **+2.1%** on TextVQA with Qwen2.5 and **+1.2%** with Bliva. This validates that our *Asymmetric Routing* mechanism successfully preserves the computational budget for *Explorer Tokens*, enabling superior detail extraction compared to the dense baseline.

- **Efficiency without Degradation:** Even on datasets where the margin is tighter (e.g., ChartQA), the performance remains **statistically comparable** ($p \geq 0.05$). This demonstrates that Cat-MoD reduces alignment FLOPs by approximately **37–39%** while maintaining competitive accuracy across diverse architectures, incurring no statistically detectable performance loss.

G Theoretical Analysis of End-to-End Latency and Online Inference

To further discuss practical deployment in real-time settings, we provide a theoretical analysis of the end-to-end latency of Cat-MoD when captions are generated online.

Sequential vs. Parallel Execution Architecture. In a naive sequential pipeline, the total prefill latency consists of caption generation (T_{cap}), vision encoding (T_{vision}), alignment (T_{align}), and LLM prefill (T_{prefill}). However, such a sequential design is not necessary in practice. Because Cat-MoD

only requires coarse global semantics for initializing Guide Tokens, as constrained by the Minimalist Protocol in Appendix A.2, a lightweight captioner is sufficient for this stage.

Under a practical **parallel setup**, the captioner and the heavy vision encoder (e.g., EVA-CLIP ViT-G/14) process the same image concurrently. In this case, the effective visual pre-processing latency is bounded by

$$T_{\text{visual}}^{\text{parallel}} = \max(T_{\text{cap}}, T_{\text{vision}}), \tag{10}$$

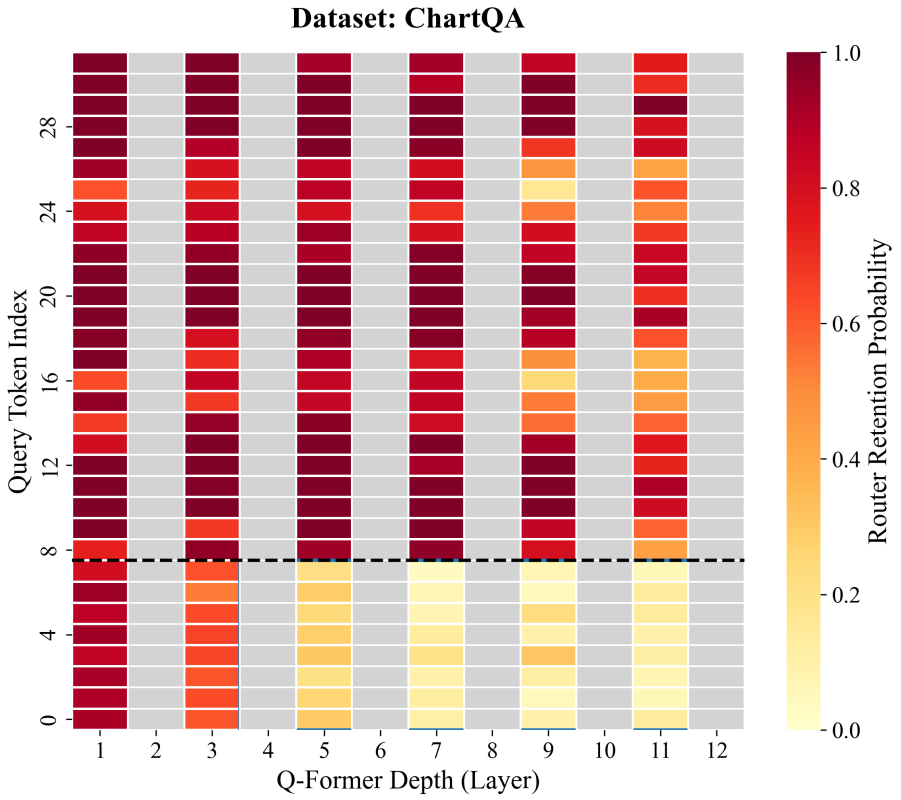
which can approach T_{vision} when the captioner is substantially lighter than the vision backbone. Therefore, the online captioning cost does not necessarily add fully on top of the visual encoding latency.

