HIPPO: Accelerating Video Large Language Models Inference via Holistic-aware Parallel Speculative Decoding

Anonymous ACL submission

Abstract

Speculative decoding (SD) has emerged as a promising approach to accelerate LLM inference without sacrificing output quality. Existing SD methods tailored for video-LLMs primarily focus on pruning redundant visual tokens to mitigate the computational burden of massive visual inputs. However, existing methods do not achieve inference acceleration comparable to text-only LLMs. We observe from extensive experiments that this phenomenon mainly stems from two limitations: (i) their pruning strategies inadequately preserve visual semantic tokens, degrading draft quality and acceptance rates; (ii) even with aggressive pruning (e.g., 90% visual tokens removed), the draft model's remaining inference cost limits overall speedup. To address these limitations, we propose HIPPO, a general **h**olist**ic**-aware **p**arallel s**p**eculative dec**o**ding framework. Specifically, HIPPO proposes (i) a semantic-aware token preservation method, which fuses global attention scores with local visual semantics to retain semantic information at high pruning ratios; (ii) a video parallel SD algorithm that decouples and overlaps draft generation and target verification phases. Experiments on four video-LLMs across six benchmarks demonstrate HIPPO's effectiveness, yielding up to $3.51\times$ speedup compared to vanilla autoregressive decoding.

1 Introduction

Recent advances in video Large Language Models (video-LLMs) (Tang et al., 2025; Bai et al., 2025b,a) extend large language models to the video domain, achieving strong performance on video question answering (Fu et al., 2025) and captioning (Chai et al., 2025) tasks. However, the autoregressive, token-by-token generation process imposes significant inference latency. This challenge is exacerbated in the video domain by the sheer volume of input tokens, which leads to substantial computational and memory overhead during attention calculation (Ji et al., 2025).

To this end, extensive research has explored token pruning strategies to accelerate video-LLM inference. These methods exploit spatiotemporal redundancy in video tokens by pruning less important tokens based on attention scores, thereby reducing computational and memory costs during inference (Huang et al., 2025; Zou et al., 2025; Ji et al., 2025). However, by removing tokens from the input, they inevitably incur information loss. This is particularly critical in high-resolution, high-frame-rate long video understanding tasks, where the disruption of spatiotemporal continuity can lead to biased video comprehension (Fu et al., 2025).

To mitigate this information loss, *Speculative Decoding* (SD) (Leviathan et al., 2023) has emerged as a lossless acceleration approach, where a lightweight draft model proposes multiple candidate tokens that are then verified in parallel by the target model. To adapt SD for the video domain, existing methods mainly prune the draft model's visual tokens using attention scores from the target model. These methods preserve lossless decoding, as the target model receives the full input for verification (Ji et al., 2025; Zhang et al., 2025b).

Despite their benefits, existing video SD methods suffer from significant limitations that prevent them from being as effective as their text-only counterparts. We identify two key limitations: **(i)** attention-based token selection suffers from position bias (Xiao et al., 2024; Gu et al., 2025), where frame boundary tokens receive inflated attention weights regardless of their semantic relevance. As shown in Figure 1 (top), under 90% pruning, this leads to retaining non-informative boundary tokens (please refer to Section 4 for statistical results). **(ii)** even with aggressive pruning (e.g., 90%), the draft model still incurs substantial computational overhead. For instance, LLaVA-OneVision (Li et al., 2025a) encodes each frame into 196 tokens,

