Rethinking Cross-Modal Attention for Visual Token Pruning in Multimodal LVLMs

Diego Weng Qiu and **Lin Shang**
State Key Laboratory for Novel Software Technology
School of Computer Science, Nanjing University
Nanjing, China

Abstract

Recent studies challenge the reliability of cross-modal attention mechanisms as an indicator for visual token pruning in Large Vision-Language Models (LVLMs). In this work, we demonstrate that such negative conclusions can be substantially influenced by an overlooked phenomenon termed task-bound coupling, wherein cross-modal attention patterns become entangled with task-specific prompts. Through controlled configurations that decouple visual inputs from downstream queries, We find attention-ranked tokens have causal task relevance, with attention distributions maintaining relative consistency across evaluated settings. Concurrently, we identify an empirical scaling phenomenon which we term Attention Fidelity Decay, where the downstream effectiveness of attention-based token selection declines in higher-capacity models or advanced architectures. Motivated by these insights, we introduce TAP-LLaVA, a simple training-free framework that leverages localized task-agnostic attention for efficient, reusable visual token pruning across multi-turn sessions. Our approach achieves competitive downstream performance while substantially lowering computational overhead. This suggests that attention can provide a reliable structural prior at moderate model scales when isolated from textual confounders.

1 Introduction

Large Vision-Language Models (LVLMs) have achieved remarkable breakthroughs across a wide spectrum of multimodal tasks (Xu et al., 2025; Lu et al., 2025; Zhu et al., 2025; Bai et al., 2025). Nevertheless, their real-world deployment remains fundamentally limited by the massive computational and memory footprint of visual tokens. Unlike text tokens, hundreds of visual tokens persist throughout the context window, inducing quadratic complexity during decoding. This structural characteristic translates into high latency and memory consumption, a challenge that is particularly exacerbated in long-horizon, multi-turn dialogues where identical visual contexts must be processed repeatedly.

To alleviate this efficiency bottleneck, visual token reduction has emerged as a crucial area of inquiry. A pivotal debate within this domain centers on whether cross-modal attention scores offer a faithful signal for identifying and preserving salient visual regions. Existing literature presents conflicting paradigms: some works successfully exploit text-guided attention for semantic grounding (Huang et al., 2024; Ye et al., 2024), while others argue that attention weights are inherently unfaithful and noisy indicators due to geometric biases and positional instability (Wen et al., 2025; Zhang et al., 2025a, 2024a; Chen et al., 2024).

We argue that this gridlock stems from a fundamental evaluation artifact rather than an intrinsic flaw of the attention mechanism itself. Intuitively, a well-aligned multimodal model should establish a stable token selection pattern for any given scene before execution, a foundational text-visual alignment space learned during pre-training (Radford et al., 2021). The perceived instability reported in prior works can be significantly confounded by a phenomenon we term **task-bound coupling**. Conventional evaluation protocols employ a singular textual prompt to simultaneously dictate the downstream task objective and trigger cross-modal attention. These dual roles heavily entangle attention distributions with query-specific semantic variations, introducing extrinsic volatility that is frequently misattributed to the unreliability of the attention mechanism. This operational entanglement echoes the classic debate on attention interpretability (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), confirming that surface-level attention distributions can be easily perturbed by textual context shifts even when the underlying representation space remains structurally sound.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2886–2897
November 6-10, 2026 ©2026 Association for Computational Linguistics

To isolate the structural behavior of attention, we decouple these roles by introducing fixed, task-agnostic prompts that operate strictly as non-interfering attention probes. Under this controlled setting, we observe a trend: cross-modal attention maps exhibit a relatively high intersection-over-union (IoU) across highly diverse downstream queries. This trend suggests that intermediate layers of LVLMs preserve a relatively stable visual representation template that remains structurally informative across distinct task contexts, yet is fundamentally obscured by task-dependent noise under conventional prompting paradigms.

Concurrently, our layer-wise investigations reveal an empirical scaling trend which we term **Attention Fidelity Decay**: while task-agnostic attention weights remain useful at moderate parameter scales (e.g., 7B models), their downstream pruning efficacy declines in larger architectures (e.g., 13B) and advanced multi-scale representation spaces. We present this as an empirical observation rather than a mechanistic claim—our diagnostics do not establish why the signal degrades, and whether visual features increasingly propagate through non-attention pathways at higher capacity remains an open question.

Driven by these complementary diagnostics, we present **TAP-LLaVA**, a training-free framework designed to operationalize task-agnostic attention for efficient visual token pruning in multi-turn dialogues. Our approach executes a single attention-probing phase utilizing a fixed prompt to isolate salient visual tokens during the initial interaction. These pruned tokens are subsequently cached and frozen across all succeeding dialogue turns, establishing a hardware-friendly and amortized $\mathcal{O}(1)$ perception overhead per additional turn. TAP-LLaVA achieves an elegant trade-off between computational efficiency and task accuracy, unlocking pronounced latency reductions without parameter tuning.

The core contributions of this work are summarized as follows:

- We identify task-bound coupling as a potential confounder in cross-modal attention analysis, showing that attention distributions converge to a consistent token selection pattern when disentangled from downstream task instructions.
- We identify Attention Fidelity Decay, an empirical scaling trend in which the effectiveness of attention-based token importance signals declines with model capacity, limiting their utility for pruning in high-capacity LVLMs.
- We develop TAP-LLaVA, a training-free framework that operationalizes task-agnostic attention for visual token pruning, enabling efficient multi-turn inference with competitive performance at moderate scales.

