COVERAGE-PRESERVING RECASTING AND HIERARCHICAL VISUAL TOKEN PRUNING FOR EFFICIENT MULTIMODEL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models increasingly rely on long visual token sequences, making token pruning essential for reducing inference overhead. However, existing methods often suffer from accuracy degradation under stringent token budgets since they insufficiently account for the spatial distribution of visual evidence and the layer-dependent characteristics of multimodal representations. Therefore, we investigate the mechanisms underlying effective token reduction and find that visual token importance in the encoder should be assessed through complementary visual cues together with crop-wise allocation to avoid overlooking informative regions, while instruction-conditioned relevance becomes reliable as visual semantics mature across decoder layers. Motivated by these findings, we propose Layered-Crop Holistic Token Pruning called LCHTP, a training-free two-stage framework that integrates spatially aware selection with layer-sensitive compression. We further develop a theoretical formulation that links pruning-induced representation distortion to prediction stability and inference complexity, which provides a principled account of the accuracy–efficiency trade-off. Extensive evaluations across diverse MLLMs and benchmarks demonstrate that LCHTP consistently preserves strong predictive performance while delivering inference acceleration.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have rapidly advanced visual perception and instruction following, substantially extending the scope of Large Language Models (LLMs) beyond text-only understanding (Liu et al., 2023). Recent advances further enhance multimodal competence by coupling stronger visual encoders with increasingly powerful language backbones (Bai et al., 2023; Gemini Team et al., 2023). However, this progress increasingly relies on higher-resolution visual representations, leading to markedly longer visual token sequences that must be processed by the LLM (Li et al., 2025; Bai et al., 2025). Processing such long sequences imposes considerable computational overhead, despite the substantial redundancy among visual tokens. Thus, visual token reduction has become a central direction for efficient MLLMs, with the objective of eliminating dispensable visual tokens while preserving the reliable reasoning evidence.

Existing token pruning methods can be broadly categorized according to the signals used for token evaluation and the mechanisms adopted for information preservation. Vision-centric approaches estimate token importance from visual representations or encoder-side statistics. FasterVLM ranks visual tokens according to encoder attention (Zhang et al., 2024), while HiRED introduces partition-aware budget allocation before token selection (Arif et al., 2025). Instruction-aware approaches exploit cross-modal relevance between visual content and language input. IVTP performs instruction-conditioned token selection (Huang et al., 2024a), while MustDrop incorporates text-guided pruning into a multi-stage reduction pipeline (Liu et al., 2024c). A complementary line of research focuses on representation consolidation rather than direct token removal. LLaVA-PruMerge merges redundant visual features around informative tokens (Shang et al., 2025), FiCoCo-V transfers information from correlated discarded tokens into retained representations (Han et al., 2026), and VisionZip combines saliency estimation with similarity-based compression (Yang et al., 2025). Collectively, these methods show that substantial acceleration is attainable through carefully designed token selection and representation compression.

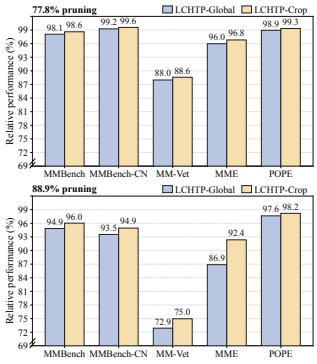

Figure 1: Comparison of crop-wise and global-ranking token selection strategy.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite these advances, existing methods still exhibit a pronounced accuracy–efficiency trade-off, particularly under stringent token budgets. First, token selection is frequently governed by highly concentrated saliency, which can bias the retained budget toward a limited subset of prominent regions. Yet visually inconspicuous areas may contain complementary objects, contextual evidence, or relational cues that are indispensable for holistic scene understanding (Zou et al., 2025; Arif et al., 2025). As shown in Fig. 1, crop-wise selection consistently outperforms global ranking, with larger gains at an 88.9% pruning ratio, which highlights the increasing importance of region-aware allocation under heavy pruning. Second, many pruning strategies operate at a fixed stage or derive token importance from only a narrow portion of the network. This treatment does not fully accommodate the hierarchical processing structure of MLLMs, in which representations at varying depths serve non-equivalent computational roles and retain distinct semantic characteristics. Coordinating pruning locations and token budgets across network depth consequently remains an underexplored dimension of efficient multimodal inference (Xing et al., 2025; Zhang et al., 2026). Finally, the majority of prior studies justify token reduction primarily through empirical performance. A unified theoretical account linking pruning-induced representation distortion, downstream prediction stability, and computational savings remains limited (Han et al., 2026). This gap leaves the accuracy preservation of aggressive pruning insufficiently characterized from a principled perspective.

To address these limitations, we first conduct extensive empirical analyses to examine the mechanisms underlying visual redundancy, spatial evidence preservation, and token importance across image regions and network depth. These investigations reveal that effective reduction requires joint consideration of spatially distributed evidence and the heterogeneous semantic roles carried by representations throughout the multimodal architecture. Motivated by these findings, we propose a training-free two-stage token pruning method LCHTP. In the first stage, LCHTP performs crop-wise pruning before LLM by combining complementary visual importance signals and distributing the retention budget across spatial regions, thereby preventing excessive concentration on a small number of salient areas. Discarded visual evidence is further recast into the retained representations to alleviate information loss without increasing the resulting sequence length. In the second stage, LCHTP applies additional instruction-guided pruning within the language decoder, allowing more intensive compression after sufficient multimodal processing. Beyond the empirical design, we establish a theoretical framework that characterizes representation distortion, perturbation propagation, prediction preservation, and inference complexity within a unified analysis. This formulation provides a principled account of the accuracy–efficiency trade-off achieved by LCHTP. The main contributions of our work are summarized as below:

- We uncover key mechanisms governing visual token pruning across the visual encoder and language decoder. In the encoder, effective token importance estimation requires complementary visual cues together with crop-wise allocation to preserve spatially distributed evidence. In the decoder, instruction relevance becomes more reliable after sufficient cross-modal interaction.
- We propose LCHTP, a two-stage pruning paradigm that jointly considers spatial coverage and hierarchical semantics. LCHTP performs crop-wise token selection with information recasting before the LLM, followed by instruction-guided pruning at an intermediate decoder stage.
- We establish theoretical guarantees for the accuracy–efficiency preservation of LCHTP. The formulation links pruning-induced representation distortion to prediction stability through the logit margin and demonstrates that reduced visual sequences can lower inference computation while preserving model predictions.

2 Understanding the Mechanisms of Token Pruning

2.1 Preliminary

Given an input image, the vision encoder partitions it into N patches. Let $Z^{(l)} = \{z_i^{(l)}\}_{i=1}^{N}$, where $z_i^{(l)}$ denote the patch-token representations at the l-th encoder layer. For multi-head self-attention, $A^{(l,h)}$ denotes the attention matrix of head h, and $\bar{A}^{(l)}$ is the corresponding head-averaged attention map. Then, the visual representations are mapped into the language-model hidden space through the multimodal projector $P(\cdot)$, yielding $v_i = P(z_i)$. Besides, we construct the ground-truth (GT) patch set G by mapping COCO object annotations and EasyOCR-localized TextVQA regions onto the 24×24 patch grid following the model visual preprocessing. Details are provided in Appendix D.1.

2.2 Empirical Findings and Observations

We first examine how visual interactions evolve across encoder depth. For each image, we randomly select one patch token as a fixed anchor and trace the corresponding row of the self-attention map $\bar{A}^{(l)}$.

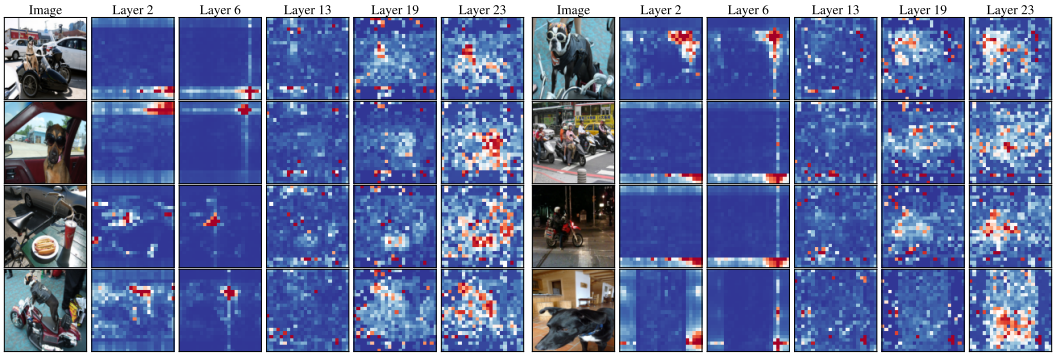

Figure 2: Self-attention evolves from local to global across visual depth. Each row tracks a randomly selected anchor patch across Layers 2, 6, 13, 19, and 23 using head-averaged self-attention.

As shown in Fig. 3, attention is strongly localized at layers 2 and 6, but becomes progressively more spatially distributed at deeper layers. We characterize this evolution using two quantitative indicators. The fraction of attention assigned to the anchor's 5×5 spatial neighborhood decreases from 0.21 at layer 2 to 0.05 at layer 23. Besides, we measure the effective number of attended patches using the participation ratio $N_{\text{eff}}^{(l)} = (\sum_i a_i^{(l)})^2 / \sum_i (a_i^{(l)})^2$, where $a_i^{(l)}$ denotes the pre-softmax attention score associated with patch i. Averaged over all examples, N_{eff} increases from 146 at layer 2 to 398 at layer 23, revealing a clear local-to-global attention evolution. This depth-dependent shift highlights the complementary roles of shallow and deep layers and motivates multi-layer token evaluation. The statistical regularities of this depth-wise evolution is further exploited in Appendix D.1.

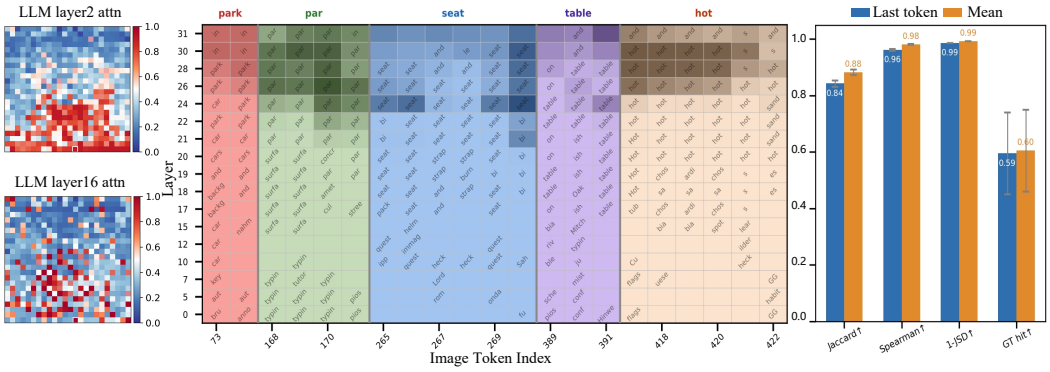

Figure 3: Left: Visual attention at shallow and intermediate decoder layers. Middle: Logit-lens visualization showing the progressive emergence of interpretable visual concepts across layers. Right: Cross-paraphrase consistency comparison of last-token and instruction-averaged queries.

Moreover, we investigate the decoder depth at which visual representations become reliable for subsequent text-guided pruning, together with a robust formulation of the instruction query. For each image, we construct five semantically equivalent instructions and compare two query representations, one derived from the final instruction token and the other averaged over all instruction positions. The resulting instruction-to-visual attention provides a relevance score for each visual token. Across paraphrase pairs, consistency is evaluated using four metrics. Top-96 Jaccard similarity measures the overlap between the selected visual-token subsets, Spearman rank correlation measures agreement over the complete visual-token ordering, 1-JSD, where JSD denotes the Jensen–Shannon divergence, measures similarity between the full attention distributions, and GT-hit measures whether the selected tokens preserve annotated critical evidence. As shown in the left panel of Fig. 3, last-token attention is strongly affected by positional bias in shallow decoder layers, whereas intermediate layers increasingly focus on semantically relevant image regions. The middle logit-lens visualization further reveals a transition from unstable image-aligned predictions to coherent visual concepts such as park, seat and table, indicating that task-aware pruning becomes reliable only after adequate cross-modal refinement. The right panel shows that instruction-level averaging yields more stable relevance estimates across paraphrases. Error bars denote 95% confidence intervals from 2,000 bootstrap resamples. Together, these observations support delaying text-guided pruning until visual semantics become sufficiently consolidated and using instruction-level query aggregation for robust token selection.

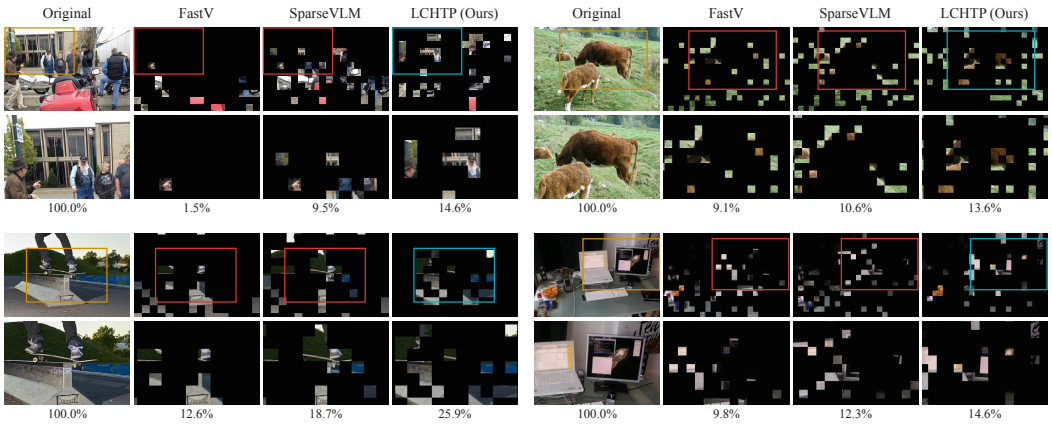

Figure 4: Qualitative comparison of visual-token retention at an 87.5% pruning rate, with black regions denoting pruned tokens and the reported percentages measuring the area-weighted fraction preserved within each key semantic region.

3 METHOD

Our analysis in Section 2 reveals two key insights: decoder layers provide complementary local and global cues, while text-guided relevance becomes reliable only after sufficient cross-modal refinement. Motivated by these findings, we develop **Layered-Crop Holistic Token Pruning (LCHTP)**, a training-free two-stage framework with layer-aware visual pruning before the LLM and instruction-guided pruning at an intermediate decoder stage.

3.1 LAYER-AWARE CROP-WISE VISUAL TOKEN PRUNING

Our spatial analysis shows that a single global ranking can over-allocate the retained budget to a few highly salient regions and consequently suppress informative evidence elsewhere. To preserve spatial coverage, we reconstruct the two-dimensional patch-token layout within each physical visual crop and perform region-aware token selection over the resulting spatial partitions. For a square patch-token grid, we organize the patch tokens into $C = R \times K$ non-overlapping regions. The quantities R and K denote the numbers of region rows and columns, respectively. The partition is then defined as

$$\mathcal{C} = \{\mathcal{C}_c\}_{c=1}^{C}, \quad \bigcup_{c=1}^{C} \mathcal{C}_c = \{1, \ldots, N\}, \quad \mathcal{C}_c \cap \mathcal{C}_{c'} = \emptyset \text{ for } c \neq c'. \tag{1}$$

Since shallow visual layers retain fine-grained local structure that may be attenuated in deeper representations, we introduce a local relevance score based on shallow patch-to-patch attention. Then, we formulate the relevance score as

$$s_i^{\text{loc}} = \begin{cases} \dfrac{1}{H} \displaystyle\sum_{h=1}^{H} \dfrac{1}{|\mathcal{C}_c| - 1} \sum_{\substack{j \in \mathcal{C}_c \\ j \neq i}} A_{j,i}^{(\ell_1, h)}, & |\mathcal{C}_c| > 1, \\ 0, & |\mathcal{C}_c| = 1. \end{cases} \tag{2}$$

Here, H denotes the number of attention heads. After excluding self-attention along the region diagonal, s_i^{loc} quantifies the average attention that token $i \in \mathcal{C}_c$ receives from the other tokens in the same region, thereby characterizing its association with surrounding visual contexts.

Deeper visual layers progressively emphasize semantically dominant content, providing a complementary signal to the fine-grained structure retained by shallow representations. Therefore, we derive a global relevance score from the attention assigned by the deep-layer `[CLS]` token to individual patch tokens. As the `[CLS]` token aggregates image-level information, it provides a natural reference for estimating global visual importance. The global relevance score is then expressed as $s_i^{\text{glob}} = \frac{1}{H} \sum_{h=1}^{H} A_{\text{CLS},i}^{(\ell_2, h)}$, where ℓ_2 denotes the selected deep layer of the vision encoder. To explicitly account for feature-space diversity, we introduce a distinctiveness score derived from the penultimate visual layer that assesses the extent to which each patch token contributes non-redundant visual information beyond the dominant feature patterns. Specifically, we characterize the feature distribution within each spatial partition by a partition-level prototype μ_c, while μ_p serves as an image-level feature centroid summarizing the overall visual content. The diversity score s_i^{div} is then

defined as follows:

$$\mu_c = \frac{1}{|\mathcal{C}_c|}\sum_{j\in\mathcal{C}_c} v_j, \quad \mu_p = \frac{1}{N}\sum_{j=1}^{N} v_j, \quad s_i^{\text{div}} = \|v_i - \mu_c\|_2 + \|v_i - \mu_p\|_2, \quad i \in \mathcal{C}_c. \tag{3}$$

The score s_i^{div} assigns higher importance to patch tokens that remain distinctive relative to both their local region and the full physical crop, thereby complementing the two attention-based relevance signals with an explicit measure of feature-space diversity.

Since the three signals encode complementary notions of token importance but operate on potentially different scales, we normalize them before fusion to ensure comparable contributions. Each score is independently min–max normalized over the patch tokens of the physical crop, and the resulting signals are integrated through a weighted linear combination:

$$\widehat{s}_i^{(q)} = \frac{s_i^{(q)} - s_{\min}^{(q)}}{s_{\max}^{(q)} - s_{\min}^{(q)} + \epsilon}, \quad q \in \{\text{loc}, \text{glob}, \text{div}\}, \quad s_i = \alpha\widehat{s}_i^{\text{loc}} + \beta\widehat{s}_i^{\text{glob}} + \gamma\widehat{s}_i^{\text{div}}. \tag{4}$$

The experiment adopts the numerical configuration $(\alpha, \beta, \gamma) = (0.4, 0.4, 0.2)$ and $\epsilon = 10^{-6}$. Given the fused relevance scores, we next allocate the first-stage token budget across spatial crops. Let K_1 denote the target number of patch tokens retained after the first pruning stage. Rather than performing a global Top-K_1 selection, we first determine a crop-specific quota according to each crop's relative capacity. For a set of candidate crops with capacities $\{n_c\}_{c=1}^{C}$, the ideal quota is defined as

$$\widetilde{k}_c = K_1\frac{n_c}{\sum_j n_j}, \qquad k_c = \lfloor \widetilde{k}_c \rfloor + \delta_c, \qquad \sum_c k_c = K_1, \qquad \delta_c \in \{0, 1\}, \tag{5}$$

where the residual budget δ_c is determined after integer rounding. Specifically, let $R = K_1 - \sum_c \left\lfloor \widetilde{k}_c \right\rfloor$ denote the number of unassigned tokens. These residual tokens are distributed by setting $\delta_c = 1$ for the R crops with the largest fractional remainders $\widetilde{k}_c - \lfloor \widetilde{k}_c \rfloor$, while setting $\delta_c = 0$ for all remaining crops. Given the resulting crop-specific quota k_c, we independently retain the k_c highest-scoring tokens within each crop, yielding the selected token set Ω_c. The selected tokens are then mapped back to their original sequence positions, yielding K_1 visual tokens after the first-stage pruning.