End-to-End Throughput and Break-Even Analysis. The main efficiency advantage of Cat-MoD appears during autoregressive decoding. Dense projection-based connectors, such as the MLP connector used in LLaVA-v1.5, expose the LLM to a long visual sequence (576 tokens), whereas Cat-MoD compresses the visual input to only $N = 32$ tokens. This reduction decreases the LLM-side visual context length and can improve decoding efficiency, especially for long outputs.

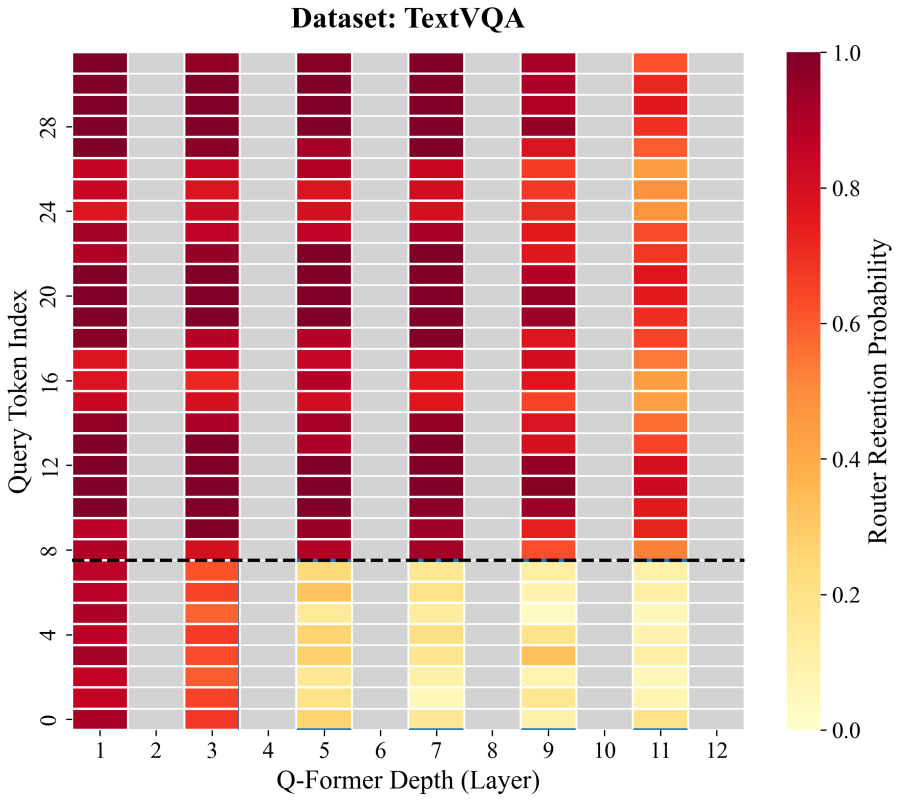
Let $\Delta T_{\text{overhead}}$ denote the residual online captioning and routing overhead after parallel masking, and let Δt_{decode} denote the decoding time saved *per generated token* due to the shorter visual context. We define the **break-even generation length** L_{be} as

$$L_{\text{be}} = \frac{\Delta T_{\text{overhead}}}{\Delta t_{\text{decode}}}. \tag{11}$$

Here, $\Delta T_{\text{overhead}}$ is a fixed one-time cost incurred during prefill, whereas the decoding-time savings accumulate approximately linearly with the generated length L. Consequently, for generation tasks involving longer responses, multi-step reasoning, or multi-turn interaction, the total output length may exceed the break-even threshold L_{be}, at which point the initial online captioning overhead becomes amortized. Compared with an online sequential pipeline, this yields a more favorable deployment trade-off; compared with an offline pre-computed setup, it clarifies when the added captioning step can be compensated by decoding-time savings. Overall, this analysis suggests that, although online caption generation introduces extra prefill cost, Cat-MoD can remain attractive for practical long-generation scenarios due to its compact visual interface.