2 Related Work

2.1 Visual Token Reduction in LVLMs

Recent efforts on improving the efficiency of large LVLMs have explored a wide spectrum of visual token reduction strategies (Shao et al.). To provide a unified perspective, we categorize existing methods along two dimensions: (i) the *source of information*, i.e., whether token selection is independent of or conditioned on textual input, and (ii) the *reduction mechanism*, i.e., whether compression is achieved via heuristic operations or learned structural modifications.

Vision-Centric Reduction. Vision-centric approaches eliminate redundancy based solely on intra-visual signals or frozen encoder priors without linguistic conditioning. Early methods exploit semantic sparsity in vision encoders, such as using `[CLS]` weights (FasterVLM (Zhang et al., 2024a)) or statistical criteria (LLaVA-PruMerge (Shang et al., 2025)). To mitigate foreground bias, subsequent works preserve complementary background regions via clustering (VisPruner (Zhang et al., 2025a), VisionZip (Yang et al., 2025)), or employ progressive matching layers (ToMe (Bolya et al., 2022)). Structural compression methods further remove redundant parameter groups through structural dependencies (YOPO (Zhang et al., 2024b), DepGraph (Fang et al., 2023)). While these prompt-agnostic methods enable stable token reuse across multi-turn sessions, the lack of textual guidance limits their adaptability to downstream task variations.

Instruction-Guided Reduction. Conversely, instruction-guided methods dynamically adjust token selection conditioned on the input text. Recent works explore removing tokens in late decoder stages (FastV (Chen et al., 2024)), constructing text-conditioned proxies (IVTP (Huang et al., 2024)), or applying learnable pruning thresholds (ATP-LLaVA (Ye et al., 2024), MADTP (Cao

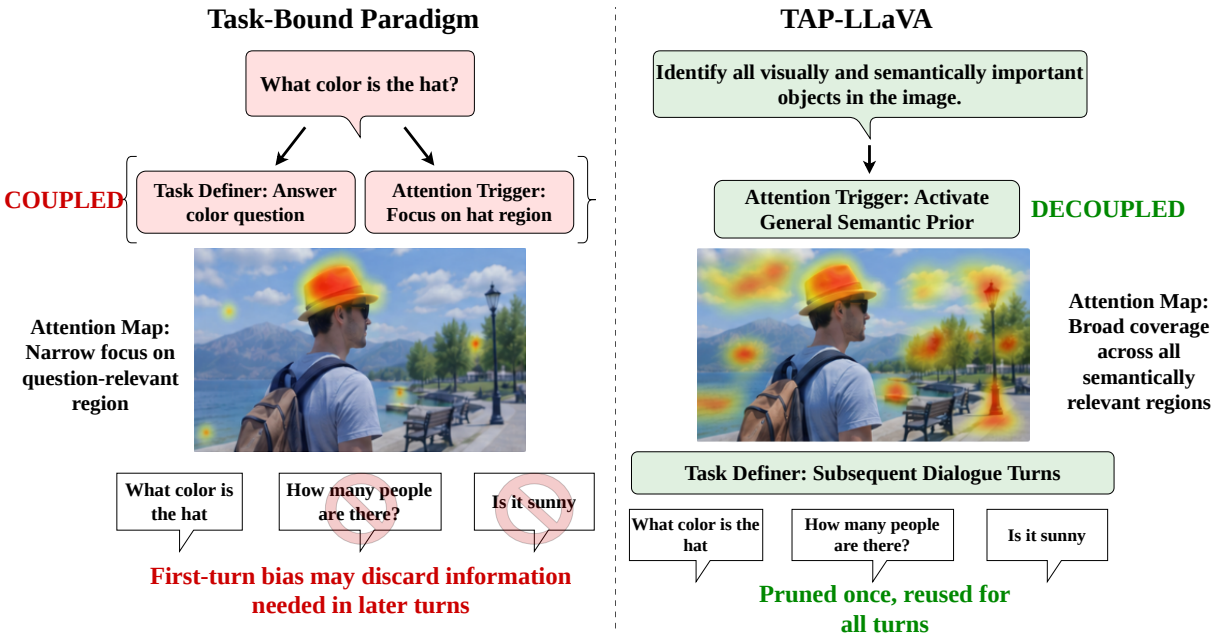


Figure 1: Framework of TAP-LLaVA. We decouple the two coupled roles of prompts in conventional paradigms: (left) Task-Bound Paradigm uses a single prompt as both task definer and attention trigger, leading to narrow, biased attention; (right) TAP-LLaVA adopts a fixed task-agnostic prompt as the dedicated attention trigger to elicit a task-decoupled, stable token selection baseline, then prunes visual tokens only once and reuses them across all dialogue turns.

et al., 2024)). Other approaches shift compression upstream by learning to compress visual features end-to-end (LLaVA-Mini (Zhang et al., 2025b), MetaCompress (Wang et al., 2026)). While achieving tight task alignment in single-turn scenarios, their selection is inherently query-dependent. Visual tokens must be re-evaluated for every new instruction, significantly hindering computational efficiency in long-horizon, multi-turn dialogues.

2.2 Revisiting Attention as a Token Importance Signal

Our work is also related to the long-standing debate on whether attention weights provide a reliable indication of model behavior. In NLP, Jain and Wallace (Jain and Wallace, 2019) argue that attention is not a faithful explanation due to weak correlation with gradient-based importance measures. In contrast, Wiegreffe and Pinter (Wiegreffe and Pinter, 2019) demonstrate that attention can encode meaningful and transferable importance under appropriate evaluation settings.