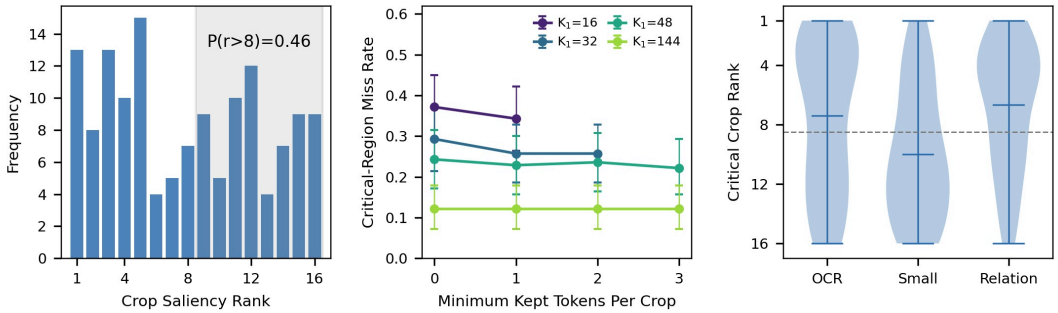

Figure 5: Left: Saliency-rank distribution of critical evidence crops. Middle: Critical-region miss rate under different minimum per-crop quotas and token budgets. Right: Critical-crop rank distributions for optical character recognition (OCR), small-object, and relational evidence.

We futher examine whether crop-level saliency reliably reflects the importance of visual evidence. Under the 4×4 spatial partition, each crop C_c contains 36 visual tokens, with saliency defined as the mean score of its constituent tokens $a_c = \frac{1}{|C_c|}\sum_{i\in C_c} s_i$. We rank the 16 crops by a_c and designate the crop with the largest overlap with the ground-truth patch set G as the critical evidence crop. As shown on the left of Fig. 5, critical evidence is widely distributed across the saliency ranking, with $P(r > 8) = 0.464$, indicating that 46.4% of critical regions occur outside the top eight crops. The right panel further shows that this phenomenon persists across evidence types and is especially pronounced for small-object evidence. We then assess whether lightweight spatial protection alleviates such evidence loss by imposing a minimum per-crop quota q under different total token budgets K_1. A sample is counted as missed when no retained token overlaps with the complete GT patch set G. As shown in the middle of Fig. 5, even a small quota markedly reduces the critical-region miss rate under aggressive token budgets, while the benefit becomes marginal at $K_1 = 144$, where the larger budget already provides broader evidence coverage. These results reveal that low crop saliency does not imply evidence dispensability. Enforcing lightweight crop-wise

protection can therefore prevent potentially critical regions from being prematurely suppressed during token selection, providing empirical support for the region-aware quota allocation in Eqs. (4) and (5).

3.2 Text-Guided Visual Recasting

Token pruning inevitably discards part of the visual representations, which may lead to the degradation of task-relevant information. To alleviate this loss without increasing the first-stage token count K_1, the discarded information is adaptively integrated into the retained patch tokens. Within each physical crop, let Ω and $\bar{\Omega}$ denote the retained and discarded token sets, respectively. Besides, the embedded text tokens are denoted as $T = \{t_j\}_{j=1}^{M}$. Each discarded token $v_d \in \bar{\Omega}$ is matched to its most similar retained token m_d and assigned a fusion weight w_d based on both visual similarity and text relevance:

$$m(d) = \arg\max_{v_s \in \Omega} \cos(v_d, v_s), \quad \tau_d = \begin{cases} \max\limits_{1 \le j \le M} \cos(v_d, t_j), & M > 0, \\ 0, & M = 0, \end{cases} \tag{6}$$
$$w_d = \left[\cos(v_d, v_{m(d)}) + \lambda \tau_d\right]_+ .$$

Here, the resulting w_d determines the contribution of the discarded token. The operator $[x]_+ = \max(x, 0)$ enforces non-negative fusion weights with the configuration $\lambda = 0.5$. For each retained token r, let $\mathcal{D}_s = \{v_d \in \bar{\Omega} \mid m(d) = v_s\}$ denote the discarded tokens assigned to it. The representations of these discarded tokens are then incorporated through normalized weighted aggregation:

$$\widetilde{v}_s = \frac{b v_s + \sum_{v_d \in \mathcal{D}_s} w_d v_d}{b + \sum_{v_d \in \mathcal{D}_s} w_d}. \tag{7}$$

The anchor coefficient b preserves the original retained representation and is fixed to 1. This fusion enriches the retained representations with complementary visual context, mitigating information degradation induced by token pruning while maintaining the efficiency of token compression.

3.3 Intermediate Decoder Pruning and Layer-Average Budget

Since early decoder layers are more susceptible to positional bias while visual semantics become progressively more discriminative toward the middle layers, and the instruction-mean query provides more stable visual grounding than the last-token query, the second-stage visual-token reduction is deferred to the intermediate decoding stage and guided by an instruction-level query.

After the first d decoder blocks, visual tokens have undergone sufficient cross-modal interaction, making them more amenable to task-aware pruning. Let $\mathcal{V}_d$ denote the positions of remaining visual-token and $\mathcal{T}_d$ the non-visual positions used to construct the instruction query. Then, we form the query for each attention head by averaging the text-position queries, while deriving the visual keys from the corresponding hidden states after the standard attention transformation. The resulting instruction-to-visual attention is averaged across heads to obtain the visual relevance score:

$$a_i = \frac{1}{H} \sum_{h=1}^{H} \frac{\exp\left(q^h (k_i^h)^\top / \sqrt{d_h}\right)}{\sum_{j \in \mathcal{V}_d} \exp\left(q^h (k_j^h)^\top / \sqrt{d_h}\right)} \tag{8}$$

Here, q^h denotes the mean query representation over all positions in $\mathcal{T}_d$ for head h, and k_i^h denotes the corresponding key representation of visual token $i \in \mathcal{V}_d$. Based on the instruction-guided relevance scores derived above, the second stage globally retains the K_2 most relevant visual tokens $\Phi = \mathrm{TopK}_{i \in \mathcal{V}}(a_i, K_2)$. In contrast to the crop-wise selection in the first stage, this global ranking allows the reduced visual sequence to adapt directly to the semantic content of the instruction. The selected tokens are then propagated through the remaining decoder layers.

Since intermediate pruning leads to varying visual sequence lengths across decoder layers, we characterize the token budget by the average visual-token load throughout the decoder. For a decoder with L layers, the two-stage schedule processes K_1 visual tokens through the first d layers and K_2 tokens through the remaining $L - d$ layers. The resulting layer-average budget is given by

$$B = \frac{dK_1 + (L - d)K_2}{L}. \tag{9}$$

This formulation enables a consistent comparison of visual-token consumption across different pruning depths and token allocations. For the faithful-tuned profile, both K_1 and d are configurable. Given a target budget B, the corresponding second-stage token count is derived directly from Eq. (9):

$$K_2^* = \frac{LB - dK_1}{L - d}. \tag{10}$$

Regarding the experimental setting, we empirically select the intermediate pruning depth d according to the model architecture and set $K_2 = \lceil 2K_1/3 \rceil$. Given a target budget B, we determine K_1 by rounding to the feasible integer value that yields the realized layer-average budget in Eq. (9) closest to B. As shown in Fig. 4, LCHTP preserves a larger fraction of visual tokens within key semantic regions than FastV and SparseVLM. The qualitative results indicate that LCHTP retains image-relevant evidence and better preserves the required core semantics.

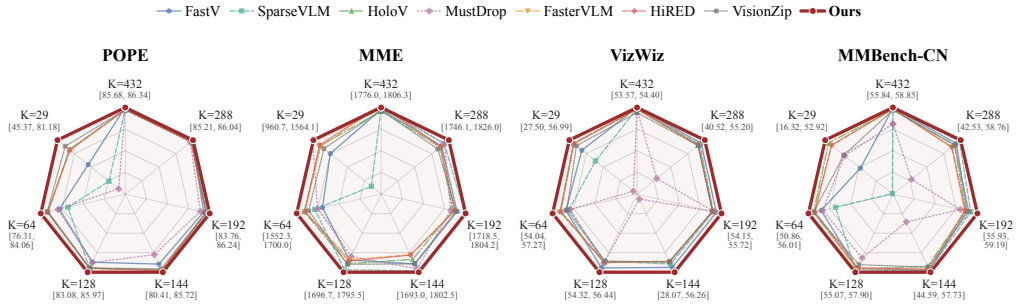

Figure 6: Performance across visual-token budgets on LLaVA-1.5-7B.

4 Theoretical Analysis

Theorem 1 (Accuracy–efficiency preservation of LCHTP)**.** *Consider an unpruned multimodal decoder with N visual tokens and L decoder blocks. Let z and $\widehat{z}$ be the output logits of the unpruned model and LCHTP, respectively. The terms $\mathcal{E}_1(K_1)$ and $\mathcal{E}_2(K_2, d)$ denote the representation distortions introduced by first-stage crop-wise selection with visual recasting and delayed instruction-guided pruning. We use $\Gamma_{a:b}$ to denote the accumulated Lipschitz propagation from decoder block a to b, and Λ_0 for the Lipschitz bound of the output projection. Under the stability, score-calibration, recasting-smoothness, semantic-stabilization, and bounded-value conditions specified in the Appendix C, the final logit perturbation satisfies*

$$\|z - \widehat{z}\|_\infty \le \Delta(K_1, K_2, d) = \Lambda_0 \left[\Gamma_{1:L}\mathcal{E}_1(K_1) + \Gamma_{d+1:L}\mathcal{E}_2(K_2, d)\right], \tag{11}$$

where $\Delta(K_1, K_2, d)$ is the upper bound on the end-to-end logit perturbation. Let $\mathfrak{m}(x, T)$ denote the logit margin of the unpruned model for image x and instruction T, defined as the gap between the largest logit and its strongest competitor. Then, The prediction ability is preserved whenever

$$\mathfrak{m}(x, T) > 2\Delta(K_1, K_2, d). \tag{12}$$

Let Ξ_0 and Ξ_2 denote the decoder inference costs of the unpruned model and LCHTP, respectively, where Ξ_2 includes the pruning overhead. When $K_2 < K_1 < N$ and the pruning overhead remains smaller than the decoder cost saved by shortening the visual sequence, the resulting inference costs satisfy $\Xi_2 < \Xi_0$. The idealized decoder speedup factor $\mathcal{S} = \Xi_0/\Xi_2$ therefore satisfies $\mathcal{S} > 1$.

Remark 1 (Implication for the two-stage design)**.** *The first-stage layer-aware crop-wise selection and visual recasting reduce $\mathcal{E}_1(K_1)$ by preserving distributed evidence and recovering information from discarded tokens. Since this distortion is propagated through all decoder blocks via $\Gamma_{1:L}$ in Eq. (11), controlling $\mathcal{E}_1(K_1)$ is particularly important. The delayed second stage reduces $\mathcal{E}_2(K_2, d)$ by using an instruction-level query after cross-modal relevance becomes more stable, while its distortion is propagated only through the shorter factor $\Gamma_{d+1:L}$. The token allocation further induces an explicit accuracy–cost trade-off: retaining a larger K_1 reduces early-stage distortion, whereas choosing $K_2 < K_1$ lowers Ξ_2 by shortening the visual sequence over the remaining $L - d$ blocks. Thus, LCHTP preserves more capacity before semantic stabilization and applies stronger compression afterward, jointly controlling $\Delta(K_1, K_2, d)$ and decoder cost.*

5 Experiment

5.1 Experimental Setup

Benchmark. We evaluate LCHTP on diverse multimodal benchmarks covering general perception, compositional reasoning, text-rich understanding, scientific reasoning, and hallucination assessment. For image-based evaluation, we consider VQAv2, GQA, MMBench and MMBench-CN, MME, POPE, VizWiz, ScienceQA, TextVQA, and MM-Vet. Together, these benchmarks provide complementary views of model capability, ranging from low-level visual perception and spatial reasoning to OCR-intensive understanding and open-ended multimodal reasoning. We further examine the transferability of our pruning strategy to video inputs using MSVD-QA and MSRVTT-QA.

Comparison Methods. We compare LCHTP with a comprehensive set of representative training-free visual token reduction methods, including ToMe, FastV, HiRED, SparseVLM, PDrop, MustDrop, FiCoCo-V, FasterVLM, VisionZip, DART, and HoloV. These approaches cover a range of reduction mechanisms, including token merging, attention-based selection, cross-modal relevance estimation, progressive pruning, and representation consolidation. We follow the default parameter settings provided in the official implementations of all comparison methods. All experiments were conducted on a Linux server equipped with seven NVIDIA A800 80GB PCIe GPUs.

5.2 Experimental Result

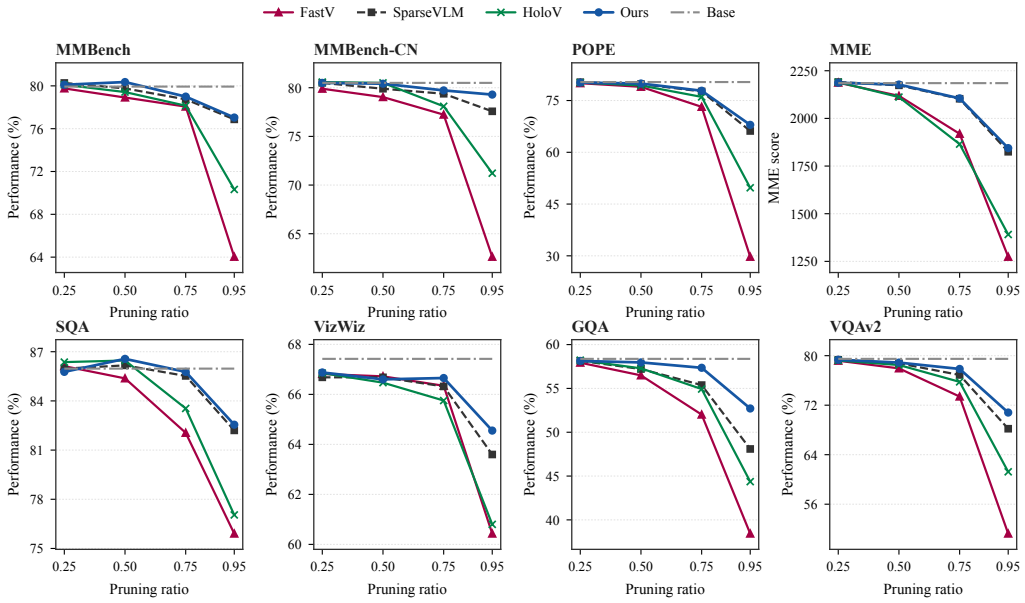

Figure 7: Performance across visual-token budgets on Qwen2.5-VL.

Table 1: Real inference efficiency comparison under different visual token budgets. Best results among token reduction methods are highlighted in blue bold.

Methods	Time	Prefill	Latency	Mem.	Acc.	F1	Time	Prefill	Latency	Mem.	Acc.	F1
Base/greedy (native 576)	21:19	71.19ms	0.1421s	13.59G	85.16%	85.94%	21:43	71.7ms	0.1447s	13.59G	85.16%	85.94%
Token Budget	*Retain 192 Tokens* (↓ 66.7%)						*Retain 64 Tokens* (↓ 88.9%)					
SparseVLM	21:35	55.38ms	0.1439s	17.46G	85.04%	85.68%	20:46	48.4ms	0.1384s	17.42G	82.60%	83.04%
FastV	20:47	42.66ms	0.1385s	13.58G	85.04%	84.65%	17:42	35.6ms	0.1180s	13.46G	81.13%	78.96%
HiRED	19:06	**40.09ms**	0.1273s	**13.41G**	84.56%	85.40%	19:31	**33.5ms**	0.1302s	**13.30G**	82.69%	82.56%
LCHTP (**Ours**)	**18:50**	51.89ms	**0.1255s**	13.46G	**85.18%**	**86.01%**	**16:21**	46.6ms	**0.1090s**	13.46G	**84.11%**	**85.00%**

Results under varying pruning ratios. Figs. 6–8 show that LCHTP consistently maintains a broader and stable performance profile as the visual-token budget becomes increasingly restrictive. On LLaVA-1.5-7B, the radar plots in Fig. 6 place LCHTP near the outer boundary across different token budgets on POPE, MME, VizWiz, and MMBench-CN, whereas several competing approaches exhibit noticeably larger contractions as fewer tokens are retained. A similar trend is observed on Qwen2.5-VL-7B in Fig. 7. FastV, SparseVLM, and HoloV remain competitive under relatively mild pruning, yet their performance degrades substantially as the pruning ratio increases, especially on POPE, MME, GQA, and VQAv2. By contrast, LCHTP exhibits a smoother degradation trajectory over the full pruning range. The fine-grained radar plots in Fig. 8 further demonstrate that the advantage is not confined to a small subset of capabilities. Across progressively reduced token budgets, LCHTP preserves a comparatively complete capability envelope over perception, OCR-related understanding and scene recognition,

Table 2: Comparison with existing visual token reduction methods on LLaVA-NeXT-7B. Best results are highlighted in blue bold.

Methods	MMB	MMB_{CN}	POPE	MME	SQA	VQA_{Text}	VizWiz	GQA	VQAv2	Average
Upper 2880 Tokens	67.4	60.5	88.7	1689	68.9	64.8	60.6	64.2	80.1	100.0%
LLaVA-NeXT-7B	*Retain 320 Tokens* (↓ 88.9%)									
FastV	55.3	47.1	63.6	1599	**68.2**	55.6	54.5	51.5	62.5	84.4%
HiRED	65.1	57.3	86.7	1627	66.8	59.8	56.9	60.1	**75.2**	95.1%
SparseVLM	61.9	54.2	84.4	1557	67.0	**60.1**	55.1	57.9	72.7	92.3%
MustDrop	63.3	57.6	84.0	1598	65.3	49.6	55.9	56.2	70.6	90.9%
FasterVLM	65.1	56.8	86.2	1651	66.7	59.1	57.5	59.6	73.9	94.8%
VisionZip	63.7	56.9	85.6	1651	66.8	58.0	57.5	59.1	73.7	94.2%
HoloV	64.5	56.9	86.1	1643	66.5	59.0	57.5	59.5	73.9	94.6%
LCHTP (**Ours**)	**65.3**	**57.7**	**87.8**	**1675**	68.1	58.4	**57.6**	**61.7**	75.1	**96.0%**

while competing methods tend to lose performance more unevenly across individual dimensions. Overall, the widening performance gap highlights the robustness of LCHTP to aggressive visual compression and validates the benefit of jointly preserving spatially distributed evidence and exploiting hierarchical semantic evolution. Moreover, Table 2 demonstrates the consistent effectiveness of LCHTP across diverse multimodal tasks, achieving the highest average relative performance of 96.0% and the best results on six of the nine evaluated tasks under an 88.9% pruning ratio.