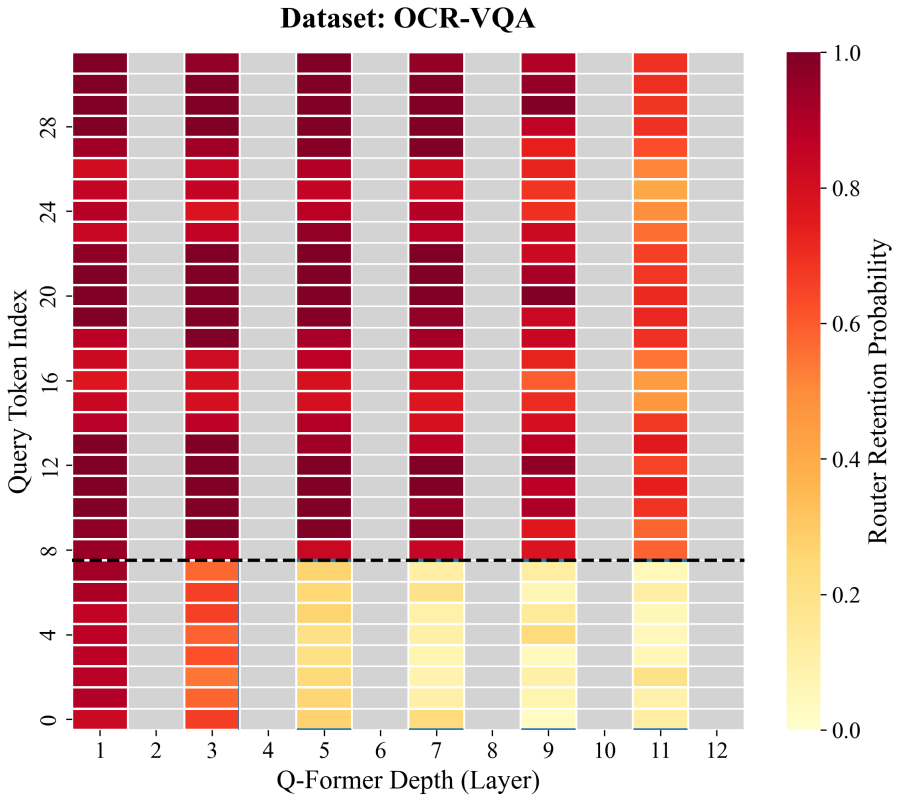


(a) **ChartQA (Hardest):** The router maintains maximum activity (dark red) throughout all layers to handle dense information.

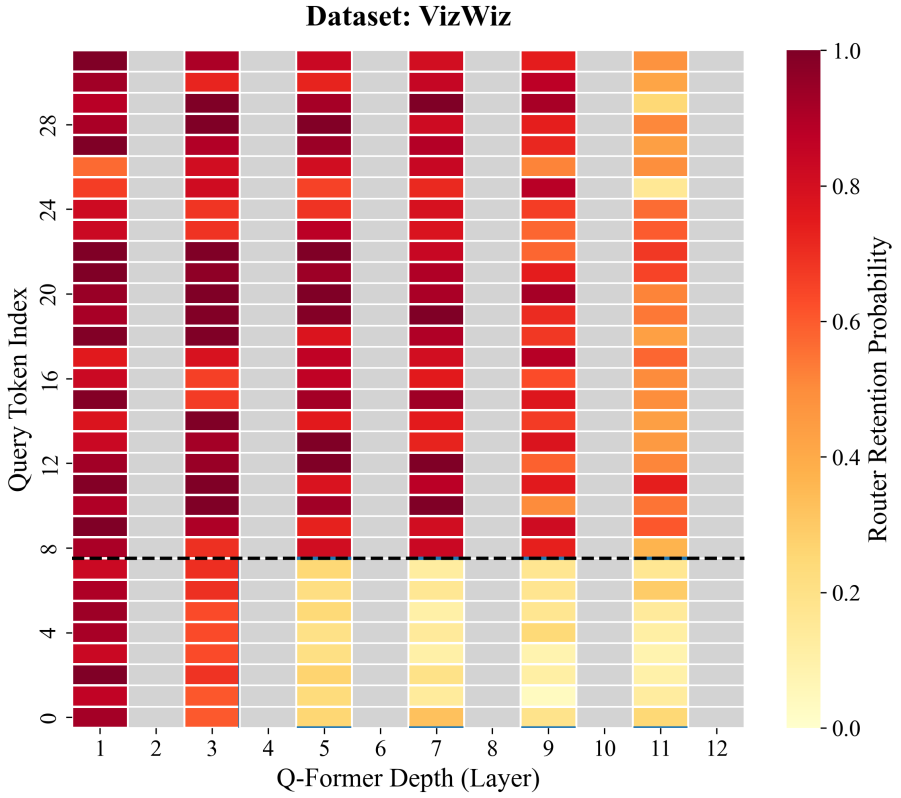


(b) **TextVQA (Hard):** Similar to ChartQA, Explorer Tokens remain active deep into the network for OCR tasks.