This debate has recently extended to multimodal token pruning. Several studies (Wen et al., 2025; Abnar and Zuidema, 2020; Xiao et al., 2024) demonstrate that attention-based scoring can exhibit positional and geometric biases, and that its effectiveness varies significantly across tasks. Consequently, several recent LVLM pruning methods have shifted away from cross-modal attention toward purely visual criteria (e.g., VisPruner (Zhang et al., 2025a), FasterVLM (Zhang et al., 2024a)).

2.3 Decoupling as the Missing Piece

Despite their differences, existing approaches share a common underlying assumption: the textual input simultaneously defines the task objective and serves as the trigger for cross-modal interaction. This implicit coupling introduces instability in attention-based token selection, as attention maps become entangled with query-specific semantic variations.

We argue that the limitations of cross-modal attention observed in prior work can be partly attributed to this task-bound coupling. By decoupling the *attention trigger* from the *task instruction*, it becomes possible to extract a more stable and reusable visual importance signal. This perspective bridges the gap between vision-centric transferability and instruction-guided adaptability, enabling training-free token reduction that shows consistency across the evaluated tasks and supports efficient multi-turn dialogue.

Model	r	TopK	Random	BottomK
LLaVA-1.5-7B	5%	47.73	43.71	32.96
	10%	53.68	48.58	34.09
	20%	57.87	53.23	36.07
LLaVA-1.5-13B	5%	49.89	46.77	34.31
	10%	56.29	53.00	35.11
	20%	59.68	57.13	39.07

Table 1: Causal token-deletion experiments on a 500-sample subset of MT-VQAv2 (downstream accuracy, %). TopK retains the highest-scoring tokens under the task-agnostic probe, Random retains tokens uniformly at matched budgets (quantity control), and BottomK retains the lowest-scoring tokens (negative control). The consistent ordering TopK $>$ Random $>$ BottomK indicates that attention-ranked tokens carry causal task relevance.

3 Diagnosing Task-Bound Coupling in Cross-Modal Attention

Existing token reduction methodologies implicitly adopt an evaluation paradigm where a singular text prompt simultaneously specifies the downstream task objective and drives cross-modal attention mechanisms. We argue that this dual functionality introduces a critical confounder: attention maps become heavily entangled with task-specific semantics, obscuring whether the observed spatial variability arises from the intrinsic properties of the attention mechanism or from dynamic variations in prompt formulations. Consequently, prior assertions regarding the unfaithfulness or instability of cross-modal attention may be misattributed.

To isolate these factors, we explicitly decouple the two functional roles of textual prompts—task specification and attention triggering—and systematically re-examine cross-modal attention behaviors across distinct architectural scales and paradigms. Our diagnosis proceeds in two stages: we first validate whether attention-selected tokens carry *causal* importance for downstream tasks through controlled deletion experiments (§3.2), and then supplement this evidence with selection-consistency and position-bias analyses that characterize the structural behavior of the prior (§3.3).

3.1 Experimental Setup

We evaluate LLaVA-1.5-7B/13B on 2,260 images from the MT-VQAv2 validation set under two distinct prompting configurations:

(1) Task-Bound Setting. Prompts consist of natural, question-driven inputs derived from multi-turn dialogues (e.g., "What is the color of the hat?"), which simultaneously define the execution objective and dynamically induce cross-modal attention across sequential turns.

(2) Task-Agnostic Setting. Prompts are constructed strictly as generic attention probes, independent of downstream execution targets. We implement two variants: (i) holistic prompts designed for global scene understanding, and (ii) diverse prompts engineered for generic reasoning patterns such as counting or spatial relation awareness (detailed in Appendix B).

Given L_{text} text tokens and N visual tokens, we extract cross-modal attention matrices from an intermediate layer $l = 20$, yielding $\mathbf{A} \in \mathbb{R}^{L_{\text{text}} \times N}$. The importance score for each visual token is computed via mean aggregation across the textual dimension:

$$s_j = \frac{1}{L_{\text{text}}} \sum_{i=1}^{L_{\text{text}}} \mathbf{A}_{i,j} \tag{1}$$

Tokens are then ranked by s_j, and the top-K scoring tokens are retained, where $K = \lceil r \cdot N \rceil$ and r is the visual token retention ratio.

3.2 Causal Validation via Token Deletion

Whereas the consistency metrics introduced below quantify the stability of token selection across prompt formulations, stability alone does not establish importance: a selection prior could remain perfectly invariant across prompts yet carry little task-relevant information. To directly interrogate whether the task-agnostic attention prior encodes downstream utility, we conduct controlled deletion experiments on a 500-sample subset of MT-VQAv2, evaluating downstream accuracy at three retention ratios ($r \in \{0.05, 0.1, 0.2\}$) on LLaVA-1.5-7B/13B; under the null hypothesis that attention scores carry no task-relevant information, the three regimes in Table 1 should perform comparably at equal budgets.

As reported in Table 1, the strict ordering TopK $>$ Random $>$ BottomK holds at every retention ratio on both scales. The Random regime constitutes the critical quantity control: it matches TopK in the number of retained tokens yet trails by 2.6–5.1 points, thereby isolating the contribution of the attention ranking itself. BottomK collapses by 14.8–21.8 points relative to TopK, indicating that the lowest-attention tokens carry disproportionately little task-relevant information. Taken together, these results provide causal evidence that task-agnostic

Table 2: Cross-modal attention consistency (measured via IoU %) and explicit position bias (measured via Pearson's r^2 %) on the MT-VQAv2 validation set for LLaVA-1.5-7B and LLaVA-1.5-13B across four visual token retention ratios ($r \in \{0.1, 0.2, 0.3, 0.4\}$).