Efficiency Analysis. Table 1 compares the real inference efficiency under different visual-token budgets. The accelerate advantage becomes more pronounced under the 64-token budget, where LCHTP further reduces the total inference time to 16:21 and achieves the lowest latency of 0.1090s, while preserving the highest accuracy and F1 among the token reduction methods. In comparison, FastV and HiRED achieve lower prefill latency but incur larger performance degradation, whereas SparseVLM introduces substantial memory overhead. Overall, LCHTP achieves a more favorable balance between inference efficiency and predictive performance.

Generalization to Video Understanding. Table 3 reports the results with 50% of the visual tokens retained. On MSVD-QA, LCHTP reaches 60.30 accuracy and a score of 3.328, outperforming the strongest competing method by 0.50 accuracy points. The advantage is more pronounced on MSRVTT-QA, where LCHTP improves accuracy from 55.90 to 57.50 over HoloV while also obtaining the highest score of 2.979. Averaged over the two datasets, LCHTP achieves 58.90 accuracy and a score of 3.154, surpassing all comparison methods as well as the unpruned base model. These results demonstrate that LCHTP transfers effectively to video inputs.

Table 3: Video QA results on Video-LLaVA-7B.

Methods	MSVD-QA		MSRVTT-QA		Average	
	Acc.	Score	Acc.	Score	Acc.	Score
Upper 2048 Tokens	59.10	3.273	54.40	2.828	56.75	3.051
Video-LLaVA-7B	*Retain 1024 Tokens (↓ 50.0%)*					
HoloV	59.80	3.317	55.90	2.918	57.85	3.118
FasterVLM	59.00	3.286	55.40	2.892	57.20	3.089
FastV	58.10	3.207	52.10	2.757	55.10	2.982
DART	57.40	3.201	53.30	2.798	55.35	3.000
LCHTP **(Ours)**	**60.30**	**3.328**	**57.50**	**2.979**	**58.90**	**3.154**

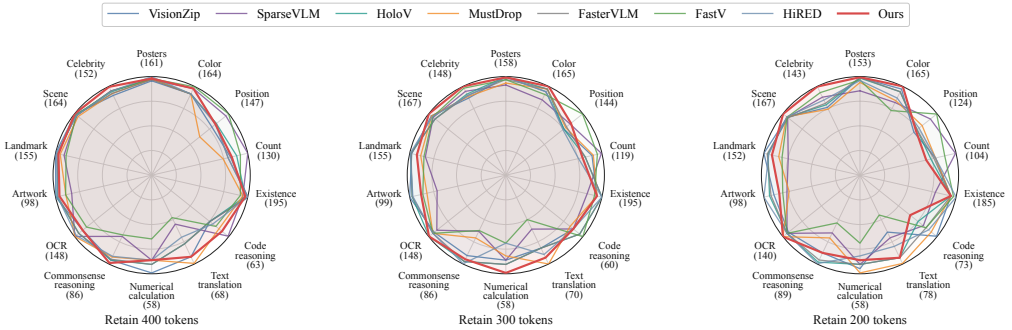

Figure 8: Performance of LLaVA-NeXT-7B across 14 MME tasks under different token budgets.

Ablation on Crop-wise pruning strategy. Building on the advantage of crop-wise selection established in Fig. 1, we further examine the effect of spatial partition granularity. As shown in Table 4, a small number of crops is already sufficient under moderate pruning, with four crops performing best at a 66.7% pruning ratio. As the pruning ratio increases, finer partitions become increasingly beneficial, with 12 and 16 crops achieving the strongest results at 77.8% and 88.9%, respectively. The trend suggests that finer spatial partitioning is particularly valuable under tighter token budgets.

Table 4: Ablation of different crop numbers.

Methods	#=4	#=8	#=12	#=16
Upper Bound	100%	100%	100%	100%
LLaVA-1.5-7B	*Token Pruning Rate = 66.7%*			
LCHTP (Ours)	**98.966%**	98.541%	98.393%	98.351%
LLaVA-1.5-7B	*Token Pruning Rate = 77.8%*			
LCHTP (Ours)	96.791%	96.777%	**96.852%**	95.845%
LLaVA-1.5-7B	*Token Pruning Rate = 88.9%*			
LCHTP (Ours)	90.021%	90.076%	90.121%	**90.306%**

6 Conclusion

We present LCHTP, a training-free two-stage visual token pruning framework that jointly preserves spatially distributed evidence and adapts to the hierarchical evolution of multimodal semantics, addressing a pressing need revealed through extensive mechanistic analyses of token redundancy and semantic evolution. Our theoretical analysis provides a principled characterization of the accuracy–efficiency trade-off. Extensive experiments across multiple MLLMs, image and video tasks, and token budgets consistently demonstrate the practical efficiency of LCHTP.

AI USE STATEMENT

In this work, generative AI tools were used exclusively to support manuscript preparation and presentation, including language refinement, improvements in clarity and readability, and the organization of experimental results into tables and figures. Generative AI was not used to develop research ideas or methods, formulate scientific claims, design experiments, analyze data, interpret results, conduct literature searches, or derive theoretical results. All AI-assisted revisions and visual materials were carefully reviewed and verified by the authors, who retain full responsibility for the final content of the manuscript.

REFERENCES

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. HiRED: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025. doi: 10.1609/aaai.v39i2.32171.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, pp. 19–35. Springer, 2024. doi: 10.1007/978-3-031-73004-7_2.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025. doi: 10.52202/085713-4899.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017. doi: 10.1109/CVPR.2017.670.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. In *Computer Vision – ECCV 2024*, pp. 390–406. Springer, 2024. doi: 10.1007/978-3-031-73010-8_23.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018. doi: 10.1109/CVPR.2018.00380.

Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Honggang Chen, Donglin Wang, Qingsen Yan, and Siteng Huang. Filter, correlate, compress: Training-free token reduction for MLLM acceleration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4601–4609, 2026. doi: 10.1609/aaai.v40i6.42460.

Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. ZipVL: Accelerating vision-language models through dynamic token sparsity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20477–20486, 2025.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. IVTP: Instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, pp. 214–230, 2024a.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024b.

Drew A. Hudson and Christopher D. Manning. 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*, pp. 6700–6709, 2019. doi: 10.1109/CVPR.2019.00686.

Lei Jiang, Weizhe Huang, Tongxuan Liu, Yuting Zeng, Jing Li, Lechao Cheng, and Xiaohua Xu. FoPru: Focal pruning for efficient large vision-language models. *arXiv preprint arXiv:2411.14164*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340, 2024.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3530–3543, 2026. doi: 10.1109/TPAMI.2025.3637265.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.342.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

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

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

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pp. 216–233. Springer, 2024d. doi: 10.1007/978-3-031-72658-3_13.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. doi: 10.52202/068431-0182.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *International Conference on Learning Representations*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.679.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774.

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867, 2025. doi: 10.1109/ICCV51701.2025.02122.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. doi: 10.1109/CVPR.2019.00851.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7614–7623, Abu Dhabi, UAE, 2025. Association for Computational Linguistics.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. VL-Cache: Sparsity and modality-aware KV cache compression for vision-language model inference acceleration. In *International Conference on Learning Representations*, 2025.

Ao Wang, Fengyuan Sun, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. [CLS] token tells everything needed for training-free efficient MLLMs. *arXiv preprint arXiv:2412.05819*, 2024.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9961–9980. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.505.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603, 2025. doi: 10.1109/CVPR52734.2025.01360.

Bingxin Xu, Yuzhang Shang, Yunhao Ge, Qian Lou, and Yan Yan. Freepruner: A training-free approach for large multimodal model acceleration. *arXiv preprint arXiv:2411.15446*, 2024.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1645–1653. ACM, 2017. doi: 10.1145/3123266.3123427.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025. doi: 10.1109/CVPR52734.2025.01843.

Te Yang, Jian Jia, Xiangyu Zhu, Weisong Zhao, Bo Wang, Yanhua Cheng, Yan Li, Shengyuan Liu, Quan Chen, Peng Jiang, et al. Enhancing instruction-following capability of visual-language models by reducing image redundancy. *arXiv preprint arXiv:2411.15453*, 2024.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025. doi: 10.1609/aaai.v39i21.34366.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 2024.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. VScan: Rethinking visual token reduction for efficient large vision-language models. *Transactions on Machine Learning Research*, 2026.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [CLS] attention is all you need for training-free visual token pruning: Make VLM inference faster. *arXiv preprint arXiv:2412.01818*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857. PMLR, 2025.

Yuke Zhu, Chi Xie, Shuang Liang, Bo Zheng, and Sheng Guo. Focusllava: A coarse-to-fine approach for efficient and effective visual token compression. *arXiv preprint arXiv:2411.14228*, 2024.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in MLLMs: Revisiting visual holistic context retention. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1327.

A Additional Experimental Details

A.1 Benchmarks and Evaluation Metrics

To comprehensively assess the model's visual understanding capabilities across multiple dimensions, we evaluate it on a range of benchmarks for image and video understanding. For ease of organization, we describe these benchmarks below in the chronological order in which they were first proposed.

VQAv2 (Goyal et al., 2017) evaluates visual question answering and extends the original VQA dataset by introducing complementary image pairs. The same question is paired with visually similar images that lead to different answers, reducing the influence of language priors and encouraging models to rely more on visual evidence.

VizWiz (Gurari et al., 2018) evaluates visual question answering in real-world assistive settings. Its images were taken by blind people using a mobile device and paired with questions reflecting actual visual information needs. The dataset contains challenging cases such as blurriness, underexposure, or incomplete framing, as well as questions that cannot be answered from existing visual content, so it focuses more on examining the visual understanding ability of the model in real scenes.

GQA (Hudson & Manning, 2019) focuses on compositional visual reasoning and constructs questions from objects, attributes, and relations represented in Visual Genome scene graphs. The questions cover object recognition, spatial relationship understanding, and multi-step reasoning, and a balanced answer distribution helps reduce the impact of language bias.

TextVQA (Singh et al., 2019) evaluates visual question answering in scenes where textual information is important for deriving the answer. It requires models to read text in images and reason jointly over the recognized text, surrounding visual context, and question semantics.

ScienceQA (Lu et al., 2022) evaluates multimodal question answering and reasoning in scientific domains. The task contained multiple choice questions from three subject areas of natural science, social science and language science, some of which provided visual context. The model needs to combine visual information with question semantics and related scientific knowledge to select the correct answer.

MME (Fu et al., 2025) evaluates both perceptual and cognitive capabilities through 14 subtasks. These tasks are mainly composed of perceptual tasks and cognitive tasks, covering basic visual perception, optical character recognition (OCR), commonsense reasoning, numerical computation, text translation, and code reasoning. The evaluation considers not only the correctness of a single question, but also the consistent correctness of related questions under the same image. We report MME scores according to the official evaluation protocol.

POPE (Li et al., 2023) evaluates object hallucination by formulating object-presence detection as a binary question-answering task. It uses three negative sample sampling strategies, random, popular and adversarial, to evaluate the performance of hallucination behavior under different conditions.

MM-Vet (Yu et al., 2024) evaluates the integrated use of multiple vision-language capabilities in open-ended tasks. The benchmark covers six core competencies: recognition, optical character recognition (OCR), knowledge, language generation, spatial cognition, and mathematics, as well as 16 combinations of these six competencies. Open-ended responses are scored by an evaluator based on a large language model.

MMBench and MMBench-CN (Liu et al., 2024d) evaluate multimodal perception and reasoning through multiple-choice questions organized by a hierarchical capability taxonomy. MMBench uses a cyclic evaluation strategy to change the order of candidate options to test the consistency of predictions. Mmbench-cn is the Chinese version corresponding to the English MMBench, which is used to evaluate the model's multimodal understanding ability in Chinese context.

MSVD-QA and MSRVTT-QA (Xu et al., 2017) evaluate video understanding through open-ended question answering. Their questions cover types such as *what*, *who*, *how*, *when*, and *where*, requiring models to reason over both visual content and temporal information. Following the evaluation protocol used by Video-ChatGPT (Maaz et al., 2024), we use an LLM-based evaluator to compare generated answers with reference answers and report Accuracy and the average answer-quality Score.

B Related Work

B.1 MLLMs and Their Bottlenecks

The rapid progress of Large Language Models (LLMs) has provided a strong foundation for multimodal reasoning and instruction following (Zhang et al., 2022; Touvron et al., 2023; Dubey et al., 2024). Building on these language backbones, Multimodal Large Language Models (MLLMs) extend language understanding and generation to visual inputs through diverse multimodal architectures (OpenAI, 2023; Gemini Team et al., 2023; Bai et al., 2023). Many open-source MLLMs further connect a vision encoder to an autoregressive language model through a modality-alignment interface and visual instruction tuning (Liu et al., 2023; 2024a). Representative systems such as LLaVA and LLaVA-NeXT demonstrate that stronger visual instruction tuning and higher-resolution image processing can substantially improve perception, OCR, and visual reasoning capabilities (Liu et al., 2023; 2024a;b). At the same time, current MLLMs still exhibit systematic visual shortcomings and hallucination-related errors, indicating that stronger language reasoning does not automatically guarantee reliable visual grounding (Tong et al., 2024; Huang et al., 2024b).

A common response to limited fine-grained visual perception is to increase visual resolution or to preserve more local image detail. Methods such as Feast Your Eyes and LLaVA-UHD explicitly adapt MLLMs to higher-resolution inputs, while LLaVA-NeXT and LLaVA-OneVision employ flexible image partitioning or high-resolution visual representations to improve downstream understanding (Luo et al., 2025; Guo et al., 2024; Liu et al., 2024b; Li et al., 2025). However, richer visual representations also lengthen the visual token sequence that must be processed by the language model. For example, LLaVA-1.5 commonly represents a 336×336 image with 576 visual tokens, whereas LLaVA-NeXT can produce substantially longer sequences for high-resolution inputs, and LLaVA-OneVision can further increase the number of visual tokens under high-resolution settings (Liu et al., 2024a;b; Li et al., 2025). Video-oriented MLLMs face an even larger burden because multiple frames contribute visual tokens to the same decoding context (Lin et al., 2024). As an additional recent example, Qwen2.5-VL adopts native dynamic-resolution visual processing for variable-size images and long videos, illustrating the continuing trend toward richer but more expensive visual inputs (Bai et al., 2025). These developments make visual-token efficiency an increasingly important problem for practical MLLM inference.

B.2 Visual Redundancy Identification

Efficient visual-token reduction first requires a criterion for distinguishing informative tokens from redundant ones. Existing approaches obtain this criterion from several complementary sources. Vision-centric methods estimate token importance from the structure of visual representations themselves. Some methods exploit encoder-side attention, particularly $[\mathrm{CLS}]$-to-patch attention, to identify visually salient tokens (Wang et al., 2024). Other approaches learn compact visual tokenizations by adaptively aggregating spatial information into a small set of representative tokens (Ryoo et al., 2021). Recent studies further explore unified reduction pipelines, redundancy-aware token aggregation, and complementary token selection based on semantic or attention cues (Yang et al., 2024; Xu et al., 2024).

Instruction-aware methods instead estimate redundancy according to the relationship between visual tokens and the current language input. These approaches commonly use text-to-vision attention or related cross-modal signals to identify visual tokens that are unlikely to affect the current response (Liu et al., 2024c; Zhu et al., 2024; Song et al., 2025). ZipVL dynamically allocates layer-wise token sparsity according to attention-score distributions, applying sparse attention to important tokens during prefilling and discarding less important tokens to reduce the KV-cache size during decoding (He et al., 2025). Other studies develop additional importance estimators, focal pruning criteria, or cache-aware sparsification strategies (Jiang et al., 2024; Tu et al., 2025; Ye et al., 2025). These methods demonstrate that visual redundancy is not determined only by image appearance; the instruction and the current model state can also change which visual evidence is useful.

More recently, VScan studies how visual-token redundancy evolves across both the visual encoding stage and the language decoding stage rather than treating token importance as a fixed property of a single layer (Zhang et al., 2026). Its analysis motivates a two-stage reduction strategy that combines global and local scanning during visual encoding with additional pruning inside the language model. This layer-aware perspective suggests that redundancy identification can depend on model depth rather than on one universal importance score. These findings establish model depth as an additional dimension for identifying visual redundancy.

B.3 Visual Token Compression and Pruning

After redundant information has been identified, existing methods reduce visual sequences through token merging, token pruning, or progressive combinations of both operations. The need for such compression becomes particularly apparent in high-resolution MLLMs, where Mini-Gemini-HD can extend high-resolution visual representations to 2,880 visual tokens, creating substantial inference overhead (Li et al., 2026). Early token-reduction research in vision transformers, represented by ToMe, merges similar tokens through lightweight matching and does not require additional training (Bolya et al., 2023). In MLLMs, LLaMA-VID and DeCo reduce visual sequence length through dedicated compression mechanisms, while IVTP introduces instruction-guided visual token pruning for more efficient language-model processing (Li et al., 2024; Yao et al., 2024; Huang et al., 2024a). These approaches show that shortening the visual sequence can substantially reduce inference cost, although different compression locations introduce different trade-offs between efficiency and information preservation.

A second line of work emphasizes training-free pruning or hybrid pruning-and-merging. LLaVA-PruMerge selects informative visual tokens and merges redundant representations according to key similarity (Shang et al., 2025). FasterVLM performs pruning before the language model by ranking visual tokens with encoder-side $[\mathrm{CLS}]$ attention, whereas FastV removes visual tokens inside early LLM layers according to attention-derived importance (Zhang et al., 2024; Chen et al., 2024). HiRED allocates token budgets across image partitions according to $[\mathrm{CLS}]$ attention and then selects informative tokens within each partition, while MustDrop combines spatial merging, text-guided pruning, and output-aware cache policies to reduce redundant visual information across multiple inference stages (Arif et al., 2025; Liu et al., 2024c). SparseVLM introduces cross-modal guidance together with adaptive sparsification, and PyramidDrop (PDrop) progressively reduces visual redundancy across multiple model stages (Zhang et al., 2025; Xing et al., 2025). VisionZip estimates visual-token importance from encoder-side attention and further consolidates redundant representations using key-similarity clustering, whereas DART adopts a duplication-aware strategy that selects pivot tokens and favors tokens with low duplication relative to those pivots (Yang et al., 2025; Wen et al., 2025). FiCoCo-V follows a training-free filter-correlate-compress design inside the vision encoder: it filters redundant visual tokens with an integrated redundancy criterion and recycles information from correlated discarded tokens into retained representations to mitigate information loss (Han et al., 2026). These methods improve inference efficiency while following different assumptions about where redundancy should be measured, how discarded information should be preserved, and whether compression should happen before or during language decoding. Operator-level methods such as FlashAttention and FlashAttention-2 reduce attention cost without changing the token sequence, making them complementary to visual-token reduction (Dao et al., 2022; Dao, 2024).