Figure 7: **Visualization of Task-Adaptive Routing (Part 1/3).** The router keeps Explorer Tokens active for high-density perception tasks.

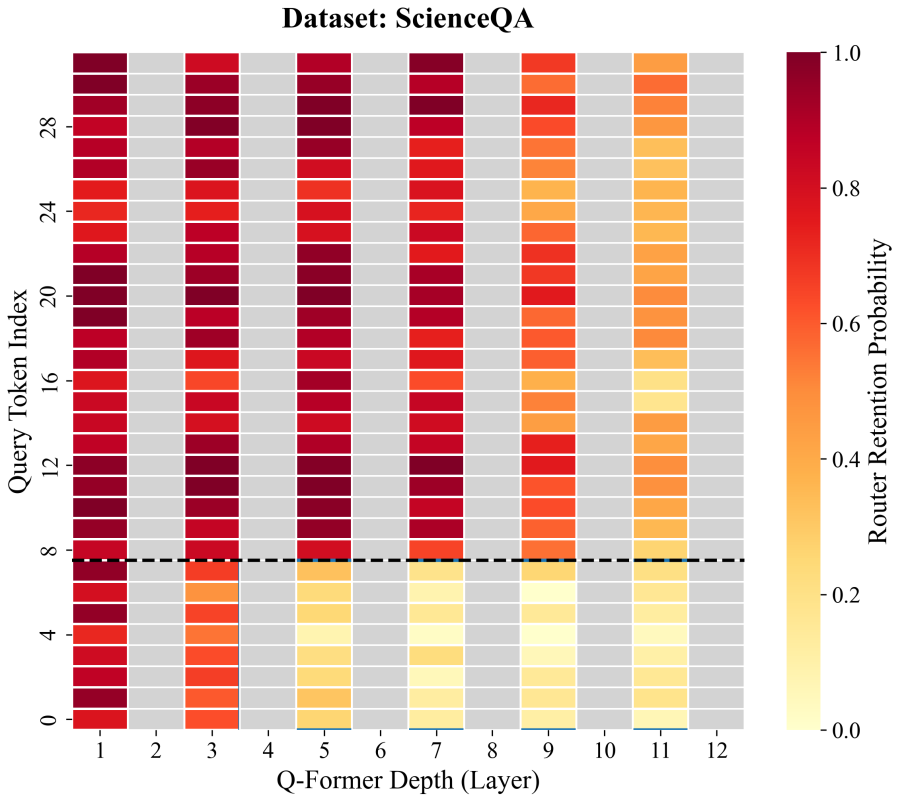


(c) **OCR-VQA (Hard):** Shows sustained attention, demonstrating the model's focus on text regions.