	LLaVA-1.5-7B				**LLaVA-1.5-13B**			
Setting	**10%**	**20%**	**30%**	**40%**	**10%**	**20%**	**30%**	**40%**
	Cross-modal attention consistency (IoU %)							
Holistic TA	74.88	79.91	83.72	86.22	79.97	84.31	87.07	88.95
Diverse TA	68.24	75.87	81.35	85.16	72.52	79.28	83.84	87.32
Task-Bound	61.34	70.46	77.37	82.37	62.68	71.86	78.20	82.83
	Position bias (Pearson's r^2 %)							
Task-Agnostic	4.16				8.50			
Task-Bound	4.03				7.48			

attention ranks visual tokens by downstream utility, complementing the consistency analyses in §3.3. We acknowledge that deletion experiments establish utility at the level of token sets rather than of individual tokens, and that gradient-based attribution or human-attention alignment would offer finer-grained mechanistic validation, which we regard as a promising future direction; we also note that this 500-sample subset differs from the full evaluation set, so absolute accuracies are not directly comparable to the main results in Table 3.

3.3 Selection Consistency and Position Bias

Having established the causal utility of the attention-ranked prior, we now ask two further questions: does task-agnostic prompting also yield a stable selection pattern, and can that stability be attributed to geometric biases? To measure selection consistency across two prompt formulations P_a and P_b on the same visual input, we compute the Intersection-over-Union (IoU) of the retained index sets:

$$\text{IoU}(P_a, P_b) = \frac{|\mathbf{V}'(P_a) \cap \mathbf{V}'(P_b)|}{|\mathbf{V}'(P_a) \cup \mathbf{V}'(P_b)|} \qquad (2)$$

A high IoU indicates that attention converges to a stable subset of informative visual tokens regardless of prompt variation ($r \in \{0.1, 0.2, 0.3, 0.4\}$ in this analysis). One caveat: prompts within each group share stylistic templates, so part of the intra-group consistency may reflect prompt similarity. The deletion experiments above address this by testing selection utility rather than stability; a prompt-similarity-controlled IoU study is left to future work.

As Table 2 shows, task-agnostic prompting yields higher IoU than the task-bound baseline across all retention ratios and scales, with the gap widening as the retention budget grows. LLaVA-1.5-13B shows the same trend, with even higher absolute IoU than its 7B counterpart. Rather than indicating better downstream performance, we read this elevated consistency as an early sign of the structural rigidity discussed in Section 5.3, where 13B accuracy under attention-based pruning falls below simpler heuristics. We use the term as a descriptive label for the observed tendency of intermediate-layer attention maps in larger models to stabilize into query-insensitive distributions, not as a mechanistic claim.

A plausible alternative is that the consistency stems from trivial geometric biases, such as a preference for central spatial regions or fixed coordinate templates. To test this, we compute the coefficient of determination between the aggregated attention weights s_j and the spatial coordinates (x_j, y_j) of the visual tokens across all evaluation instances. As Table 2 reports, coordinates explain only a negligible fraction of the attention variance, and these values are nearly identical between the task-bound and task-agnostic settings within each model. Since coordinate-dependent biases are static while IoU diverges between the two paradigms, linear coordinate dependencies alone cannot explain the observed stability; the gains under task-agnostic prompting instead point to token selection utility. This check captures only first-order linear position effects; richer characterizations (e.g., adjacent-token proximity or attention-sink analysis) would be needed to fully rule out

geometric confounders.

4 Methodology

The analysis in Section 3 reveals that cross-modal attention, when disentangled from task-specific prompts, converges to a stable and reusable token selection prior. This observation suggests that the conventional query-dependent token reduction paradigm may be unnecessarily entangled with task execution.

In this section, we formalize a minimal and principled alternative: instead of repeatedly re-estimating token importance for each query, we extract a *task-agnostic token selection prior once* and reuse it throughout the entire interaction. This leads to TAP-LLaVA, a training-free framework that factorizes visual processing into a one-time perception stage and a reusable execution stage.

4.1 Problem Formulation

Let $\mathcal{M}$ denote a pre-trained LVLM. Given an image $\mathbf{I}$, the vision encoder produces a sequence of visual tokens:

$$\mathbf{V} = \{v_1, v_2, \ldots, v_N\} \in \mathbb{R}^{N \times D} \tag{3}$$

In a multi-turn dialogue, the model receives a sequence of task-specific prompts:

$$\mathcal{Q} = \{P_1, P_2, \ldots, P_T\} \tag{4}$$

Existing methods implicitly define token selection as a query-conditioned function:

$$\mathbf{V}'_t = f(\mathbf{V}, P_t) \tag{5}$$

which introduces two limitations: (i) repeated computation across turns, and (ii) sensitivity to task-specific variations.

We reformulate the problem by replacing the query-dependent token-selection mapping with a task-agnostic token selection function:

$$\mathbf{V}' = f(\mathbf{V}, P_{\text{TA}}) \tag{6}$$

where P_{TA} is a fixed task-agnostic prompt. This reformulation allows token selection to be performed once and reused across dialogue turns.