Recent work also places greater emphasis on spatial coverage and stage-wise reduction. HoloV argues that attention-first pruning can over-concentrate retained tokens in locally salient regions and therefore allocates token budgets across spatial crops to preserve more holistic visual context (Zou et al., 2025). VScan, meanwhile, explicitly divides reduction between visual encoding and language decoding, using global/local visual scans before an intermediate LLM pruning stage (Zhang et al., 2026). Together, these studies highlight two complementary considerations for efficient MLLMs: preserving spatially distributed visual evidence and adapting token reduction across model stages.

C Theoretical Analysis and Derivation

This section establishes a principled theoretical characterization of the two-stage pruning mechanism instantiated by LCHTP. The analysis focuses on a single question: why can LCHTP substantially reduce visual-token computation while preserving the predictive behavior of the unpruned model? The derivation connects spatially protected evidence retention, second-order visual recasting, delayed instruction-conditioned pruning, perturbation propagation, and decoder complexity within one unified framework.

C.1 Problem formulation

The input image is denoted by x. The vision encoder produces N projected visual tokens. The visual sequence is

$$V = (v_1, \ldots, v_N), \qquad v_i \in \mathbb{R}^D. \tag{13}$$

The symbol D denotes the hidden dimension of each projected visual token. The text instruction contains M embedded tokens. The text sequence is

$$T = (t_1, \ldots, t_M), \qquad t_j \in \mathbb{R}^D. \tag{14}$$

The multimodal decoder contains L transformer blocks. The hidden sequence before the first decoder block is denoted by $X^{(0)}$, and the mapping implemented by decoder block ℓ is denoted by F_ℓ. The unpruned decoder therefore satisfies

$$X^{(0)} = [T; V], \qquad X^{(\ell)} = F_\ell\Big(X^{(\ell-1)}\Big), \qquad \ell = 1, \ldots, L. \tag{15}$$

The notation $[T; V]$ denotes sequence concatenation. The index r denotes the decoder position used to predict the next output token. The matrix W denotes the output projection. The unpruned logit vector is $z = WX_r^{(L)}$.

C.2 LCHTP TWO-STAGE TOKEN REDUCTION

The first stage preserves the physical two-dimensional organization of the visual grid. The N visual positions are partitioned into C non-overlapping crops. The crop partition satisfies

$$\mathcal{C} = \{C_c\}_{c=1}^{C}, \qquad \bigcup_{c=1}^{C} C_c = \{1, \ldots, N\}, \qquad C_c \cap C_{c'} = \varnothing \quad \text{for } c \neq c'. \tag{16}$$

The integer $n_c = |C_c|$ denotes the number of visual tokens in crop C_c. The quantity H denotes the number of attention heads in the vision encoder. The shallow and deep visual layers used by LCHTP are indexed by ℓ_1 and ℓ_2, respectively. The attention matrix of head h at visual layer ℓ is denoted by $A^{(\ell,h)}$. The index 0 denotes the visual classification token used by the deep-layer global relevance score. The three score components are indexed numerically. The index $q = 1$ denotes shallow local relevance, $q = 2$ denotes deep global relevance, and $q = 3$ denotes feature-space diversity. For token $i \in C_c$, the first two component scores are

$$s_i^{(1)} = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{n_c - 1} \sum_{\substack{j \in C_c \\ j \neq i}} A_{ji}^{(\ell_1,h)}, \tag{17}$$

$$s_i^{(2)} = \frac{1}{H} \sum_{h=1}^{H} A_{0i}^{(\ell_2,h)}. \tag{18}$$

The crop prototype μ_c and the image-level prototype μ are

$$\mu_c = \frac{1}{n_c} \sum_{j \in C_c} v_j, \qquad \mu = \frac{1}{N} \sum_{j=1}^{N} v_j. \tag{19}$$

The diversity score is

$$s_i^{(3)} = \|v_i - \mu_c\|_2 + \|v_i - \mu\|_2. \tag{20}$$

For each component $q \in \{1, 2, 3\}$, the quantities s_q^- and s_q^+ denote the minimum and maximum component scores over the current physical crop. The normalized score is

$$\widehat{s}_i^{(q)} = \frac{s_i^{(q)} - s_q^-}{s_q^+ - s_q^- + \varepsilon}, \qquad q \in \{1, 2, 3\}. \tag{21}$$

The constant $\varepsilon > 0$ prevents division by zero. The fused LCHTP score is

$$s_i = \alpha \widehat{s}_i^{(1)} + \beta \widehat{s}_i^{(2)} + \gamma \widehat{s}_i^{(3)}, \qquad (\alpha, \beta, \gamma) = (0.4, 0.4, 0.2). \tag{22}$$

The first stage retains K_1 visual tokens. The ideal quota assigned to crop C_c is denoted by $\widetilde{k}_c$, and the integer quota is denoted by k_c. The quota allocation satisfies

$$\widetilde{k}_c = K_1 \frac{n_c}{\sum_{j=1}^{C} n_j}, \qquad k_c = \lfloor \widetilde{k}_c \rfloor + \delta_c, \qquad \delta_c \in \{0, 1\}, \qquad \sum_{c=1}^{C} k_c = K_1. \tag{23}$$

The largest-remainder rule determines δ_c. The retained support in crop C_c is denoted by Ω_c. The global retained support after the first stage is denoted by Ω. These sets satisfy

$$\Omega_c = \mathrm{TopK}_{i \in C_c}(s_i, k_c), \qquad \Omega = \bigcup_{c=1}^{C} \Omega_c, \qquad |\Omega| = K_1. \tag{24}$$

The complement Ω^c denotes the discarded visual positions. Each discarded visual token v_d indexed by $d \in \Omega^c$ is assigned to the most similar retained visual token v_s indexed by $s \in \Omega$. The corresponding retained-token index is

$$m(d) = \arg\max_{s \in \Omega} \cos(v_d, v_s). \tag{25}$$

The quantity τ_d measures the maximum cosine similarity between the discarded visual token v_d and the text-token set. The nonnegative recasting weight associated with v_d is denoted by w_d and is given by

$$\tau_d = \max_{1 \le j \le M} \cos(v_d, t_j), \qquad w_d = \left[\cos(v_d, v_{m(d)}) + \lambda \tau_d\right]_+. \tag{26}$$

The notation $[a]_+ = \max(a, 0)$ denotes the positive-part operator, and $\lambda \ge 0$ controls the contribution of text relevance. The discarded visual tokens assigned to the retained visual token v_s form the set

$$D_s = \{v_d : d \in \Omega^c,\ m(d) = s\}. \tag{27}$$

The scalar $b > 0$ denotes the anchor weight assigned to the original retained visual token v_s. The corresponding recast representation is

$$\widetilde{v}_s = \frac{b v_s + \sum_{v_d \in D_s} w_d v_d}{b + \sum_{v_d \in D_s} w_d}. \tag{28}$$

The first-stage visual sequence is

$$\widetilde{V} = (\widetilde{v}_s)_{s \in \Omega}, \qquad |\widetilde{V}| = K_1. \tag{29}$$

The first d decoder blocks operate on the sequence containing $\widetilde{V}$, where $1 \le d < L$. The visual and text positions at depth d are denoted by $\mathcal{V}_d$ and $\mathcal{T}_d$, respectively. The query vector at text position r in attention head h is denoted by q_r^h, and the key vector at visual position i is denoted by k_i^h. The head dimension is denoted by d_h. The instruction-level query for head h is

$$\bar{q}_h = \frac{1}{|\mathcal{T}_d|} \sum_{r \in \mathcal{T}_d} q_r^h. \tag{30}$$

The instruction-to-visual relevance score assigned to visual position i at decoder depth d is

$$a_i^{(d)} = \frac{1}{H} \sum_{h=1}^{H} \frac{\exp\left(\bar{q}_h (k_i^h)^\top / \sqrt{d_h}\right)}{\sum_{j \in \mathcal{V}_d} \exp\left(\bar{q}_h (k_j^h)^\top / \sqrt{d_h}\right)}. \tag{31}$$

The vector $a^{(d)} = (a_i^{(d)})_{i \in \mathcal{V}_d}$ denotes the complete instruction-derived relevance distribution at depth d. The head-wise normalization in Eq. (31) implies $\sum_{i \in \mathcal{V}_d} a_i^{(d)} = 1$ and $a_i^{(d)} \ge 0$. The second stage retains K_2 visual positions. The retained support is

$$\Phi = \mathrm{TopK}_{i \in \mathcal{V}_d}(a_i^{(d)}, K_2), \qquad |\Phi| = K_2. \tag{32}$$

The hidden sequence immediately after the second pruning step is denoted by $\widehat{X}^{(d)}$. The sequence $\widehat{X}^{(d)}$ contains every text position in $\mathcal{T}_d$ and only the visual positions in Φ. For every $\ell > d$, $\widehat{X}^{(\ell)}$ denotes the hidden sequence produced by applying the remaining decoder blocks to $\widehat{X}^{(d)}$. The resulting two-stage logit vector is denoted by $\widehat{z}$. The layer-average visual-token budget is denoted by B. The budget identity is

$$B = \frac{dK_1 + (L - d)K_2}{L}. \tag{33}$$

C.3 Regularity conditions and auxiliary quantities

The analysis introduces a normalized task-influence weight ρ_i for every original visual token. The quantity ρ_i measures the first-order contribution of token i to the final answer representation. The

influence weights satisfy

$$\rho_i \geq 0, \qquad \sum_{i=1}^{N} \rho_i = 1. \tag{34}$$

For any visual index set S, the quantity $\rho(S)$ denotes the total influence carried by S. The definition is $\rho(S) = \sum_{i \in S} \rho_i$. For any current visual index set, the notation S^c denotes the complement of S. The notation $|S|$ denotes set cardinality. For perturbation comparisons between sequences with different visual-token counts, the analysis uses the canonical original visual-index space: each retained or recast visual state is placed at its original visual index, while each removed visual position contributes the zero vector. This index alignment makes every hidden-state difference well-defined without changing the identities of the retained tokens. The notation $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices, while $\|\cdot\|_\infty$ denotes the maximum absolute coordinate. The operator $\mathbb{E}$ denotes expectation, and the operator $\mathbb{P}$ denotes probability.

Assumption 1 (Stable decoder dynamics)**.** *Each decoder block is Lipschitz continuous on the hidden-state region traversed by the unpruned and pruned computations. The scalar $\Lambda_\ell \geq 0$ denotes the Lipschitz constant of decoder block ℓ. The constants satisfy*

$$\|F_\ell(U) - F_\ell(V)\|_2 \leq \Lambda_\ell \|U - V\|_2. \tag{35}$$

For integers $a \leq b$, the product $\Gamma_{a:b}$ denotes the accumulated Lipschitz factor between blocks a and b. The definition is

$$\Gamma_{a:b} = \prod_{\ell=a}^{b} \Lambda_\ell. \tag{36}$$

The scalar $\Lambda_0 \geq 0$ bounds the output projection. The output projection satisfies

$$\|Wu\|_\infty \leq \Lambda_0 \|u\|_2. \tag{37}$$

Assumption 2 (Crop-wise score calibration)**.** *For each crop C_c, we introduce a positive scaling coefficient ϑ_c and an offset term ν_c to relate the fused score s_i to the latent task influence ρ_i through an affine approximation. The scalar $\xi_1 \geq 0$ denotes the uniform calibration error. The calibration condition is*

$$|\rho_i - (\vartheta_c s_i + \nu_c)| \leq \xi_1, \qquad i \in C_c. \tag{38}$$

Assumption 3 (Smooth local recasting)**.** *For each retained visual token v_s indexed by $s \in \Omega$, the set*

$$G_s = \{v_s\} \cup D_s \tag{39}$$

denotes the local recasting group associated with v_s. The normalization factor ζ_s is

$$\zeta_s = b + \sum_{v_d \in D_s} w_d. \tag{40}$$

The normalized group weights are

$$\pi_{ss} = \frac{b}{\zeta_s}, \qquad \pi_{sd} = \frac{w_d}{\zeta_s} \quad \textit{for } v_d \in D_s. \tag{41}$$

The weights satisfy

$$\sum_{v_x \in G_s} \pi_{sx} = 1, \qquad \sum_{v_x \in G_s} \pi_{sx}(v_x - \widetilde{v}_s) = 0. \tag{42}$$

The scalar $r_s \geq 0$ denotes the maximum feature-space radius of the recasting group G_s. The radius condition is

$$\max_{v_x \in G_s} \|v_x - \widetilde{v}_s\|_2 \leq r_s. \tag{43}$$

The function $\varphi : \mathbb{R}^D \to \mathbb{R}^D$ denotes a generic local downstream contribution map induced by attention and subsequent linear transformations. The scalar $\chi \geq 0$ bounds the spectral norm of its Hessian. The smoothness condition is

$$\sup_v \|\nabla^2 \varphi(v)\|_2 \leq \chi. \tag{44}$$

Assumption 4 (Instruction-query concentration and semantic stabilization)**.** *For a fixed attention head, the analysis suppresses the head index and writes the query vector at text position j as q_j. The query vector is modeled as a shared semantic component plus a zero-mean perturbation. The*

semantic component is denoted by $q^\star$, and the perturbation is denoted by η_j. The model is

$$q_j = q^\star + \eta_j, \qquad \mathbb{E}[\eta_j] = 0, \qquad \mathbb{E}\|\eta_j\|_2^2 \le \sigma^2. \tag{45}$$

The scalar $\varrho \in [0, 1]$ bounds pairwise perturbation correlation. The correlation condition is

$$\mathbb{E}[\eta_i^\top \eta_j] \le \varrho\sigma^2, \qquad i \ne j. \tag{46}$$

The vector $p^{\star,(d)}$ denotes the oracle task-relevance distribution over the K_1 visual positions present at decoder depth d, and $p_i^{\star,(d)}$ denotes its component at visual position i. The vector $a^{(d)}$ denotes the instruction-derived relevance distribution at the same depth. For any visual support S, the expressions $p^{\star,(d)}(S)$ and $a^{(d)}(S)$ denote the sums of the corresponding relevance entries over S. The scalar $\delta_d \ge 0$ denotes their discrepancy. The semantic-stabilization condition is

$$\|a^{(d)} - p^{\star,(d)}\|_1 \le \delta_d. \tag{47}$$

The integer $d_0 < d$ denotes an earlier reference depth. The delayed relevance estimate is assumed to be at least as stable as the earlier estimate. The stability condition is

$$\delta_d \le \delta_{d_0}. \tag{48}$$

Assumption 5 (Attention-value boundedness and compute model)**.** *The vector u_i denotes the value contribution associated with intermediate visual position i. The scalar $U \ge 0$ satisfies*

$$\|u_i\|_2 \le U. \tag{49}$$

The integer Q denotes the number of non-visual tokens processed by each decoder block. The positive constants κ_1 and κ_2 collect architecture-dependent coefficients for linear projection/MLP computation and quadratic attention computation, respectively. The function $\Xi(n)$ denotes the idealized cost of one decoder block processing n visual tokens. Its definition is

$$\Xi(n) = \kappa_1(Q + n)D^2 + \kappa_2(Q + n)^2 D. \tag{50}$$

The scalar $\Xi_\circ \ge 0$ denotes the additional overhead introduced by scoring, indexing, and token selection.

For crop C_c, the set $S_c^\star$ denotes the size-k_c subset that maximizes the true task influence $\rho(S)$. The quantity $\epsilon_{11}(K_1)$ denotes the influence mass that cannot be captured even by the crop-wise oracle under the allocated quotas. The definition is

$$\epsilon_{11}(K_1) = 1 - \sum_{c=1}^{C} \sum_{i \in S_c^\star} \rho_i. \tag{51}$$

The nonnegative coefficients ω_s weight the local recasting groups and satisfy

$$\omega_s \ge 0, \qquad \sum_{s \in \Omega} \omega_s = 1. \tag{52}$$

The positive constants c_1 and c_2 convert influence loss and recasting approximation error into a common decoder-input perturbation scale. The constants are chosen to upper-bound the representation-level deviation induced by the first-stage replacement of the original visual sequence with the recast sequence. The resulting first-stage distortion bound is

$$\mathcal{E}_1(K_1) = c_1 \left[\epsilon_{11}(K_1) + 2K_1\xi_1\right] + \frac{c_2\chi}{2} \sum_{s \in \Omega} \omega_s r_s^2. \tag{53}$$

The oracle size-K_2 support at depth d is denoted by $\Phi^\star$. The set is

$$\Phi^\star = \mathrm{TopK}\Big(p^{\star,(d)}, K_2\Big). \tag{54}$$

The quantity $\epsilon_{21}(K_2, d)$ denotes the oracle relevance mass discarded by a size-K_2 support at depth d. The definition is

$$\epsilon_{21}(K_2, d) = 1 - p^{\star,(d)}(\Phi^\star). \tag{55}$$

The second-stage distortion bound is

$$\mathcal{E}_2(K_2, d) = 2U \left[\epsilon_{21}(K_2, d) + 3\delta_d\right]. \tag{56}$$

The additional discrepancy term in Eq. (56) accounts for the conversion from oracle relevance mass to the instruction-derived relevance mass that actually determines the second-stage aggregation.

C.4 UNIFIED THEOREM

Theorem 2 (Accuracy–efficiency preservation under two-stage visual-token reduction)**.** *Under Assumptions 1–5, the following eight statements characterize how the two-stage design controls approximation error while reducing decoder computation.*

(I) Crop-wise influence preservation and spatial anti-collapse. *The crop-wise selection stage preserves task influence locally before it aggregates the retained tokens globally. For every crop C_c, the LCHTP support Ω_c satisfies*

$$\sum_{i\in\Omega_c}\rho_i \geq \sum_{i\in S_c^\star}\rho_i - 2k_c\xi_1. \tag{57}$$

The retained support therefore satisfies

$$\rho(\Omega) \geq 1 - \epsilon_{11}(K_1) - 2K_1\xi_1, \tag{58}$$

and the total discarded influence satisfies

$$\rho(\Omega^c) \leq \epsilon_{11}(K_1) + 2K_1\xi_1. \tag{59}$$

If all crops have equal capacity and $K_1 \geq C$, then every crop receives at least one retained token. The quota consequently satisfies

$$k_c \geq \left\lfloor \frac{K_1}{C} \right\rfloor \geq 1. \tag{60}$$

Method-design implication. *The bound in Eq. (57) isolates the error introduced when the fused score s_i is used as a surrogate for the latent task influence ρ_i. The quantity ξ_1 measures the local calibration mismatch, so a smaller ξ_1 makes the crop-wise Top-k_c ranking approach the crop-wise oracle ranking. The quantity $\epsilon_{11}(K_1)$ measures the task influence that no size-K_1 crop-wise allocation can recover under the imposed quotas, so a smaller $\epsilon_{11}(K_1)$ directly increases the retained influence in Eq. (58). The spatial quota in Eq. (60) additionally prevents a globally dominant region from consuming the entire token budget. The first-stage design therefore separates two sources of error: the unavoidable budget tail $\epsilon_{11}(K_1)$ and the score-calibration penalty $2K_1\xi_1$. This separation explains why crop-wise protection can preserve low-saliency but task-critical evidence without abandoning saliency-based selection inside each crop.*

(II) Second-order approximation induced by visual recasting. *For each retained visual token v_s indexed by $s \in \Omega$, the original weighted group contribution is*

$$\mathcal{A}_s = \sum_{v_x\in G_s}\pi_{sx}\varphi(v_x). \tag{61}$$

The recast group contribution is

$$\widetilde{\mathcal{A}}_s = \varphi(\widetilde{v}_s). \tag{62}$$

The groupwise recasting error satisfies

$$\|\mathcal{A}_s - \widetilde{\mathcal{A}}_s\|_2 \leq \frac{\chi}{2}\sum_{v_x\in G_s}\pi_{sx}\|v_x - \widetilde{v}_s\|_2^2 \leq \frac{\chi}{2}r_s^2. \tag{63}$$

The scalar ϵ_{12} denotes the weighted aggregate of the groupwise recasting errors and is defined by

$$\epsilon_{12} = \sum_{s\in\Omega}\omega_s\|\mathcal{A}_s - \widetilde{\mathcal{A}}_s\|_2. \tag{64}$$

The aggregate error satisfies

$$\epsilon_{12} \leq \frac{\chi}{2}\sum_{s\in\Omega}\omega_s r_s^2. \tag{65}$$

Method-design implication. *The recasting mechanism does not treat a discarded token as pure information loss. The barycentric identity in Eq. (42) forces the weighted first-order displacement around $\widetilde{v}_s$ to vanish, so the dominant local approximation term becomes second order rather than first order. The radius r_s measures how tightly the discarded tokens assigned to v_s cluster around the recast representative. A smaller r_s therefore decreases the local error quadratically through Eq. (63). The curvature constant χ measures how nonlinear the downstream contribution map is in the local neighborhood, so a smaller χ further reduces the penalty of replacing the group with a single recast token. The aggregate inequality in Eq. (65) shows that the first stage can preserve information from*

discarded tokens without increasing the post-pruning sequence length. This property explains why the recasting step improves fidelity while maintaining the computational advantage created by token removal.