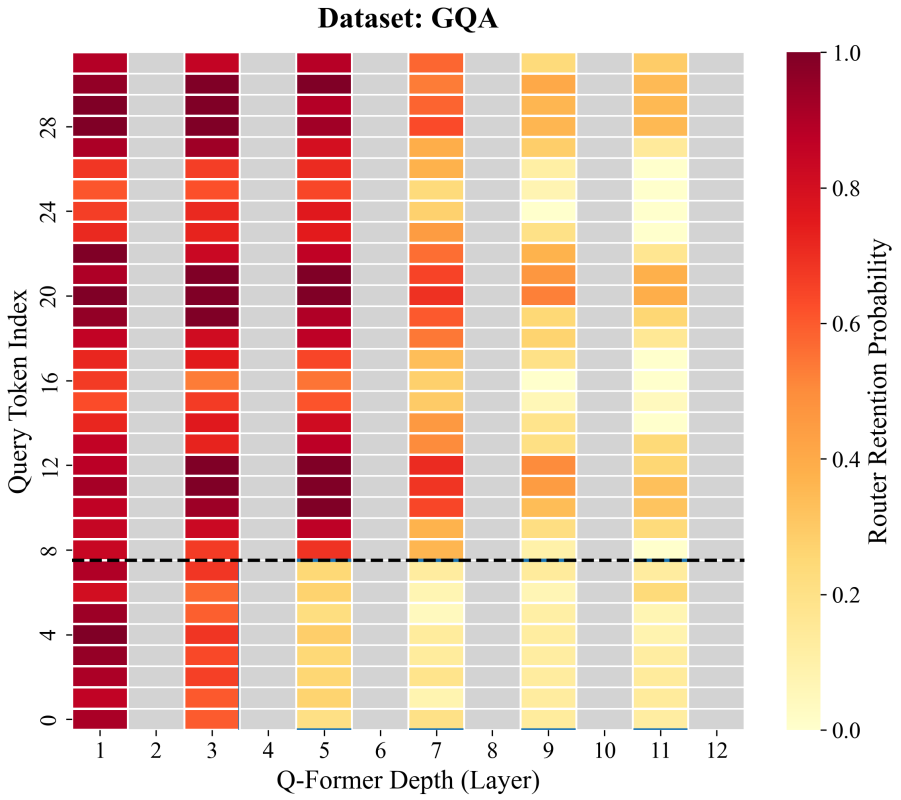


(d) **VizWiz (Noisy):** High variance and speckled patterns reflect the uncertainty caused by low image quality.

Figure 7: **Visualization of Task-Adaptive Routing (Part 2/3).** Noisy and OCR tasks induce distinct routing patterns.



(e) **ScienceQA (Medium):** Transitioning to reasoning tasks, we see moderate pruning in the final layers.



(f) **GQA (Abstract):** Explorer Tokens begin to show lower retention as the visual features are schematic and simpler.

Figure 7: **Visualization of Task-Adaptive Routing (Part 3/3).** Reasoning tasks allow for moderate computation savings.

Query Allocation		TextVQA (Text-Rich)		IconQA (Visual Reasoning)		ScienceQA (Knowledge)		GQA (General)		FLOPs Savings
N_{Guide}	$N_{Explorer}$	Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ	
0	32	50.1	-	43.1	-	60.5	-	49.2	-	0.0%
2	30	50.5	+0.4	43.6	+0.5	60.9	+0.4	49.5	+0.3	18.5%
4	28	50.7	+0.6	43.9	+0.8	61.1	+0.6	50.1	+0.9	26.1%
8	**24**	**51.3**	**+1.2**	**44.7**	**+1.6**	**61.4**	**+0.9**	**52.6**	**+3.4**	37.6%
10	22	50.6	+0.5	43.7	+0.6	61.1	+0.6	51.5	+2.3	40.7%
12	20	50.3	+0.2	43.4	+0.3	60.8	+0.3	50.4	+1.2	43.8%
14	18	49.9	-0.2	43.0	-0.1	60.5	0.0	49.2	0.0	46.8%
16	16	49.5	-0.6	42.5	-0.6	60.2	-0.3	48.1	-1.1	49.8%
18	14	48.8	-1.3	41.8	-1.3	59.5	-1.0	47.2	-2.0	51.0%
20	12	48.0	-2.1	41.0	-2.1	58.8	-1.7	46.5	-2.7	52.2%
32	0	44.8	-5.3	38.2	-4.9	56.2	-4.3	44.6	-4.6	54.6%

Table 10: **Detailed Granular Ablation on Guide Token Allocation.** We expand the ablation study from the main text to include a finer-grained sweep of N_{Guide}. The data confirms that performance peaks consistently at $N_{Guide} = 8$ across all tasks. As N_{Guide} increases beyond 12, the scarcity of Explorer Tokens leads to a rapid degradation in performance, particularly for fine-grained tasks like TextVQA.