4.2 Prior Extraction via Task-Agnostic Attention

We instantiate the token selection function using cross-modal attention induced by a fixed prompt P_{TA}. The concatenated input $[\mathbf{V}, P_{\text{TA}}]$ is processed by $\mathcal{M}$, and we extract the cross-modal attention matrix at layer l:

$$\mathbf{A} \in \mathbb{R}^{L_{\text{TA}} \times N} \tag{7}$$

where L_{TA} denotes the number of tokens in P_{TA}.

We compute token importance scores via mean aggregation over the prompt tokens:

$$s_j = \frac{1}{L_{\text{TA}}} \sum_{i=1}^{L_{\text{TA}}} \mathbf{A}_{i,j} \tag{8}$$

Let π be the index permutation that sorts s_j in descending order. Given a retention ratio r, we define:

$$K = \lceil r \cdot N \rceil \tag{9}$$

The selected visual tokens are defined as:

$$\mathbf{V}' = \{v_{\pi(1)}, \ldots, v_{\pi(K)}\} \tag{10}$$

This selection is performed once and reused for all subsequent dialogue turns.

4.3 Prior Reuse for Multi-Turn Inference

Let P_t denote the user prompt at turn t, and let $\mathbf{R}_t$ denote the sequence of text tokens generated by the language model as the response at turn t.

For each dialogue turn t, we directly reuse $\mathbf{V}'$ without re-evaluating token importance. The model input is constructed as:

$$\mathbf{X}_t = [\mathbf{V}', P_1, \mathbf{R}_1, \ldots, P_{t-1}, \mathbf{R}_{t-1}, P_t] \tag{11}$$

where all elements are concatenated at the token level following standard multimodal input formatting.

This formulation separates visual token selection from subsequent language generation: the former is performed once using the task-agnostic prompt P_{TA}, while the latter proceeds via standard language modeling over a fixed visual context.

4.4 Efficiency Analysis

This reformulation yields two fundamental advantages over query-dependent token reduction:

Constant Perception Cost. Dynamic token reduction frameworks require recomputing attention maps for every new query, incurring a cumulative complexity of $\mathcal{O}(T)$ with respect to dialogue length. In contrast, TAP-LLaVA performs the

Model	Method	MT-VQA-v2				MT-GQA			
		Acc_1	Acc_2	Acc_3	Avg	Acc_1	Acc_2	Acc_3	Avg
LLaVA-1.5-7B	Base	72.02	73.73	75.49	73.75	73.07	76.10	76.97	75.38
	Random	60.84	61.80	63.04	61.89	70.57	71.30	68.30	70.06
	Sample	60.93	62.88	63.68	62.50	68.13	71.70	70.93	70.25
	FastV	52.38	52.33	53.21	52.64	53.73	50.20	50.50	51.48
	PruMerge	61.34	62.83	64.89	63.02	64.90	68.10	69.37	67.46
	VisPruner	64.27	65.30	66.69	65.42	66.63	70.70	71.10	69.48
	Ours	65.71	68.55	68.92	**67.73**	69.47	72.53	71.87	**71.29**
LLaVA-1.5-13B	Base	76.16	77.64	78.26	77.35	73.10	79.00	78.43	76.84
	Random	63.75	64.33	64.69	64.26	69.23	71.63	73.33	71.40
	Sample	64.16	65.97	65.55	65.22	70.03	72.67	72.80	**71.83**
	FastV	57.73	57.01	57.71	57.48	57.57	54.30	55.87	55.91
	PruMerge	63.44	65.03	67.21	65.23	65.43	68.17	71.10	68.23
	VisPruner	67.07	68.17	68.17	67.80	68.13	71.83	71.93	70.63
	Ours	67.15	67.69	69.07	**67.97**	68.93	72.10	72.67	71.23
LLaVA-NeXT-7B	Base	82.47	83.97	83.98	83.47	78.57	81.07	81.57	80.40
	Random	72.45	73.95	73.40	73.27	74.97	78.03	76.40	76.47
	Sample	74.12	73.48	73.71	73.77	74.10	77.10	77.10	76.10
	VisPruner	76.86	74.88	74.81	75.54	75.03	77.03	77.27	76.44
	Ours	74.76	76.77	75.82	**75.78**	72.80	79.50	78.10	**76.80**
LLaVA-NeXT-13B	Base	83.27	83.68	84.27	83.74	78.87	82.47	82.47	81.27
	Random	74.59	74.90	74.86	74.78	74.70	76.83	76.80	76.11
	Sample	75.40	75.31	76.57	75.76	73.50	77.63	77.73	76.27
	VisPruner	77.54	76.09	75.17	**76.27**	74.13	76.60	78.50	**76.41**
	Ours	77.61	74.95	74.85	75.80	73.07	76.67	78.50	76.08

Table 3: Main results on MT-VQA-v2 and MT-GQA benchmarks across different model architectures and scales. All pruning methods operate at a 10% visual token retention ratio. Best training-free token reduction results are **bolded**.

cross-modal attention extraction exactly once during the initial turn. This factorized execution effectively amortizes the initial perception cost, bringing the incremental computational complexity of subsequent turns down to a constant $\mathcal{O}(1)$ with respect to the dialogue session length.

Static KV Cache Reuse. Because our compressed token set $\mathbf{V}'$ remains invariant across subsequent turns, its corresponding Key-Value (KV) representations can be computed once and permanently cached. This avoids the redundant KV cache reconstruction and memory fragmentation issues typically caused by dynamically shifting token sequences (Zhang et al., 2023), offering a highly hardware-friendly footprint for multi-turn acceleration.