(III) Variance reduction of the instruction-level query. *For a fixed attention head, the mean instruction query is*

$$\bar{q} = \frac{1}{M}\sum_{j=1}^{M} q_j. \tag{66}$$

The mean query satisfies

$$\mathbb{E}\|\bar{q} - q^\star\|_2^2 \le \sigma^2 \frac{1 + (M-1)\varrho}{M} = \sigma^2\left[\varrho + \frac{1-\varrho}{M}\right]. \tag{67}$$

A query taken from a single text position satisfies

$$\mathbb{E}\|q_M - q^\star\|_2^2 \le \sigma^2. \tag{68}$$

When $\varrho = 0$, the mean query satisfies

$$\mathbb{E}\|\bar{q} - q^\star\|_2^2 \le \frac{\sigma^2}{M}, \tag{69}$$

so the corresponding upper bound is reduced by a factor of M relative to the single-position noise scale.

Method-design implication. *The decomposition $q_j = q^\star + \eta_j$ separates the shared instruction semantics from token-specific perturbations. The average in Eq. (66) preserves the shared component $q^\star$ while averaging the perturbations η_j. The factor $\varrho + (1-\varrho)/M$ in Eq. (67) makes the benefit explicit: the uncorrelated portion of the query noise shrinks as $1/M$, while only the correlated component remains as an irreducible floor. The single-position query in Eq. (68) does not receive this averaging benefit. The instruction-level query therefore supplies the second-stage ranking with a more stable semantic direction whenever the token-specific perturbations are not perfectly correlated. The resulting relevance scores become less sensitive to the identity of one terminal text token, which reduces the risk that paraphrasing or positional effects alter the Top-K_2 visual support.*

(IV) Controlled semantic loss at the delayed pruning depth. *The oracle relevance mass discarded by the actual second-stage support Φ satisfies*

$$p^{\star,(d)}(\Phi^c) \le \epsilon_{21}(K_2, d) + 2\delta_d. \tag{70}$$

The instruction-derived relevance mass discarded by Φ satisfies

$$a^{(d)}(\Phi^c) \le \epsilon_{21}(K_2, d) + 3\delta_d. \tag{71}$$

The relevance-weighted visual aggregation before second-stage pruning is

$$o^{(d)} = \sum_{i \in \mathcal{V}_d} a_i^{(d)} u_i. \tag{72}$$

The corresponding aggregation after retaining Φ and renormalizing the retained relevance mass is

$$\widehat{o}^{(d)} = \sum_{i \in \Phi} \frac{a_i^{(d)}}{1 - a^{(d)}(\Phi^c)} u_i. \tag{73}$$

The perturbation induced by the second-stage pruning satisfies

$$\|o^{(d)} - \widehat{o}^{(d)}\|_2 \le \mathcal{E}_2(K_2, d). \tag{74}$$

If the later depth satisfies

$$\epsilon_{21}(K_2, d) \le \epsilon_{21}(K_2, d_0), \tag{75}$$

and at least one of Eqs. (48) and (75) is strict, then

$$\mathcal{E}_2(K_2, d) < \mathcal{E}_2(K_2, d_0). \tag{76}$$

Method-design implication. *The quantity $\epsilon_{21}(K_2, d)$ represents the relevance tail that remains unavoidable even when an oracle chooses the best size-K_2 support at depth d. The discrepancy δ_d represents the additional error caused by estimating that oracle relevance with the instruction-derived distribution $a^{(d)}$. Eq. (70) shows that the actual support pays at most two relevance-estimation penalties beyond the oracle tail, while Eq. (71) adds one further discrepancy when the proof converts*

oracle mass into the instruction-derived mass that drives the aggregation. The distortion bound in Eq. (74) therefore decreases when semantic concentration reduces $\epsilon_{21}(K_2, d)$ or when cross-modal stabilization reduces δ_d. The delayed pruning design exploits exactly this depth dependence. The second stage waits until the decoder has developed a more concentrated and more reliably estimated task-relevance distribution, so the same K_2 budget can produce a smaller representation perturbation than an equally aggressive earlier pruning step.

(V) End-to-end logit stability and prediction preservation. *The total logit perturbation satisfies*

$$\|z - \widehat{z}\|_\infty \le \Delta(K_1, K_2, d), \tag{77}$$

where

$$\Delta(K_1, K_2, d) = \Lambda_0 \left[\Gamma_{1:L}\mathcal{E}_1(K_1) + \Gamma_{d+1:L}\mathcal{E}_2(K_2, d)\right]. \tag{78}$$

The first term in Eq. (78) propagates the first-stage perturbation through all L decoder blocks, while the second term propagates the second-stage perturbation only through the remaining $L - d$ blocks. The unpruned model prediction is denoted by $y^\star$. The corresponding logit margin is

$$\mathfrak{m}(x, T) = z_{y^\star} - \max_{y \ne y^\star} z_y. \tag{79}$$

The two-stage computation preserves the unpruned prediction whenever

$$\mathfrak{m}(x, T) > 2\Delta(K_1, K_2, d). \tag{80}$$

Let $\mathcal{D}$ denote the evaluation distribution over image–instruction–label triples. The symbol y denotes the ground-truth output label or token. Let $\widehat{y}_0$ denote the unpruned prediction, and let $\widehat{y}_2$ denote the two-stage prediction. The disagreement probability satisfies

$$\mathbb{P}_{\mathcal{D}}[\widehat{y}_2 \ne \widehat{y}_0] \le F_{\mathfrak{m}}(2\Delta(K_1, K_2, d)), \tag{81}$$

where $F_{\mathfrak{m}}(u) = \mathbb{P}_{\mathcal{D}}[\mathfrak{m}(x, T) \le u]$ denotes the cumulative distribution function of the unpruned logit margin. If $\mathcal{R}_0$ and $\mathcal{R}_2$ denote the task error rates of the unpruned and two-stage computations, respectively, then

$$\mathcal{R}_2 \le \mathcal{R}_0 + F_{\mathfrak{m}}(2\Delta(K_1, K_2, d)). \tag{82}$$

Method-design implication. *Eq. (78) shows that the two pruning stages affect the final prediction through different propagation horizons. A reduction in $\mathcal{E}_1(K_1)$ has a global effect because the corresponding perturbation traverses every decoder block, whereas a reduction in $\mathcal{E}_2(K_2, d)$ is weighted only by $\Gamma_{d+1:L}$. The margin condition in Eq. (80) converts these representation-level errors into a prediction-level criterion: the two-stage prediction cannot change when the original decision margin dominates twice the maximum logit perturbation. The distributional bounds in Eqs. (81) and (82) further show that only examples with sufficiently small unpruned margins are vulnerable to prediction changes under the derived perturbation bound. The architecture therefore benefits from preserving more information in the early stage and postponing aggressive pruning until a later depth, because the later perturbation has both a smaller semantic term and a shorter propagation path.*

(VI) Existence of a beneficial nonuniform two-stage allocation. *The normalized pruning depth is*

$$\theta = \frac{d}{L}. \tag{83}$$

For a fixed layer-average budget B, the scalar $u \ge 0$ transfers visual tokens from the later stage to the earlier stage while preserving the same average budget. The stage budgets are

$$K_1(u) = B + (1 - \theta)u, \qquad K_2(u) = B - \theta u. \tag{84}$$

The parameterization in Eq. (84) satisfies

$$\theta K_1(u) + (1 - \theta)K_2(u) = B. \tag{85}$$

The propagated hidden-state distortion without the common output factor Λ_0 is

$$J(u, d) = \Gamma_{1:L}\mathcal{E}_1(K_1(u)) + \Gamma_{d+1:L}\mathcal{E}_2(K_2(u), d). \tag{86}$$

Suppose that $\mathcal{E}_1$ and the first argument of $\mathcal{E}_2$ are differentiable and strictly decreasing in their token budgets. The notation $\partial_1\mathcal{E}_2$ denotes differentiation with respect to the first argument of $\mathcal{E}_2$. If

$$(1 - \theta)\Gamma_{1:L}\left|\mathcal{E}_1'(B)\right| > \theta\Gamma_{d+1:L}\left|\partial_1\mathcal{E}_2(B, d)\right|, \tag{87}$$

then the directional derivative at the uniform allocation satisfies

$$\left.\frac{\partial J(u, d)}{\partial u}\right|_{u=0} < 0. \tag{88}$$

A scalar $u_0 > 0$ therefore exists such that

$$J(u, d) < J(0, d) \qquad \textit{for every } u \in (0, u_0]. \tag{89}$$

If the delayed depth additionally satisfies

$$\Gamma_{d+1:L}\mathcal{E}_2(B, d) \le \Gamma_{d_0+1:L}\mathcal{E}_2(B, d_0), \tag{90}$$

then the same average budget admits

$$J(u, d) < J(0, d) \le J(0, d_0). \tag{91}$$

Method-design implication. *The parameter u describes a budget redistribution rather than a budget increase. Since a positive u gives additional visual tokens to the first stage and removes a compensating number of tokens from the second stage, Eq. (85) keeps the average budget unchanged. The inequality in Eq. (87) compares the propagated marginal accuracy benefit of one additional early-stage token with the propagated marginal cost of removing later-stage capacity. When the early-stage benefit dominates, Eq. (88) becomes negative and a small positive transfer strictly lowers the total distortion. The condition in Eq. (90) then adds the benefit of delaying the second pruning depth. The result therefore explains why a nonuniform schedule with $K_1 > K_2$ can outperform a uniform allocation even when both schedules consume the same layer-average number of visual tokens.*

(VII) Strict decoder-compute reduction. *The cost of the unpruned decoder is*

$$\Xi_0 = L\Xi(N). \tag{92}$$

The cost of the two-stage decoder is

$$\Xi_2 = d\Xi(K_1) + (L - d)\Xi(K_2) + \Xi_\circ. \tag{93}$$

The net decoder-compute saving satisfies

$$\begin{aligned}\Xi_0 - \Xi_2 = {} & L\kappa_1 D^2(N - B) \\ & + L\kappa_2 D\left[2Q(N - B) + N^2 - \theta K_1^2 - (1 - \theta)K_2^2\right] - \Xi_\circ.\end{aligned} \tag{94}$$

If $K_2 < K_1 < N$ and

$$\Xi_\circ < L\Xi(N) - d\Xi(K_1) - (L - d)\Xi(K_2), \tag{95}$$

then

$$\Xi_2 < \Xi_0. \tag{96}$$

The idealized decoder speedup factor is

$$\mathcal{S} = \frac{\Xi_0}{\Xi_2} > 1. \tag{97}$$

Under the experimental relation $K_2 = \frac{2}{3}K_1$, the budget identity gives

$$K_1 = \frac{3B}{2 + \theta}, \qquad K_2 = \frac{2B}{2 + \theta}. \tag{98}$$

The visual-only quadratic workload then satisfies

$$\theta K_1^2 + (1 - \theta)K_2^2 = B^2\frac{4 + 5\theta}{(2 + \theta)^2}. \tag{99}$$

The corresponding reduction factor relative to the unpruned N-token quadratic workload is

$$\frac{N^2}{\theta K_1^2 + (1 - \theta)K_2^2} = \left(\frac{N}{B}\right)^2 \frac{(2 + \theta)^2}{4 + 5\theta}. \tag{100}$$

Method-design implication. *The cost model in Eq. (50) contains a term that grows linearly with sequence length and an attention term that grows quadratically with sequence length. Eq. (94) separates the total saving into the linear gain, the quadratic gain, and the pruning overhead $\Xi_\circ$. The quantity $N - B$ controls the linear saving, while the term $N^2 - \theta K_1^2 - (1 - \theta)K_2^2$ controls the visual part of the quadratic saving. The second-stage reduction is especially valuable because the smaller budget K_2 is used for all remaining $L - d$ blocks. The condition in Eq. (95) requires the implementation overhead to remain smaller than the compute removed from the decoder. The resulting inequality in Eq. (96) therefore formalizes the acceleration mechanism: LCHTP reduces actual decoder workload rather than merely replacing one expensive operation with another.*

(VIII) Marginal-distortion balance at an interior optimum. *Let B_0 denote a prescribed upper bound on the layer-average visual-token budget. The continuous relaxation of the distortion-minimization problem is*

$$\min_{K_1,K_2,d} \Delta(K_1, K_2, d) \tag{101}$$

subject to

$$\theta K_1 + (1-\theta)K_2 \le B_0, \qquad K_2 \le K_1 \le N. \tag{102}$$

For a fixed interior depth d, any interior optimum with an active budget constraint satisfies

$$\frac{-\Gamma_{1:L}\mathcal{E}_1'(K_1)}{\theta} = \frac{-\Gamma_{d+1:L}\partial_1\mathcal{E}_2(K_2,d)}{1-\theta}. \tag{103}$$

Method-design implication. *The numerator on the left side of Eq. (103) measures the propagated reduction in first-stage distortion produced by one additional early-stage token, while the numerator on the right side measures the corresponding propagated reduction in second-stage distortion produced by one additional later-stage token. The denominators θ and $1-\theta$ convert these gains into marginal benefits per unit of layer-average budget. The equality therefore characterizes a balanced operating point at which neither stage can reduce the total distortion more efficiently by receiving an infinitesimal additional budget. The condition gives a principled interpretation of the LCHTP allocation instead of treating K_1, K_2, and d as unrelated pruning hyperparameters. The discrete implementation can approximate this continuous balance by selecting the feasible integer budgets whose realized average budget is closest to the target.*

Proof. **Proof of (I).** For crop C_c, the affine surrogate influence is

$$\widetilde{\rho}_i = \vartheta_c s_i + \nu_c. \tag{104}$$

Assumption 2 implies the pointwise deviation

$$|\rho_i - \widetilde{\rho}_i| \le \xi_1, \qquad i \in C_c. \tag{105}$$

For every size-k_c subset $S \subseteq C_c$, Eq. (105) implies

$$\left|\sum_{i\in S}\rho_i - \sum_{i\in S}\widetilde{\rho}_i\right| \le k_c\xi_1. \tag{106}$$

The positivity of ϑ_c preserves the ordering induced by s_i, so the Top-k_c support Ω_c also maximizes $\sum_{i\in S}\widetilde{\rho}_i$ among all size-k_c subsets of C_c. The optimality relation is

$$\sum_{i\in\Omega_c}\widetilde{\rho}_i \ge \sum_{i\in S_c^\star}\widetilde{\rho}_i. \tag{107}$$

Eqs. (106) and (107) therefore give

$$\begin{aligned}\sum_{i\in\Omega_c}\rho_i &\ge \sum_{i\in\Omega_c}\widetilde{\rho}_i - k_c\xi_1 &(108)\\ &\ge \sum_{i\in S_c^\star}\widetilde{\rho}_i - k_c\xi_1 &(109)\\ &\ge \sum_{i\in S_c^\star}\rho_i - 2k_c\xi_1. &(110)\end{aligned}$$

Eq. (110) proves Eq. (57). The crop supports are disjoint, and Eq. (23) gives $\sum_c k_c = K_1$. Summing Eq. (110) over all crops yields

$$\begin{aligned}\rho(\Omega) &= \sum_{c=1}^{C}\sum_{i\in\Omega_c}\rho_i\\ &\ge \sum_{c=1}^{C}\sum_{i\in S_c^\star}\rho_i - 2\xi_1\sum_{c=1}^{C}k_c\\ &= 1 - \epsilon_{11}(K_1) - 2K_1\xi_1. &(111)\end{aligned}$$

Eq. (111) proves Eq. (58), and the identity $\rho(\Omega^c) = 1 - \rho(\Omega)$ proves Eq. (59). When all crops have equal capacity, Eq. (23) gives $\widetilde{k}_c = K_1/C$. The largest-remainder rule therefore implies

$$k_c \geq \left\lfloor \frac{K_1}{C} \right\rfloor. \tag{112}$$

The assumption $K_1 \geq C$ makes the right side of Eq. (112) at least one, which proves Eq. (60).

Proof of (II). For each $v_x \in G_s$, define the local displacement

$$\Delta_{sx} = v_x - \widetilde{v}_s. \tag{113}$$

Taylor's theorem with integral remainder gives

$$\varphi(v_x) = \varphi(\widetilde{v}_s) + D\varphi(\widetilde{v}_s)[\Delta_{sx}] + R_{sx}, \tag{114}$$

$$R_{sx} = \int_0^1 (1-t) D^2\varphi(\widetilde{v}_s + t\Delta_{sx})[\Delta_{sx}, \Delta_{sx}]\, dt. \tag{115}$$

The operators $D\varphi$ and $D^2\varphi$ denote the first and second Fréchet derivatives of φ. Assumption 3 and Eq. (44) imply

$$\begin{aligned} \|R_{sx}\|_2 &\leq \int_0^1 (1-t)\chi \|\Delta_{sx}\|_2^2\, dt \\ &= \frac{\chi}{2}\|\Delta_{sx}\|_2^2. \end{aligned} \tag{116}$$

The barycentric identity in Eq. (42) yields

$$\sum_{v_x \in G_s} \pi_{sx}\Delta_{sx} = 0. \tag{117}$$

Linearity of $D\varphi(\widetilde{v}_s)$ therefore gives

$$\sum_{v_x \in G_s} \pi_{sx} D\varphi(\widetilde{v}_s)[\Delta_{sx}] = 0. \tag{118}$$

Substituting Eq. (114) into Eq. (61) and using Eq. (118) gives

$$\begin{aligned} \mathcal{A}_s &= \varphi(\widetilde{v}_s) + \sum_{v_x \in G_s} \pi_{sx} R_{sx} \\ &= \widetilde{\mathcal{A}}_s + \sum_{v_x \in G_s} \pi_{sx} R_{sx}. \end{aligned} \tag{119}$$

The triangle inequality and Eq. (116) imply

$$\begin{aligned} \|\mathcal{A}_s - \widetilde{\mathcal{A}}_s\|_2 &\leq \sum_{v_x \in G_s} \pi_{sx} \|R_{sx}\|_2 \\ &\leq \frac{\chi}{2} \sum_{v_x \in G_s} \pi_{sx} \|v_x - \widetilde{v}_s\|_2^2 \\ &\leq \frac{\chi}{2} r_s^2, \end{aligned} \tag{120}$$

where the last step uses Eq. (43) and $\sum_{v_x \in G_s} \pi_{sx} = 1$. Eq. (120) proves Eq. (63). Multiplying Eq. (120) by ω_s, summing over $s \in \Omega$, and using Eq. (64) yields

$$\epsilon_{12} \leq \frac{\chi}{2} \sum_{s \in \Omega} \omega_s r_s^2, \tag{121}$$

which proves Eq. (65).