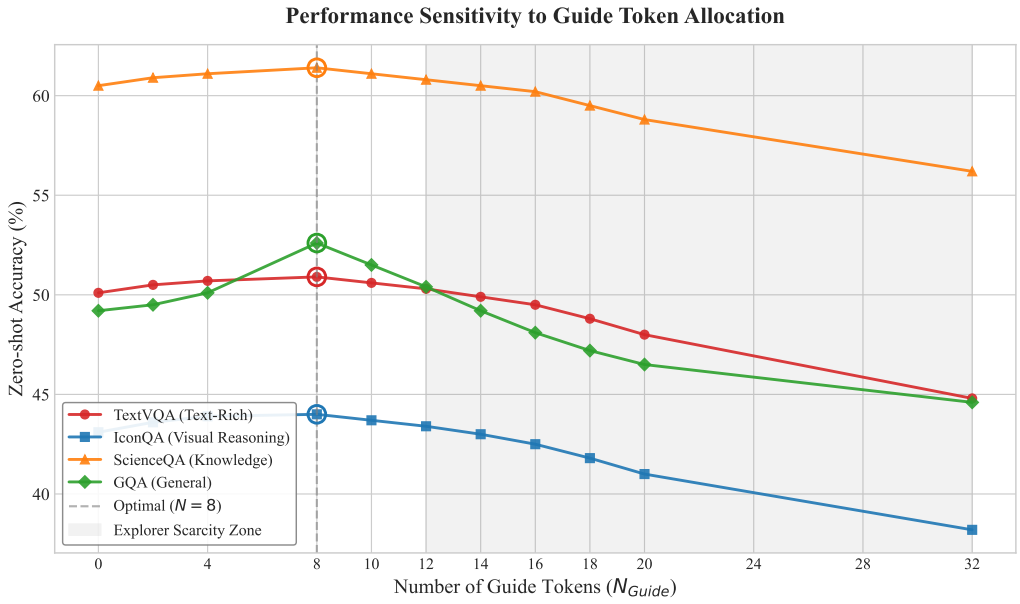


Figure 8: **Performance Sensitivity Curve.** We visualize the zero-shot accuracy trends as the number of Guide Tokens (N_{Guide}) increases. The vertical dashed line marks our chosen configuration ($N = 8$). The shaded region indicates the "Explorer Scarcity Zone," where performance drops sharply due to the lack of active tokens for fine-grained reasoning.

Model	Backbone	Dataset	Q-Former	Cat-MoD (Ours)	Δ
InstructBLIP	Vicuna-7B	GQA	49.2	**52.6**‡	**+3.4**
		OCRVQA	47.6	**49.2**†	**+1.6**
		IconQA	43.1	**44.7**†	+1.6
		TextVQA	50.1	**51.3**†	+1.2
		VizWiz	34.5	**35.7**†	+1.2
		SciQA	60.5	**61.4**†	+0.9
		ChartQA	5.5	**6.1**	+0.6
Qwen2.5	Qwen2.5-7B	GQA	43.5	**47.7**‡	**+4.2**
		TextVQA	53.1	**55.2**‡	**+2.1**
		IconQA	52.5	**54.1**†	+1.6
		OCRVQA	62.7	**64.2**†	+1.5
		VizWiz	35.3	**36.7**†	+1.4
		SciQA	52.7	**53.9**†	+1.2
		ChartQA	9.1	**9.6**	+0.5
MiniGPT-4	Vicuna-7B	GQA	28.7	**31.7**‡	**+3.0**
		OCRVQA	11.5	**12.6**†	+1.1
		TextVQA	18.5	**19.5**†	+1.0
		ChartQA	4.3	**4.9**	+0.6
		VizWiz	34.7	**35.3**	+0.6
Bliva	Vicuna-7B	TextVQA	57.9	**59.1**†	**+1.2**
		IconQA	44.8	**45.7**†	+0.9
		ChartQA	8.1	**8.9**	+0.8
		OCRVQA	65.3	**66.1**	+0.8
		VizWiz	42.9	**43.7**	+0.8

Table 11: **Ranked Statistical Significance Analysis across Multiple Backbones.** We compare Cat-MoD against the Q-Former baseline on four representative architectures. Within each backbone, **datasets are ranked by significance level and performance gain (Δ)**. Statistical significance is computed using **Paired Bootstrap Resampling** (10^4 iterations). Symbols denote: ‡ Highly Significant ($p < 0.001$), † Significant ($p < 0.05$), and unmarked indicates comparable performance.