5 Experiments

5.1 Experimental Setup

We evaluate our framework on two multi-turn benchmarks constructed from the validation splits of VQA v2 (Antol et al., 2015) and GQA (Hudson and Manning, 2019). In both cases, questions that share the same image are grouped into ordered 3-turn sessions; within a session, each turn answers a question about the shared image, while the questions and model-generated answers of preceding turns accumulate as text-only dialogue history. This yields 2,260 sessions for MT-VQAv2 and 3,000 for MT-GQA. We report turn-wise accuracy (Acc 1–Acc 3) together with the session average (Avg). Validation is conducted across four architectural configurations: LLaVA-1.5-7B/13B and LLaVA-Next-7B/13B (Liu et al., 2023, 2024a,b). Performance is compared against Base (full tokens), Random dropping, Sample (uniform down-

Method	LLaVA-1.5-7B			LLaVA-NeXT-7B		
	TFLOPs	PT (ms)	E2ET (ms)	TFLOPs	PT (ms)	E2ET (ms)
Base	25.16	68.6	296.7	111.07	156.6	1090.9
Random	4.39	71.7	193.9	13.78	156.8	375.9
Sample	4.43	71.6	194.8	13.78	203.7	390.9
VisPruner	4.77	123.0	246.2	15.69	401.5	588.9
Ours	12.96	121.7	247.4	51.75	467.1	657.9

Table 4: Efficiency comparison on MT-VQAv2 (10% retention, 3-turn sessions) on a V100. FLOPs are computed analytically per model configuration and are cumulative over a 3-turn session, including the one-time attention probe for TAP-LLaVA and per-turn recomputation for query-dependent methods. Preprocess Time (PT) and End-to-End Time (E2ET) are per-turn values averaged over full benchmark runs.

Rounds	Base (s)	Ours (s)	Per-turn (s)	Speedup
1	0.263	0.363	0.363	0.72×
3	0.784	0.669	0.223	1.17×
5	1.373	1.014	0.203	1.35×
10	2.855	1.873	0.187	1.53×
20	6.183	3.927	0.196	1.58×

Table 5: Cumulative latency and amortized per-turn cost of TAP-LLaVA versus the full-token baseline on LLaVA-1.5-7B (MT-VQAv2, 10% retention), measured over sessions of increasing dialogue length.

sampling), and VisPruner (Zhang et al., 2025a) (visual self-attention). All pruning methods enforce a strict 90% visual token reduction ratio. For TAP-LLaVA, task-agnostic attention maps are isolated from layer $l = 20$ using fixed probes.

5.2 Main Performance and Efficiency Analysis

As shown in Table 3, TAP-LLaVA consistently outperforms heuristic baselines at the 7B scale. On LLaVA-1.5-7B, the clear gains in session-average accuracy demonstrate that the task-agnostic attention prior carries task-relevant signal. The advantage holds on LLaVA-Next-7B despite its multi-scale token space. Conversely, the picture reverses at 13B: the gap narrows, and uniform downsampling (Sample) matches or outperforms our method, most clearly on LLaVA-Next-13B. This convergence marks the empirical boundary of cross-modal attention in high-capacity regimes. Uniform downsampling is thus competitive for ultra-large models, while TAP-LLaVA remains a highly effective, training-free structural prior for moderate-scale models (e.g., 7B), which dominate latency-sensitive, resource-constrained edge deployment.

On **efficiency trade-offs**, Table 4 reports computational cost, memory footprint, and per-session latency. TAP-LLaVA's cumulative 3-session FLOPs (12.96T on 7B) exceed Random (∼4.4T) because the one-time attention probe runs a full prefill pass over the original visual sequence; this cost is $\mathcal{O}(1)$ in dialogue length and does not accumulate. Peak memory is comparable across methods (13.8–14.9GB on 7B); TAP-LLaVA's modest overhead comes from the transient probe.

Table 5 tracks cumulative latency on LLaVA-1.5-7B as dialogue length grows. The one-time probe makes TAP-LLaVA slower than the baseline for single-turn interactions (0.363s vs. 0.263s, 0.72×), but the amortized per-turn cost falls from 0.363s to 0.187–0.196s as the session extends, yielding a 1.17× speedup at 3 rounds and 1.58× at 20 rounds. Real-world multi-turn dialogues (customer service, interactive assistants, robotic Q&A) commonly involve 3 or more turns, so this crossover covers the dominant deployment scenario; for 1–2 turn interactions, methods with lower one-time overhead such as VisPruner may be preferable. Beyond latency, a static token set enables full KV cache reuse across turns, avoiding the cache reconstruction and memory fragmentation of methods whose retained set changes every round.

5.3 Deconstructing the 13B Scale Bottleneck

To reconcile the empirical paradox where LLaVA-1.5-13B manifests higher static IoU consistency (Table 2) yet delivers degraded downstream task accuracy (Table 3), we track the layer-wise execution performance as visualized in Figure 3.

Specifically, LLaVA-1.5-7B exhibits a distinctive, well-defined bell-shaped dependency on layer selection, with performance peaking sharply at Layer 20. This indicates that the cross-modal semantic prior in early single-scale models is highly concentrated within specific intermediate layers.