Proof of (III). Eq. (45) implies

$$\bar{q} - q^\star = \frac{1}{M} \sum_{j=1}^{M} \eta_j. \tag{122}$$

The squared norm expands as

$$\left\| \sum_{j=1}^{M} \eta_j \right\|_2^2 = \sum_{j=1}^{M} \|\eta_j\|_2^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \eta_i^\top \eta_j. \tag{123}$$

Taking expectations and applying Eqs. (45) and (46) gives

$$\begin{aligned} \mathbb{E}\|\bar{q} - q^\star\|_2^2 &= \frac{1}{M^2} \left[\sum_{j=1}^{M} \mathbb{E}\|\eta_j\|_2^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \mathbb{E}[\eta_i^\top \eta_j] \right] \\ &\leq \frac{M\sigma^2 + M(M-1)\varrho\sigma^2}{M^2} \\ &= \sigma^2 \left[\varrho + \frac{1-\varrho}{M} \right]. \end{aligned} \tag{124}$$

Eq. (124) proves Eq. (67). Eq. (45) directly gives

$$\mathbb{E}\|q_M - q^\star\|_2^2 = \mathbb{E}\|\eta_M\|_2^2 \leq \sigma^2, \tag{125}$$

which proves Eq. (68). Setting $\varrho = 0$ in Eq. (124) proves Eq. (69).

Proof of (IV). The set Φ maximizes the estimated relevance mass $a^{(d)}(S)$ among all size-K_2 supports, whereas $\Phi^\star$ maximizes the oracle relevance mass $p^{\star,(d)}(S)$ over the same family. For any visual support $S \subseteq \mathcal{V}_d$, Eq. (47) gives

$$\begin{aligned} \left| a^{(d)}(S) - p^{\star,(d)}(S) \right| &= \left| \sum_{i \in S} \left(a_i^{(d)} - p_i^{\star,(d)} \right) \right| \\ &\leq \sum_{i \in S} \left| a_i^{(d)} - p_i^{\star,(d)} \right| \\ &\leq \|a^{(d)} - p^{\star,(d)}\|_1 \leq \delta_d. \end{aligned} \tag{126}$$

Eq. (126) implies

$$p^{\star,(d)}(S) \geq a^{(d)}(S) - \delta_d \tag{127}$$

and

$$a^{(d)}(S) \geq p^{\star,(d)}(S) - \delta_d. \tag{128}$$

The Top-K_2 optimality of Φ under $a^{(d)}$ gives

$$a^{(d)}(\Phi) \geq a^{(d)}(\Phi^\star). \tag{129}$$

Combining Eqs. (127)–(129) yields

$$\begin{aligned} p^{\star,(d)}(\Phi) &\geq a^{(d)}(\Phi) - \delta_d \\ &\geq a^{(d)}(\Phi^\star) - \delta_d \\ &\geq p^{\star,(d)}(\Phi^\star) - 2\delta_d. \end{aligned} \tag{130}$$

Taking complements and using Eq. (55) gives

$$\begin{aligned} p^{\star,(d)}(\Phi^c) &= 1 - p^{\star,(d)}(\Phi) \\ &\leq 1 - p^{\star,(d)}(\Phi^\star) + 2\delta_d \\ &= \epsilon_{21}(K_2, d) + 2\delta_d, \end{aligned} \tag{131}$$

which proves Eq. (70).

Eq. (126) applied to the complement Φ^c gives

$$\begin{aligned} a^{(d)}(\Phi^c) &\leq p^{\star,(d)}(\Phi^c) + \delta_d \\ &\leq \epsilon_{21}(K_2, d) + 3\delta_d. \end{aligned} \tag{132}$$

Eq. (132) proves Eq. (71). Define the discarded instruction-derived relevance mass by

$$\varepsilon_d = a^{(d)}(\Phi^c). \tag{133}$$

Eqs. (72) and (73) give

$$o^{(d)} - \widehat{o}^{(d)} = \sum_{i \notin \Phi} a_i^{(d)} u_i - \frac{\varepsilon_d}{1-\varepsilon_d} \sum_{i \in \Phi} a_i^{(d)} u_i. \tag{134}$$

Assumption 5 and the normalization $\sum_{i \in \mathcal{V}_d} a_i^{(d)} = 1$ imply

$$\begin{aligned} \|o^{(d)} - \widehat{o}^{(d)}\|_2 &\le \sum_{i \notin \Phi} a_i^{(d)} \|u_i\|_2 + \frac{\varepsilon_d}{1-\varepsilon_d} \sum_{i \in \Phi} a_i^{(d)} \|u_i\|_2 \\ &\le U\varepsilon_d + \frac{\varepsilon_d}{1-\varepsilon_d} U(1-\varepsilon_d) \\ &= 2U\varepsilon_d \\ &\le 2U\left[\epsilon_{21}(K_2, d) + 3\delta_d\right] = \mathcal{E}_2(K_2, d). \end{aligned} \tag{135}$$

Eq. (135) proves Eq. (74). Eqs. (48) and (75) make both terms in Eq. (56) nonincreasing when the pruning depth moves from d_0 to d. If at least one of the two inequalities is strict, then

$$\begin{aligned} \mathcal{E}_2(K_2, d) &= 2U\left[\epsilon_{21}(K_2, d) + 3\delta_d\right] \\ &< 2U\left[\epsilon_{21}(K_2, d_0) + 3\delta_{d_0}\right] = \mathcal{E}_2(K_2, d_0), \end{aligned} \tag{136}$$

which proves Eq. (76).

Proof of (V). The first-stage distortion bound $\mathcal{E}_1(K_1)$ applies before the decoder blocks act. Assumption 1 therefore gives the pre-pruning deviation at depth d:

$$\|X^{(d)} - \widetilde{X}^{(d)}\|_2 \le \Gamma_{1:d} \mathcal{E}_1(K_1), \tag{137}$$

where $\widetilde{X}^{(d)}$ denotes the hidden sequence obtained from the first-stage recast sequence before the second-stage pruning operation. Eq. (74) bounds the additional second-stage perturbation. The triangle inequality therefore gives

$$\|X^{(d)} - \widehat{X}^{(d)}\|_2 \le \Gamma_{1:d} \mathcal{E}_1(K_1) + \mathcal{E}_2(K_2, d). \tag{138}$$

Applying Assumption 1 to the remaining $L - d$ blocks gives

$$\begin{aligned} \|X^{(L)} - \widehat{X}^{(L)}\|_2 &\le \Gamma_{d+1:L}\left[\Gamma_{1:d}\mathcal{E}_1(K_1) + \mathcal{E}_2(K_2, d)\right] \\ &= \Gamma_{1:L}\mathcal{E}_1(K_1) + \Gamma_{d+1:L}\mathcal{E}_2(K_2, d). \end{aligned} \tag{139}$$

Assumption 1 and Eq. (37) then imply

$$\begin{aligned} \|z - \widehat{z}\|_\infty &= \|W(X_r^{(L)} - \widehat{X}_r^{(L)})\|_\infty \\ &\le \Lambda_0 \|X_r^{(L)} - \widehat{X}_r^{(L)}\|_2 \\ &\le \Delta(K_1, K_2, d), \end{aligned} \tag{140}$$

which proves Eq. (77).

For every competing label $y \neq y^\star$, Eq. (77) gives

$$\widehat{z}_{y^\star} \ge z_{y^\star} - \Delta(K_1, K_2, d), \tag{141}$$

$$\widehat{z}_y \le z_y + \Delta(K_1, K_2, d). \tag{142}$$

Subtracting Eq. (142) from Eq. (141) yields

$$\begin{aligned} \widehat{z}_{y^\star} - \widehat{z}_y &\ge z_{y^\star} - z_y - 2\Delta(K_1, K_2, d) \\ &\ge \mathfrak{m}(x, T) - 2\Delta(K_1, K_2, d). \end{aligned} \tag{143}$$

The condition in Eq. (80) makes the right side of Eq. (143) positive for every competing label, so the two-stage decoder preserves the unpruned argmax. The disagreement event therefore obeys

$$\{\widehat{y}_2 \neq \widehat{y}_0\} \subseteq \{\mathfrak{m}(x, T) \le 2\Delta(K_1, K_2, d)\}. \tag{144}$$

Taking probabilities under $\mathcal{D}$ proves Eq. (81). The task-error event satisfies

$$\{\widehat{y}_2 \neq y\} \subseteq \{\widehat{y}_0 \neq y\} \cup \{\widehat{y}_2 \neq \widehat{y}_0\}. \tag{145}$$

The union bound applied to Eq. (145) and Eq. (81) proves Eq. (82).

Proof of (VI). Substituting Eq. (84) into the layer-average budget gives

$$\begin{aligned} \theta K_1(u) + (1-\theta) K_2(u) = \theta\left[B + (1-\theta)u\right] \\ + (1-\theta)\left[B - \theta u\right] \end{aligned}$$

$$= B, \tag{146}$$

which proves Eq. (85). Differentiating Eq. (86) at fixed d gives

$$\begin{aligned}\frac{\partial J(u,d)}{\partial u} &= \Gamma_{1:L}\mathcal{E}_1'(K_1(u))\frac{dK_1(u)}{du} \\ &\quad + \Gamma_{d+1:L}\partial_1\mathcal{E}_2(K_2(u),d)\frac{dK_2(u)}{du} \\ &= (1-\theta)\Gamma_{1:L}\mathcal{E}_1'(K_1(u)) - \theta\Gamma_{d+1:L}\partial_1\mathcal{E}_2(K_2(u),d). \end{aligned} \tag{147}$$

At $u = 0$, the identities $K_1(0) = K_2(0) = B$ and the monotonicity assumptions give

$$\begin{aligned}\left.\frac{\partial J(u,d)}{\partial u}\right|_{u=0} &= -(1-\theta)\Gamma_{1:L}|\mathcal{E}_1'(B)| \\ &\quad + \theta\Gamma_{d+1:L}|\partial_1\mathcal{E}_2(B,d)|. \end{aligned} \tag{148}$$

Eq. (87) makes the right side of Eq. (148) strictly negative, which proves Eq. (88). Differentiability yields the local expansion

$$J(u,d) = J(0,d) + u\left.\frac{\partial J(u,d)}{\partial u}\right|_{u=0} + o(u). \tag{149}$$

The strict negativity in Eq. (88) implies that a sufficiently small $u_0 > 0$ exists for which Eq. (89) holds. The condition in Eq. (90) gives

$$J(0,d) = \Gamma_{1:L}\mathcal{E}_1(B) + \Gamma_{d+1:L}\mathcal{E}_2(B,d) \le J(0,d_0), \tag{150}$$

and Eqs. (89) and (150) prove Eq. (91).

Proof of (VII). Substituting Eq. (50) into Eqs. (92) and (93) gives

$$\Xi_0 = L\left[\kappa_1(Q+N)D^2 + \kappa_2(Q+N)^2D\right], \tag{151}$$

$$\begin{aligned}\Xi_2 &= \kappa_1 D^2\left[d(Q+K_1) + (L-d)(Q+K_2)\right] \\ &\quad + \kappa_2 D\left[d(Q+K_1)^2 + (L-d)(Q+K_2)^2\right] + \Xi_\circ. \end{aligned} \tag{152}$$

The budget identity in Eq. (33) implies

$$dK_1 + (L-d)K_2 = LB. \tag{153}$$

The linear compute difference therefore becomes

$$L(Q+N) - \left[d(Q+K_1) + (L-d)(Q+K_2)\right] = L(N-B). \tag{154}$$

Using $\theta = d/L$, the quadratic difference satisfies

$$\begin{aligned}&L(Q+N)^2 - d(Q+K_1)^2 - (L-d)(Q+K_2)^2 \\ &\quad = L\left[2Q(N-B) + N^2 - \theta K_1^2 - (1-\theta)K_2^2\right]. \end{aligned} \tag{155}$$

Multiplying Eqs. (154) and (155) by the corresponding coefficients in Eq. (50) and subtracting $\Xi_\circ$ proves Eq. (94). The condition in Eq. (95) makes $\Xi_0 - \Xi_2$ positive, which proves Eq. (96) and consequently Eq. (97).

When $K_2 = \frac{2}{3}K_1$, Eq. (33) gives

$$\begin{aligned}B &= \theta K_1 + (1-\theta)\frac{2}{3}K_1 \\ &= \frac{2+\theta}{3}K_1. \end{aligned} \tag{156}$$

Solving Eq. (156) gives Eq. (98). Substituting Eq. (98) into the depth-averaged quadratic visual workload gives

$$\begin{aligned}\theta K_1^2 + (1-\theta)K_2^2 &= \frac{B^2}{(2+\theta)^2}\left[9\theta + 4(1-\theta)\right] \\ &= B^2\frac{4+5\theta}{(2+\theta)^2}, \end{aligned} \tag{157}$$

which proves Eq. (99). Dividing N^2 by Eq. (99) proves Eq. (100).

Proof of (VIII). For a fixed depth d, the continuous relaxation treats K_1 and K_2 as positive real variables. The scalar $\lambda_B \ge 0$ denotes the Lagrange multiplier associated with the active

average-budget constraint. The Lagrangian is

$$\mathcal{L} = \Lambda_0 \Gamma_{1:L} \mathcal{E}_1(K_1) + \Lambda_0 \Gamma_{d+1:L} \mathcal{E}_2(K_2, d) + \lambda_B \left[\theta K_1 + (1-\theta) K_2 - B_0\right]. \tag{158}$$

Interior stationarity with respect to K_1 gives

$$\Lambda_0 \Gamma_{1:L} \mathcal{E}_1'(K_1) + \lambda_B \theta = 0. \tag{159}$$

Interior stationarity with respect to K_2 gives

$$\Lambda_0 \Gamma_{d+1:L} \partial_1 \mathcal{E}_2(K_2, d) + \lambda_B (1-\theta) = 0. \tag{160}$$

Rearranging Eqs. (159) and (160) yields

$$\lambda_B = -\frac{\Lambda_0 \Gamma_{1:L} \mathcal{E}_1'(K_1)}{\theta}, \tag{161}$$

$$\lambda_B = -\frac{\Lambda_0 \Gamma_{d+1:L} \partial_1 \mathcal{E}_2(K_2, d)}{1-\theta}. \tag{162}$$

Equating Eqs. (161) and (162) and cancelling the common factor Λ_0 proves Eq. (103). □

The theorem consolidates the complete theoretical argument into a single chain. The first stage protects spatially distributed evidence before cross-modal reasoning begins. The recasting step converts locally coherent discarded information into a second-order approximation error rather than a purely deletion-induced first-order loss. The instruction-level query reduces sensitivity to token-specific perturbations, and the delayed second stage removes visual tokens only after task-relevant semantics have become more concentrated. The final perturbation bound links these two approximation stages to the prediction margin, while the complexity analysis links the same token schedule to a strict reduction in decoder computation. The resulting framework explains how LCHTP can occupy a favorable accuracy–efficiency regime without requiring the early elimination of semantically unresolved visual evidence.

D Additional Experiment Results

D.1 Additional empirical findings

To further uncover the statistical regularities of this depth-wise evolution, we analyze local interaction, evidence coverage, and cross-layer consistency across a broad range of samples. Local spatial interaction is defined as the average fraction of outgoing self-attention that each patch assigns to its spatial neighborhood. For evidence coverage, we consider two complementary evidence sets: *detail evidence* comprising small-object and localized-text regions, and *main-entity evidence* corresponding to the largest object region occupying at least 10% of the image. At each layer l, we retain the Top-144 patches ranked by self-attention mass as T_l, and compute cumulative coverage over their union as $\mathrm{Cov}^{(l)}(E) = |E \cap \bigcup_{t \le l} T_t|/|E|$, where E denotes either evidence set. As shown in Fig. 9 left, local spatial interaction peaks at layer 6 and then declines sharply, whereas cumulative coverage steadily increases from 0.282 to 0.875 for detail evidence and from 0.199 to 0.748 for main-entity evidence. This opposing trend reveals a progressive shift from localized interaction toward broader evidence aggregation across depth. The right of Fig. 9 quantifies cross-layer consistency using the Jaccard similarity between Top-144 token sets. The similarity is only 0.13 between layers 2 and 22, but increases to 0.46 between layers 18 and 22, indicating substantial differences between shallow and deep selections and increasing consistency among deeper layers. Together, these statistics show that successive depths capture distinct yet complementary visual evidence.

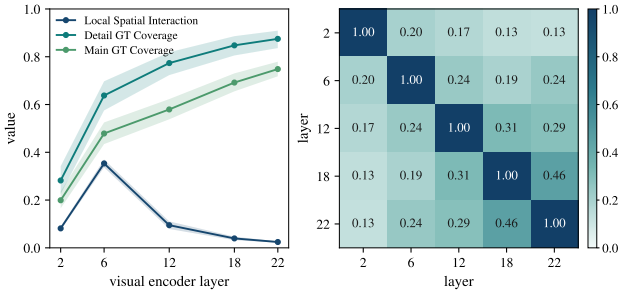

Figure 9: Left: Local spatial interaction and cumulative coverage of detail and main-entity evidence across layers. Right: Jaccard similarity between Top-144 salient token sets.

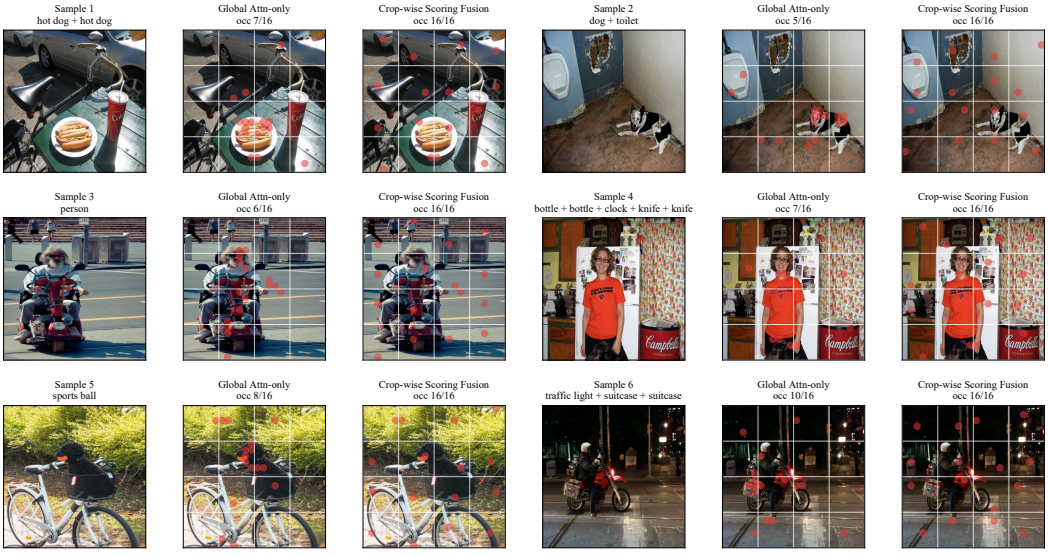

Figure 10: Global ranking collapses spatial coverage. We compare attention-only Top-K selection method against our scoring fusion strategy with a crop-wise quota, under $K{=}16$ out of 576 tokens. Across six samples, attention-only selection concentrates nearly all retained tokens on a single salient object, occupying only 5–10 of the 16 crops, whereas the crop-wise quota spreads tokens across all 16 crops, preserving multi-region relational evidence. Here, *occ* denotes the number of occupied crops, i.e., crops containing at least one retained token.