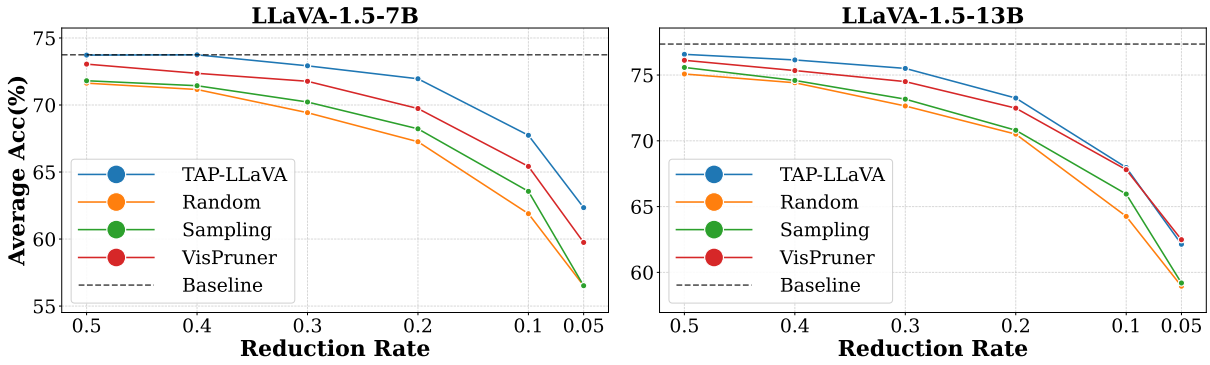


Figure 2: Performance comparison of various compression strategies on the VQAv2 dataset, reporting the average accuracy of LLaVA-1.5-7B and LLaVA-1.5-13B across different visual token retention ratios.

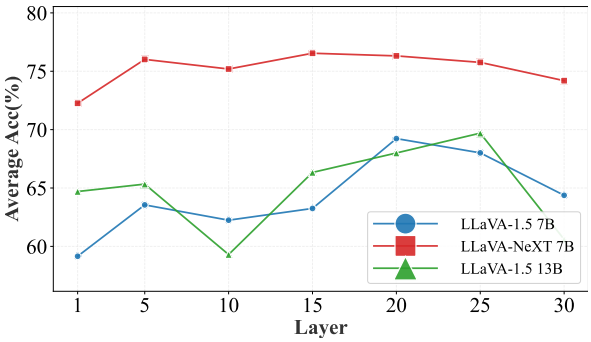


Figure 3: Results of TAP-LLaVA on the MT-VQAv2 benchmark, illustrating the average accuracy of LLaVA-1.5-7B and LLaVA-NeXT-7B with attention extracted from different transformer layers.

Conversely, for the advanced LLaVA-Next-7B architecture, the accuracy curve flattens out dramatically into a prolonged plateau, while its optimal extraction layer shifts upward to Layer 15.

This transition from a localized sharp peak to a flattened, shifted plateau provides preliminary empirical evidence for our Attention Fidelity Decay hypothesis. As network parameters expand and visual structures incorporate multi-scale representations, the visual-textual alignment signal undergoes severe semantic dilution and layer-wise drift across deep residual streams. Consequently, the exceptionally high static IoU observed in the 13B configuration represents structural rigidity—a highly constrained, layout-fixed attention pattern insensitive to task variations—rather than high-fidelity grounding. This structural ossification may help explain the observed scaling bottleneck encountered by surface-level attention pruning in high-capacity regimes, which is ultimately surpassed by uniform spatial heuristics.

Practically, these trends yield a lightweight layer-selection guideline for applying TAP-LLaVA to a new backbone: sweep candidate extraction layers (e.g., 10–25) on roughly 100 calibration sessions and adopt the peak layer. The sweep is a one-off calibration cost that is amortized over all subsequent dialogues, and in our experiments this protocol recovers the optima reported above (Layer 20 for LLaVA-1.5-7B, Layer 25 for LLaVA-1.5-13B, and Layer 15 for LLaVA-NeXT-7B).

6 Conclusion

In this work, we revisited the debate on cross-modal attention as an indicator for visual token pruning in LVLMs. We identified task-bound coupling as a confounder that injects text-driven volatility into attention fields, and showed, via causal token-deletion experiments with fixed task-agnostic prompts, that attention-ranked tokens carry task relevance while selection remains stable across diverse downstream tasks. Building on this, we introduced TAP-LLaVA, a minimal training-free framework that probes once, freezes the token selection, and reuses the static KV cache across multi-turn sessions.

Our layer-wise diagnostics also revealed a scaling limitation we term Attention Fidelity Decay: as model capacity grows, the downstream effectiveness of attention-based selection declines. Scaling token reduction to larger models may therefore require signals beyond surface attention maps, such as hidden features or residual activations.

7 Limitations

While TAP-LLaVA indicates strong efficiency–accuracy trade-offs in training-free settings, several limitations clarify its design scope and suggest directions for future work.

Heuristic Nature of Attention-Based Scoring. TAP-LLaVA builds upon cross-modal attention as a proxy for visual token importance, which remains a heuristic signal rather than a theoretically grounded estimator. One possible explanation for its degraded efficacy at larger scales—which our experiments do not establish—is that a growing share of information propagates through residual and feed-forward pathways, weakening the interpretability of attention weights; verifying or refuting this hypothesis remains open. Consequently, our method should be viewed as a lightweight and practical approximation rather than a fundamentally new token selection paradigm. Developing more principled importance estimators, potentially leveraging gradient-based or representation-level signals, is an important direction for future work.