Global Ranking Collapses Spatial Coverage. To examine how global ranking affects spatial coverage, we global compare attention-only Top-K selection with crop-wise allocation under the same budget of $K = 16$ out of 576 visual tokens. We partition each image into 16 spatial crops and measure the number of crops containing at least one retained token. As shown in Fig. 2, global ranking concentrates tokens around a few salient regions, covering only $5-10$ crops across the six examples, whereas crop-wise allocation preserves all 16 regions under the same budget. For instance, attention-only selection in Samples 2 and 5 is largely concentrated around the dog and rider, respectively, while crop-wise allocation extends token coverage to surrounding objects and contextual regions, preserving a more complete representation of the scene. This contrast reveals that global competition allows dominant regions to absorb multiple token slots and suppress complementary evidence elsewhere. A lightweight crop-wise quota mitigates this spatial concentration and preserves distributed visual evidence, motivating region-aware allocation before global budget completion.

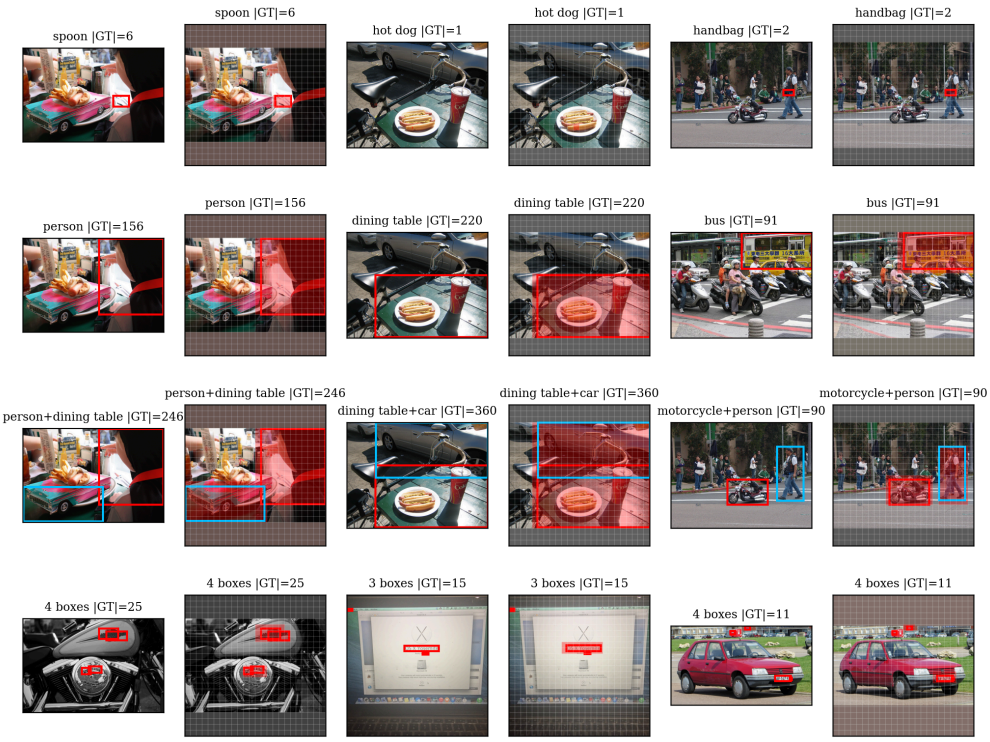

Figure 11: Visualization of ground-truth evidence regions. Examples include small-object, main-entity, relational, and text/OCR evidence, together with their corresponding mappings onto the visual patch grid.

Ground-truth region construction. We represent the critical visual evidence of each sample as a ground-truth patch set G on the 24×24 visual-token grid. To ensure exact spatial alignment with the model input, each source annotation undergoes the same preprocessing as the corresponding image. Specifically, the image is first padded to a square using its mean pixel color and then resized to 336×336. The annotation coordinates are transformed accordingly and projected onto the patch grid, where each cell corresponds to a 14×14 image region. For a transformed bounding box $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, the covered columns are $c_0 = \left\lfloor \frac{x_{\min}}{14} \right\rfloor$ and $c_1 = \left\lfloor \frac{x_{\max}-1}{14} \right\rfloor$, with r_0 and r_1 computed analogously for the rows. The subtraction of one pixel prevents a box ending exactly at a cell boundary from extending into the neighboring patch. After clamping all indices to the valid grid range, the ground-truth set is defined as $G = \{24r + c \mid r_0 \le r \le r_1, c_0 \le c \le c_1\}$.

When multiple evidence regions are present, their corresponding patch sets are merged to form G. This construction converts source-level annotations into discrete visual-token indices, allowing evidence coverage to be evaluated directly in the same representation space used for token selection. We construct G from different annotation sources according to the type of visual evidence. Small-object, main-entity, and relation samples are derived from COCO annotations. Small-object samples

retain a single qualifying small instance, while main-entity samples use the largest non-crowd instance. Relation samples contain the two largest qualifying non-crowd instances, whose regions are treated as the subject and object and jointly define the target evidence. For text samples, localized evidence regions are obtained from TextVQA images using EasyOCR. We retain recognized text boxes with confidence $c \geq 0.2$ and at least two decoded characters, and merge the patches covered by all retained boxes into G. Fig. 11 illustrates this annotation-to-patch mapping. Each example shows the source image with its evidence boxes alongside the preprocessed image with the 24×24 grid overlaid. The highlighted cells correspond exactly to the patches in G, and $|G|$ denotes the number of patches covered by the annotated evidence.

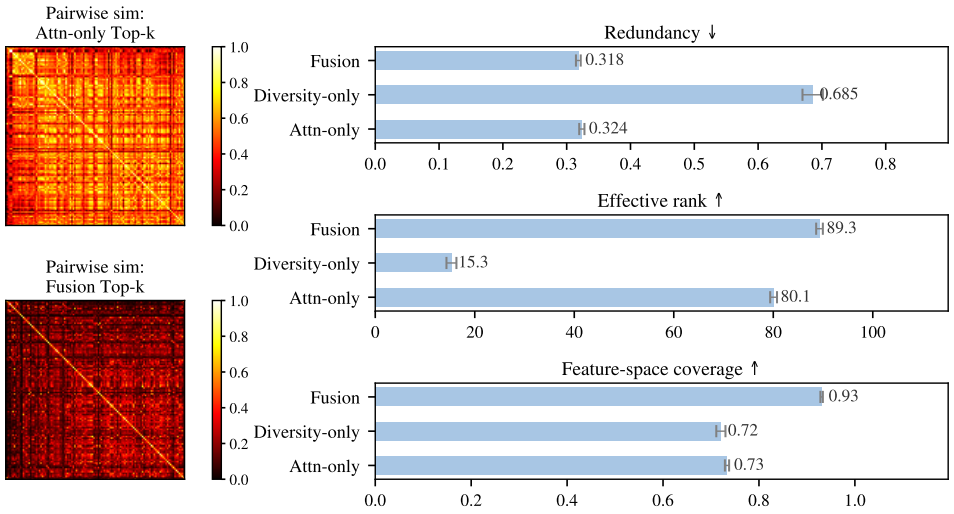

Figure 12: Left: Pairwise feature similarity of tokens selected by attention-only and fusion strategies. Right: Comparison of redundancy, effective rank, and feature-space coverage under the same token budget $K = 144$.

To examine whether attention relevance alone preserves sufficiently diverse visual information, we compare attention-only selection, diversity-only selection based on s_i^{div}, and joint relevance-diversity selection under the same token budget $K = 144$. For a retained set S with L_2-normalized projected features $\{v_i\}_{i \in S}$, we characterize its feature structure using three metrics. Feature redundancy R is defined as the mean pairwise cosine similarity $R = \frac{1}{K(K-1)} \sum_{\substack{i,j \in S \\ i \neq j}} \cos(v_i, v_j)$, where a lower value indicates less feature duplication. Effective rank E characterizes the dimensional richness of the retained features. Given the singular values $\{\sigma_r\}$ of V_S, we normalize them as $p_r = \sigma_r / \sum_j \sigma_j$ and compute $E = \exp(-\sum_r p_r \log p_r)$ where a larger value indicates that information spans more independent feature directions. Feature-space coverage C is defined as $C = \frac{1}{N} \sum_{i=1}^{N} \max_{j \in S} \cos(v_i, v_j)$, which quantifies how well the retained tokens represent the original visual feature space. As shown in Fig. 12, the pairwise-similarity maps on the left reveal stronger feature overlap under attention-only selection, whereas jointly considering relevance and diversity yields a less redundant feature structure. The three statistics in the right-hand bar plots quantify this difference from complementary perspectives. Compared with attention-only selection, joint relevance-diversity selection reduces R from 0.324 to 0.318, increases E from 80.1 to 89.3, and substantially raises C from 0.73 to 0.93, indicating reduced feature duplication, richer independent feature directions, and broader representation of the original feature space. These results show that relevance and diversity provide complementary cues for token selection, motivating their joint modeling to preserve informative yet non-redundant visual representations.

D.2 Additional Comparison Results

Tables 5 and 6 provide comprehensive comparisons on LLaVA-1.5-7B and Qwen2.5-VL-7B under progressively reduced visual-token budgets. LCHTP consistently achieves the highest average relative performance across the 192-, 128-, and 64-token settings on both models, with its advantage remaining pronounced under the most restrictive 64-token budget, where it retains 95.9% and 96.2%

of the unpruned performance, respectively. Tables 7–10 further provide a fine-grained analysis on MME across token budgets of 400, 320, 300, and 200. LCHTP achieves the highest total MME score among all token reduction methods at every evaluated budget, remaining close to the unpruned model at 400 tokens and preserving the strongest overall performance even when the budget is reduced to 200 tokens. The category-wise results further indicate that the preserved performance is distributed across perception, scene understanding, OCR-related tasks, and reasoning capabilities rather than being concentrated in a limited subset of tasks.

Table 5: Comparison with existing visual token reduction methods across multimodal benchmarks. Best results are highlighted in blue bold.

Methods	MMB	MMB_{CN}	POPE	MME	SQA	VQA_{Text}	VizWiz	GQA	VQAv2	Average
Upper Bound, 576 Tokens	64.4	58.3	85.9	1784	69.5	58.2	54.4	62.0	76.7	100.0%
LLaVA-1.5 7B	*Retain 192 Tokens* (↓ 66.7%)									
ToMe	60.5	–	72.4	1563	65.2	52.1	–	54.3	68.0	89.3%
FastV	63.2	57.3	84.7	1757	65.7	52.5	54.9	58.6	74.5	96.8%
LLaVA-PruMerge	59.6	52.9	71.3	1632	**66.3**	54.3	50.1	54.3	70.6	90.9%
HiRED	62.8	57.0	85.3	1742	65.2	47.4	54.2	59.3	75.3	95.7%
SparseVLM	64.1	58.0	85.6	1801	64.7	56.1	54.3	**60.3**	**75.9**	98.3%
PDrop	63.2	56.8	82.3	1766	64.2	56.1	51.1	57.1	75.1	95.9%
MustDrop	61.7	55.9	83.8	1719	65.2	56.5	54.5	57.1	73.1	96.0%
FiCoCo-V	62.3	55.3	82.5	1732	65.9	55.7	51.0	58.5	74.4	95.6%
FasterVLM	62.8	56.8	85.3	1744	65.2	57.3	54.2	59.3	75.3	97.5%
VisionZip	62.0	57.7	85.2	1751	65.4	57.3	54.2	59.3	75.0	97.6%
HoloV	62.6	57.0	85.3	1732	65.1	57.4	54.2	59.3	75.3	97.5%
LCHTP **(Ours)**	**64.4**	**59.2**	**86.2**	**1804**	66.0	**57.5**	**55.7**	59.8	75.1	**99.3%**
LLaVA-1.5 7B	*Retain 128 Tokens* (↓ 77.8%)									
ToMe	53.3	–	62.8	1343	59.6	49.1	–	52.4	63.0	81.1%
FastV	63.4	57.2	83.1	1720	65.5	50.6	55.7	56.9	72.7	95.5%
LLaVA-PruMerge	58.1	51.7	67.2	1554	66.0	54.3	50.3	53.3	68.8	89.0%
HiRED	62.1	57.0	84.9	1729	65.2	46.1	54.5	57.8	73.9	94.8%
SparseVLM	63.5	57.7	84.9	1782	64.8	54.9	54.3	58.1	**74.1**	97.1%
PDrop	61.1	56.6	82.3	1644	64.2	55.1	51.0	56.0	72.9	94.0%
MustDrop	61.5	55.1	83.2	1697	65.6	56.3	54.7	56.3	72.5	95.4%
FiCoCo-V	61.1	54.3	82.2	1711	65.4	55.6	49.4	57.6	73.1	94.3%
FasterVLM	62.1	57.2	84.8	1718	65.1	56.6	54.5	57.8	73.9	96.7%
VisionZip	61.3	56.4	84.8	1748	65.0	56.8	54.3	57.7	73.8	96.5%
HoloV	62.0	57.3	84.8	1745	65.1	56.8	54.7	57.8	74.0	97.0%
LCHTP **(Ours)**	**64.2**	**57.9**	**86.0**	**1796**	**66.2**	**57.1**	**56.4**	**58.9**	74.0	**98.7%**
LLaVA-1.5 7B	*Retain 64 Tokens* (↓ 88.9%)									
ToMe	43.7	–	52.5	1138	50.0	45.3	–	48.6	57.1	70.8%
FastV	60.8	53.8	79.0	1552	65.8	47.8	54.9	53.7	68.3	91.0%
LLaVA-PruMerge	55.3	49.1	65.3	1549	65.5	54.0	50.1	51.9	67.4	87.1%
HiRED	60.5	54.9	82.3	1657	65.4	44.2	55.9	55.0	70.6	92.3%
SparseVLM	60.0	50.9	76.3	1601	65.6	51.8	54.3	52.5	66.3	90.4%
PDrop	33.3	50.5	55.9	1092	64.2	45.9	50.7	41.9	69.2	76.3%
MustDrop	58.2	53.5	79.4	1583	66.3	54.2	54.0	52.7	67.1	91.5%
FiCoCo-V	60.3	53.0	76.0	1591	64.8	53.6	49.8	52.4	**71.3**	90.7%
FasterVLM	60.4	55.1	82.3	1657	65.6	55.3	56.0	55.1	70.6	94.5%
VisionZip	59.6	55.2	82.5	1636	65.8	**55.5**	54.7	55.2	70.6	94.1%
HoloV	60.5	55.2	82.2	1661	65.7	55.4	55.8	55.1	70.7	94.5%
LCHTP **(Ours)**	**61.0**	**56.0**	**84.1**	**1700**	**66.4**	55.4	**57.3**	**56.2**	71.1	**95.9%**

Table 6: Comparison with existing visual token reduction methods across multimodal benchmarks. Best results are highlighted in blue bold.

Methods	MMB	MMB_{CN}	POPE	MME	SQA	VQA_{Text}	VizWiz	GQA	VQAv2	Average
Upper Bound, 576 Tokens	79.9	80.5	80.4	2185	86.0	76.9	67.4	58.4	79.5	100%
Qwen2.5-VL 7B	*Retain 192 Tokens* (↓ 66.7%)									
FastV	77.9	78.2	76.2	1998	82.6	75.7	66.6	54.3	75.6	95.8%
HiRED	79.1	79.7	77.6	1979	85.8	72.8	66.0	56.2	77.1	96.7%
SparseVLM	78.9	79.1	77.3	2081	85.3	75.3	66.3	54.8	76.3	97.1%
MustDrop	79.7	79.5	**80.3**	2159	**86.3**	75.5	66.3	57.0	78.2	98.9%
FasterVLM	79.4	79.4	79.6	1965	85.5	71.1	65.6	56.3	76.9	96.6%
VisionZip	79.2	79.8	77.1	1951	85.3	72.8	66.2	56.2	77.2	96.5%
HoloV	79.8	80.1	77.6	1958	85.8	72.8	66.0	56.4	77.1	96.8%
LCHTP **(Ours)**	**80.4**	**80.6**	80.2	**2163**	86.1	**76.4**	**66.6**	**57.6**	**78.4**	**99.5%**
Qwen2.5-VL 7B	*Retain 128 Tokens* (↓ 77.8%)									
FastV	77.5	77.0	72.2	1868	81.4	73.3	65.9	51.0	72.3	92.7%
HiRED	78.7	77.7	74.6	1842	82.9	68.9	66.1	54.5	75.2	93.8%
SparseVLM	77.8	78.6	72.2	1945	82.9	73.1	65.7	51.3	72.9	93.6%
MustDrop	79.4	79.6	78.9	2089	85.6	74.1	66.4	55.6	77.2	97.7%
FasterVLM	78.3	77.8	76.0	1841	84.4	63.8	65.2	54.1	74.4	93.1%
VisionZip	78.3	78.2	74.0	1811	83.3	68.2	66.0	54.9	75.3	93.6%
HoloV	78.5	77.4	75.2	1797	82.4	69.4	65.8	54.4	75.1	93.5%
LCHTP **(Ours)**	**80.3**	**80.5**	**79.2**	**2173**	**85.8**	**74.7**	**66.7**	**57.8**	**77.5**	**99.0%**
Qwen2.5-VL 7B	*Retain 64 Tokens* (↓ 88.9%)									
FastV	72.9	73.4	58.4	1548	77.8	66.8	63.8	45.8	64.3	84.1%
HiRED	75.3	75.7	66.6	1641	81.3	58.6	64.7	50.0	69.9	87.4%
SparseVLM	75.2	75.8	67.6	1769	81.2	69.8	64.9	49.0	69.0	89.5%
MustDrop	78.1	78.5	76.2	1953	85.3	69.4	65.3	52.8	74.5	94.4%
FasterVLM	74.7	74.5	68.0	1632	79.3	52.8	62.6	50.3	69.1	85.8%
VisionZip	74.8	75.0	65.8	1598	82.0	57.9	64.2	50.2	70.1	86.9%
HoloV	75.1	75.2	66.1	1617	81.1	59.5	64.2	49.7	69.6	87.0%
LCHTP **(Ours)**	**78.2**	**79.0**	**77.1**	**2027**	**85.8**	**70.7**	**66.7**	**56.5**	**74.6**	**96.2%**

Table 7: Fine-grained MME results under a visual token budget of 400.

Task	Base	FastV	HiRED	SparseVLM	MustDrop	FasterVLM	VisionZip	HoloV	Ours
Existence	195.00	180.00	195.00	185.00	185.00	195.00	195.00	190.00	190.00
Count	128.33	120.00	103.33	130.00	96.67	103.33	106.67	115.00	108.33
Position	151.67	146.67	121.67	141.67	91.67	123.33	115.00	123.33	121.67
Color	165.00	163.33	150.00	160.00	150.00	150.00	150.00	150.00	160.00
Posters	161.22	157.48	160.88	156.12	156.80	155.78	154.08	157.82	157.82
Celebrity	146.47	143.82	143.53	144.41	139.71	140.00	136.47	142.65	151.18
Scene	161.75	160.25	159.50	159.00	156.50	160.50	162.75	161.00	163.75
Landmark	144.50	138.00	153.50	141.25	147.50	152.25	155.00	151.25	150.25
Artwork	82.75	87.75	96.00	81.50	87.50	95.50	97.75	96.50	94.00
OCR	140.00	125.00	145.00	147.50	147.50	140.00	132.50	140.00	132.50
Commonsense Reasoning	72.14	57.86	78.57	59.29	79.29	81.43	82.14	82.14	85.00
Numerical Calculation	50.00	37.50	50.00	50.00	50.00	52.50	57.50	52.50	50.00
Text Translation	37.50	32.50	47.50	37.50	67.50	52.50	62.50	52.50	62.50
Code Reasoning	52.50	52.50	50.00	62.50	55.00	47.50	47.50	47.50	57.50
Total	**1688.84**	**1602.66**	**1654.49**	**1655.74**	**1610.63**	**1649.63**	**1654.86**	**1662.20**	**1684.50**

Table 8: Fine-grained MME results under a visual token budget of 320.

Task	Base	FastV	HiRED	SparseVLM	MustDrop	FasterVLM	VisionZip	HoloV	Ours
Existence	195.00	180.00	190.00	170.00	185.00	195.00	190.00	195.00	185.00
Count	128.33	108.33	88.33	120.00	121.67	101.67	95.00	95.00	96.67
Position	151.67	143.33	115.00	138.33	120.00	120.00	111.67	115.00	125.00
Color	165.00	158.33	155.00	140.00	160.00	155.00	155.00	160.00	160.00
Posters	161.22	153.06	158.16	146.94	150.00	156.80	156.12	157.82	156.46
Celebrity	146.47	145.00	142.94	143.53	131.47	137.65	132.06	137.94	145.88
Scene	161.75	158.50	166.00	160.25	162.75	155.25	161.00	155.25	166.00
Landmark	144.50	133.00	146.75	131.25	137.25	150.75	150.25	153.00	148.25
Artwork	82.75	84.75	98.50	72.00	80.50	97.00	99.00	97.00	90.25
OCR	140.00	140.00	110.00	140.00	132.50	147.50	145.00	147.50	147.50
Commonsense Reasoning	72.14	57.14	76.43	55.00	62.14	81.43	80.71	81.43	81.43
Numerical Calculation	50.00	47.50	60.00	50.00	47.50	52.50	55.00	52.50	50.00
Text Translation	37.50	32.50	62.50	40.00	62.50	55.00	65.00	57.50	67.50
Code Reasoning	52.50	57.50	57.50	50.00	45.00	45.00	55.00	52.50	55.00
Total	**1688.84**	**1598.95**	**1627.12**	**1557.30**	**1598.28**	**1650.55**	**1650.81**	**1657.44**	**1674.94**

Table 9: Fine-grained MME results under a visual token budget of 300.

Task	Base	FastV	HiRED	SparseVLM	MustDrop	FasterVLM	VisionZip	HoloV	Ours
Existence	195.00	180.00	195.00	165.00	185.00	195.00	185.00	195.00	185.00
Count	128.33	113.33	90.00	118.33	108.33	98.33	106.67	98.33	95.00
Position	151.67	143.33	106.67	120.00	111.67	113.33	108.33	108.33	121.67
Color	165.00	150.00	155.00	140.00	160.00	160.00	155.00	160.00	165.00
Posters	161.22	153.06	157.14	144.56	147.62	153.40	157.48	157.82	156.80
Celebrity	146.47	145.00	132.06	139.71	130.88	132.94	128.53	135.00	147.94
Scene	161.75	158.50	159.00	157.25	160.75	153.00	161.25	153.00	166.75
Landmark	144.50	132.00	154.25	129.75	137.25	152.75	151.75	153.50	143.50
Artwork	82.75	85.50	95.75	70.75	81.25	96.50	99.00	96.50	87.25
OCR	140.00	140.00	140.00	132.50	140.00	140.00	137.50	147.50	147.50
Commonsense Reasoning	72.14	53.57	85.71	54.29	60.71	83.57	77.86	83.57	81.43
Numerical Calculation	50.00	40.00	40.00	50.00	47.50	52.50	50.00	52.50	57.50
Text Translation	37.50	35.00	62.50	42.50	70.00	57.50	57.50	57.50	65.00
Code Reasoning	52.50	60.00	52.50	52.50	50.00	57.50	52.50	57.50	52.50
Total	**1688.84**	**1589.30**	**1625.58**	**1517.13**	**1590.97**	**1646.33**	**1628.37**	**1656.06**	**1672.84**

Table 10: Fine-grained MME results under a visual token budget of 200.

Task	Base	FastV	HiRED	SparseVLM	MustDrop	FasterVLM	VisionZip	HoloV	Ours
Existence	195.00	180.00	185.00	145.00	175.00	175.00	175.00	175.00	175.00
Count	128.33	91.67	81.67	103.33	81.67	80.00	81.67	76.67	71.67
Position	151.67	123.33	95.00	113.33	100.00	91.67	86.67	91.67	90.00
Color	165.00	120.00	143.33	135.00	138.33	155.00	160.00	155.00	165.00
Posters	161.22	147.62	146.94	130.95	143.88	151.70	151.36	150.68	152.04
Celebrity	146.47	132.94	120.88	125.59	106.18	115.59	109.12	113.24	142.65
Scene	161.75	157.75	158.50	156.75	158.75	156.00	157.50	157.50	166.50
Landmark	144.50	123.25	144.25	113.50	128.50	146.75	152.00	145.50	139.25
Artwork	82.75	78.75	98.00	74.50	72.25	90.75	87.75	92.25	82.25
OCR	140.00	132.50	132.50	132.50	140.00	132.50	130.00	132.50	140.00
Commonsense Reasoning	72.14	47.86	88.57	57.86	62.86	85.00	77.86	87.14	77.86
Numerical Calculation	50.00	40.00	47.50	52.50	57.50	52.50	55.00	52.50	50.00
Text Translation	37.50	35.00	62.50	55.00	77.50	72.50	50.00	72.50	72.50
Code Reasoning	52.50	62.50	65.00	60.00	67.50	62.50	72.50	55.00	47.50
Total	**1688.84**	**1473.17**	**1569.64**	**1455.81**	**1509.91**	**1567.46**	**1546.42**	**1557.14**	**1572.21**

D.3 Parameter Sensitivity Analysis

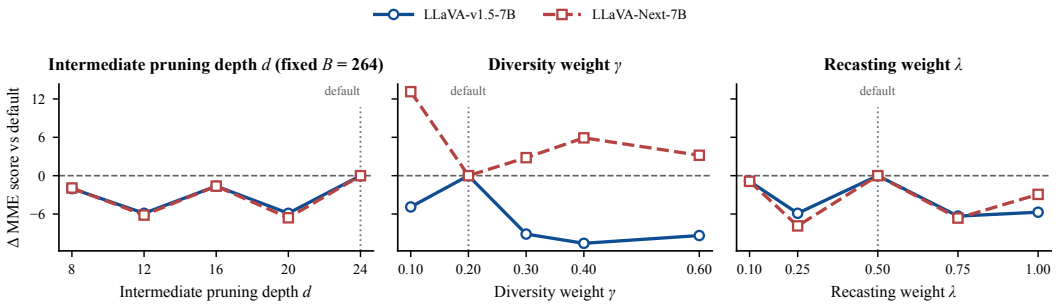

Figure 13: Hyperparameter sensitivity around the default configuration.

We evaluate the sensitivity of LCHTP on the full MME test set using LLaVA-1.5-7B and LLaVA-Next-7B, where only one hyperparameter is varied while all remaining settings are fixed. We vary the intermediate pruning depth over $d \in \{8, 12, 16, 20, 24\}$, the diversity weight over $\gamma \in \{0.10, 0.20, 0.30, 0.40, 0.60\}$, and the recasting weight over $\lambda \in \{0.10, 0.25, 0.50, 0.75, 1.00\}$. When varying γ, we keep α and β fixed at 0.4 and normalize the three importance weights by $\alpha + \beta + \gamma$ to obtain their effective contributions to the fused score. During the depth sweep, we fix $K_1 = 288$ and maintain a layer-average budget of approximately $B = 264$ by adjusting K_2 accordingly, so that the comparison primarily reflects pruning timing and stage-wise token allocation rather than changes in the overall token budget. As shown in Fig. 13, the intermediate pruning depth exhibits the smallest variation, with both models attaining their best performance at $d = 24$. The diversity weight is more model-dependent, with $\gamma = 0.20$ performing best on LLaVA-1.5-7B while

$\gamma = 0.10$ gives the highest score on LLaVA-Next-7B, indicating that the preferred diversity contribution may vary across architectures. In contrast, both models achieve their strongest performance with the recasting weight set to $\lambda = 0.50$. Despite these local fluctuations, the overall variation remains small, with the maximum relative score range below 0.814%, indicating that LCHTP is generally robust around the default configuration and supporting $d = 24$, $\gamma = 0.20$, and $\lambda = 0.50$ as a stable reference setting. Thus, for all model architectures, this configuration can serve as a strong reference point, around which further model-specific tuning may yield additional performance gains. However, for models with substantially different decoder depths, the intermediate pruning depth should be reselected empirically within the middle-to-late layers to identify a more suitable operating point. For the first-stage score computation, all three importance signals are extracted from the vision encoder. Specifically, the local relevance score s_i^{loc} is computed from the shallow-layer attention at ℓ_1, corresponding to the 6th visual layer, while the global relevance score s_i^{glob} is obtained from the penultimate layer ℓ_2. The diversity score s_i^{div} is computed from the patch features of the same penultimate visual layer ℓ_2 after the multimodal projection. This configuration allows the three signals to complementarily capture shallow local interactions, deep global semantics, and feature-space diversity, while remaining independent of the intermediate decoder pruning depth d.

D.4 Additional Token Pruning Visualizations

Figs. 14–16 present additional qualitative comparisons across multiple image samples under pruning ratios of 87.5%, 70.0%, and 50.0%, with FastV, SparseVLM, and LCHTP evaluated under the same token budget. Across all three pruning levels, LCHTP consistently retains a larger fraction of the highlighted regions of interest, and the advantage becomes particularly evident as the token budget becomes more restrictive. The visual comparisons further show that the retained tokens of LCHTP are distributed more effectively over semantically relevant object regions, providing qualitative evidence that its coverage-preserving selection better maintains spatially dispersed visual evidence rather than concentrating the limited token budget on only a few salient locations.

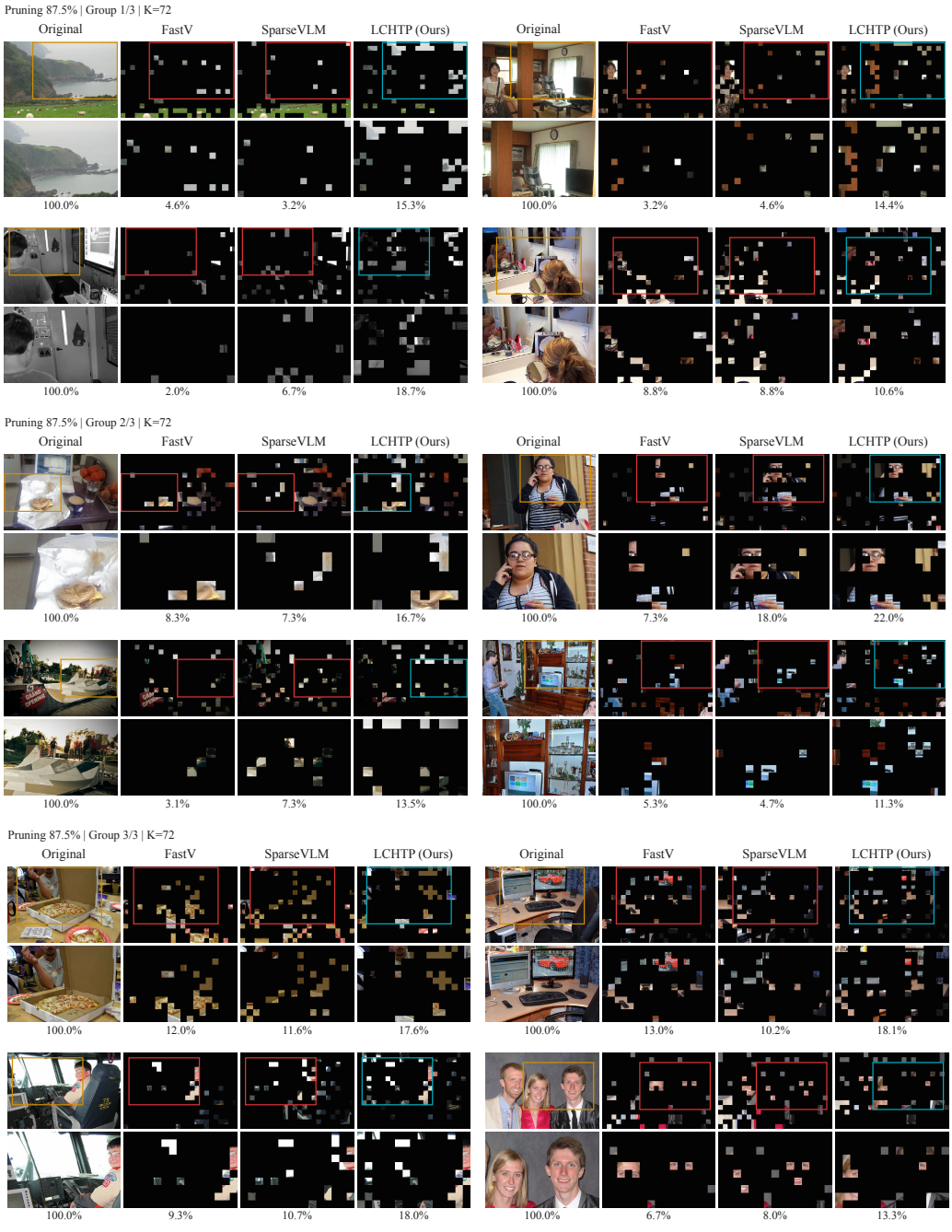

Figure 14: Additional qualitative comparisons at an 87.5% pruning ratio ($K = 72$). The three plates correspond to Groups 1–3. LCHTP consistently preserves a larger fraction of the highlighted ROI than FastV and SparseVLM at the same token budget.



Figure 15: Additional qualitative comparisons at a 70.0% pruning ratio ($K = 173$). The three plates correspond to Groups 1–3. Compared with the baselines, LCHTP retains more of the highlighted object regions while keeping the same number of visual tokens.

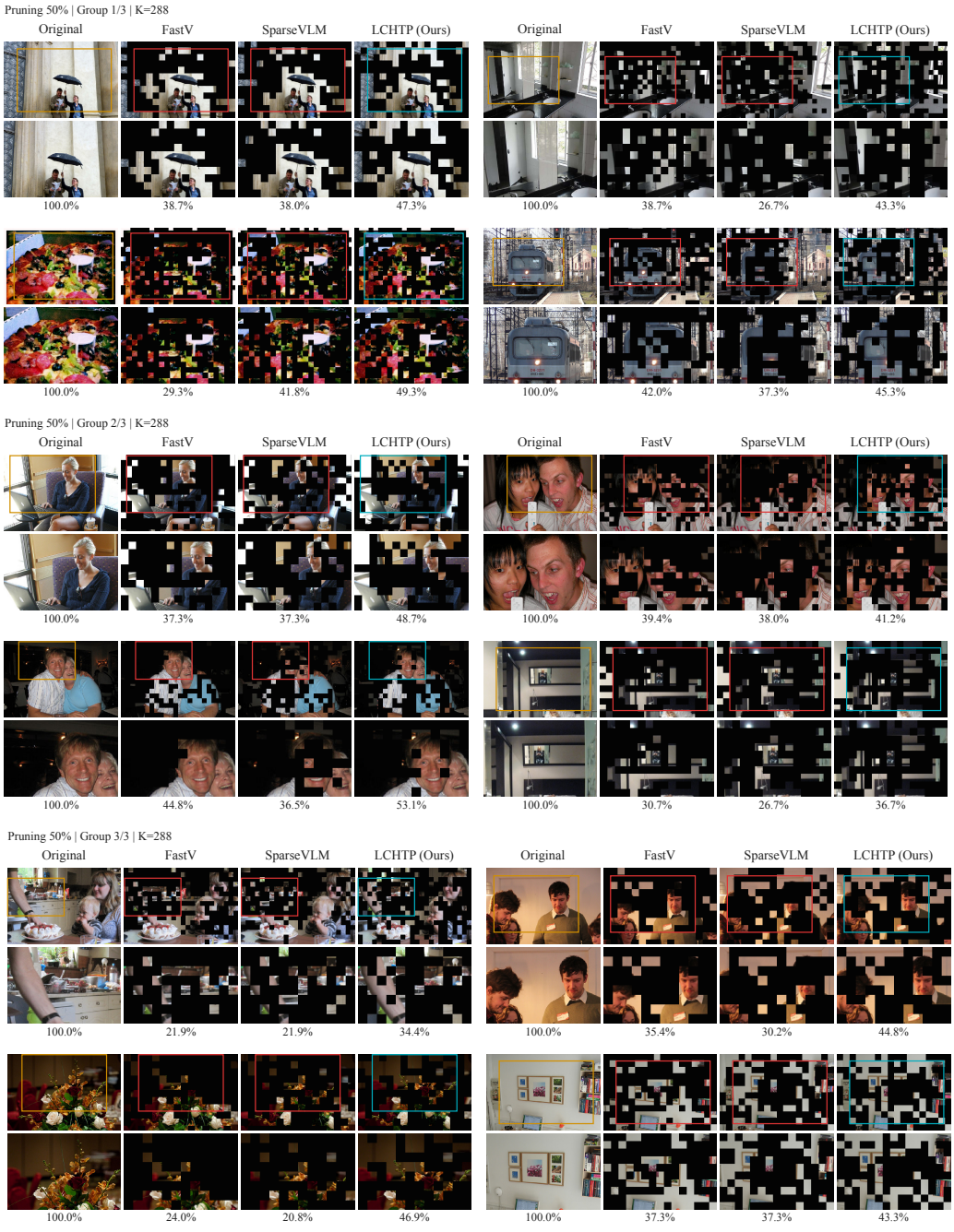

Figure 16: Additional qualitative comparisons at a 50.0% pruning ratio ($K = 288$). The three plates correspond to Groups 1–3. The advantage of LCHTP remains visible at the less aggressive pruning setting, with higher ROI coverage across all shown examples.