Scope in Multi-Turn Settings. Our framework is specifically designed for multi-turn multimodal dialogue, where token relevance exhibits cross-turn consistency and can be amortized over interactions. As a result, TAP-LLaVA introduces a one-time initialization overhead to estimate token importance, which may not be optimal for single-turn or latency-critical scenarios. This reflects a deliberate design trade-off favoring long-horizon efficiency over immediate responsiveness.

Static Token Selection Strategy. To maximize computational savings, TAP-LLaVA adopts a static token selection scheme, where visual tokens are fixed after initialization. While effective in practice, this design limits the model's ability to adapt to newly emerging information needs in later dialogue turns. In scenarios requiring fine-grained or dynamically shifting attention, this may lead to suboptimal coverage. Incorporating lightweight dynamic re-selection mechanisms remains an important avenue for improving flexibility without sacrificing efficiency.

Prompt-Dependent Prior Estimation. The task-agnostic prompt used to derive the initial attention prior is manually designed and may implicitly encode biases inherited from pretraining data. Although we observe empirical robustness across prompt variations, the resulting token distribution is not entirely independent of prompt formulation. Future work could explore prompt-free or model-intrinsic approaches for estimating token importance, such as leveraging internal representations or self-consistency signals.

Layer Sensitivity of Attention Extraction. The reliability of task-agnostic attention as a pruning signal varies considerably across extraction layers, with accuracy peaking at Layer 20 for LLaVA-1.5-7B, Layer 25 for LLaVA-1.5-13B, and Layer 15 for LLaVA-NeXT-7B, and degrading sharply outside these ranges (Figure 3). The optimal layer is not known a priori for a new model and must be located through a small-scale sweep, which introduces a modest per-model calibration cost.

Scale–Architecture Confounds and Backbone Coverage. Our scale analysis is observational rather than controlled. The LLaVA-1.5 pair isolates parameter count, but the LLaVA-NeXT pair additionally changes the base LLM (Mistral for 7B, Vicuna for 13B) and the vision encoder, so the degradation observed on 13B models may partially reflect architecture and pretraining differences rather than scale alone. Moreover, all experiments are conducted on the LLaVA model family; validating on other backbones such as Qwen2.5-VL or InternVL is left to future work.

Potential Risks and Societal Impact. As with other token pruning approaches, reducing visual inputs may lead to the omission of subtle but important details, particularly in safety-critical or fine-grained scenarios. Moreover, attention-based selection may favor visually salient regions while overlooking less prominent yet semantically important content, potentially amplifying existing biases. Therefore, TAP-LLaVA should be applied with caution in high-stakes applications requiring comprehensive and faithful visual understanding.

Acknowledgements

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China(No.JYB2025XDXM118).

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4190–4197.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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*.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. 2024. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15710–15719.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. 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.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. 2023. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16091–16101.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. 2024. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European conference on computer vision*, pages 214–230. Springer.

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.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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

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

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2025. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19618–19627.

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

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22857–22867.

Kele Shao, TAO Keda, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research*.

Yi Wang, Haofei Zhang, Qihan Huang, Anda Cao, Gongfan Fang, Wei Wang, Xuan Jin, Jie Song, Mingli Song, and Xinchao Wang. 2026. Rethinking token reduction for large vision-language models. *arXiv preprint arXiv:2603.21701*.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15537–15549.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pages 21875–21895.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2025. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1877–1893.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 19792–19802.

X Ye, Y Gan, Y Ge, XP Zhang, and Y Tang. 2024. Atp-llava: adaptive token pruning for large vision language models (2024). *URL https://arxiv.org/abs/2412*, 447.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2025a. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20857–20867.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024a. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pages arXiv–2412.

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

Zeliang Zhang, Phu Pham, Wentian Zhao, Kun Wan, Yu-Jhe Li, Jianing Zhou, Daniel Miranda, Ajinkya Kale, and Chenliang Xu. 2024b. Treat visual tokens as text? but your mllm only needs fewer efforts to see. *arXiv preprint arXiv:2410.06169*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. 2023. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30678–30701, Vienna, Austria. Association for Computational Linguistics.

A Computational Resources and Evaluation Protocol

All experiments were conducted on a single GPU server equipped with eight Tesla V100-SXM2-32GB GPUs. Models were loaded in their original precision from their official checkpoints. The software environment was a Python 3 conda environment with PyTorch 2.4.0 (CUDA 12.1), transformers 4.51.0, numpy 1.26.4, and peft 0.15.2.

Because all compared methods are training-free and evaluation is deterministic (greedy decoding, no sampling or temperature), each reported accuracy is the outcome of a single run on the official checkpoint. We therefore report single-run results without variance bars; no metric is an average over multiple stochastic executions, and no hyperparameter was selected on the test set. Total GPU-hours are not reported separately: the experiments were run incrementally over the course of the project on a shared machine, and the inference footprint of the training-free sweep is not tracked as an aggregate.

B Prompt Templates

We provide the full list of task-agnostic prompts used in our experiments. These prompts are designed to avoid instance-specific objectives and instead serve as general attention probes.

Holistic Task-Agnostic Prompts.

- Identify all visually and semantically important objects in the image.
- Describe this image in detail, covering all visible elements.
- What are the main elements and their spatial relationships in this picture?
- Analyze the content of this image comprehensively, including objects, colors, and text.

Diverse Task-Agnostic Prompts.

- List all distinct objects visible in this image.
- Count the total number of distinct items present in this scene.
- Describe the spatial arrangement of elements from left to right.
- Classify the overall scene category and setting of this image.