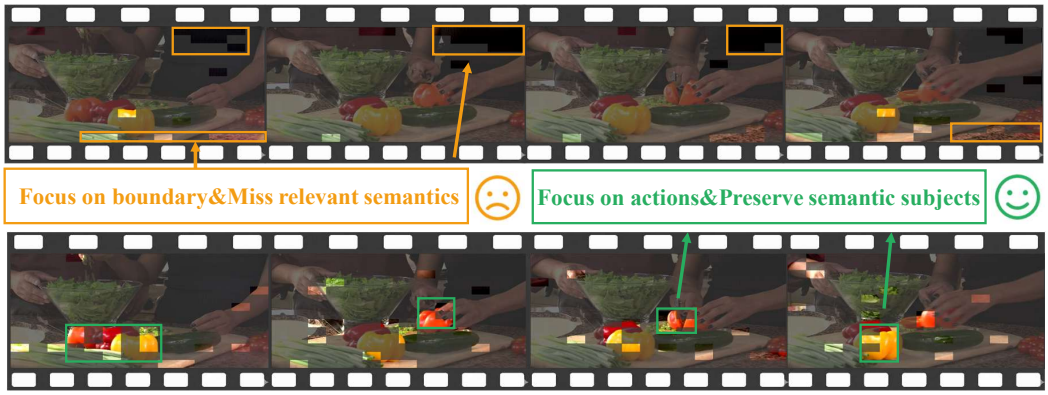


Figure 1: Visualization of video frames under 90% pruning ratio. (i) The upper part attention-only pruning suffers from the position bias phenomenon, where inflated attention weights are allocated to non-informative boundary tokens **(highlighted by orange boxes)**. (ii) The lower part semantic-aware token preservation preserves informative visual tokens **(highlighted by green boxes)**, retaining the core action despite the high pruning ratio.

yielding over 1M tokens for a two-minute video at 60 FPS. After 90% pruning, the draft model still processes over 100K tokens, resulting in non-negligible inference latency that limits the speedup of serial SD.

To address these limitations, we propose a general framework, namely **h**olist**i**c-aware **p**arallel s**p**eculative dec**o**ding (HIPPO). Specifically, HIPPO addresses the two limitations and proposes two key designs: **(i)** a semantic-aware token preservation method that integrates **global attention scores** with **local visual semantics** to adaptively allocate the pruning budget across spatial crops. As shown in the lower part of Figure 1, this method preserves semantically informative visual tokens even at high pruning ratios (e.g., 90%) by prioritizing tokens that are salient under both global and local criteria. This global-local coordination prevents semantic information loss and ensures that the retained tokens encode coherent scene-level context rather than merely isolated salient attention regions. **(ii)** a video parallel SD algorithm that enables synchronous execution of the draft and target models, effectively overlapping the draft model's inference cost during the verification phase. While the target model verifies candidate tokens from the previous round, the draft model concurrently generates the next round of candidate tokens. This parallel execution allows both models' computations to overlap, thereby mitigating the limited speedup issue that arises when the draft process accounts for a significant portion of the total latency.

We summarize our contributions as follows:

(i) **Semantic-aware video token preservation.** HIPPO integrates global attention scores with local visual semantics to adaptively allocate pruning budgets, enabling effective preservation of semantically informative tokens at high pruning ratios while mitigating position bias.

(ii) **Video parallel speculative decoding.** HIPPO proposes a parallel execution framework where the draft model generates candidates concurrently while the target model performs verification, effectively overlapping their computation to hide the draft overhead.

(iii) **Significant and versatile speedups.** Experiments on four video-LLMs across six benchmarks show that HIPPO achieves up to $3.51\times$ speedup over auto-regressive decoding, demonstrating its effectiveness and versatility.

2 Related Work

2.1 Speculative Decoding for LLMs

Speculative decoding has emerged as an effective approach to accelerate LLM inference without compromising generation quality. Existing works focus on improving draft model efficiency and increasing acceptance rates (Miao et al., 2023; Sun et al., 2023; Cai et al., 2024). Medusa (Cai et al., 2024) adds extra decoding heads at the top of the target model to generate drafts. Lookahead (Fu et al., 2024) caches the generation trajectory (n-grams) as the drafts. PEARL (Liu et al., 2025) parallelizes the execution of the draft and target models to achieve superior acceleration. Eagle (Li et al., 2025b) balances efficiency and acceptance rates by training a single-layer transformer as the draft model.

In the video domain, existing methods focus on exploiting the redundancy of visual tokens. They use signals from the target model to prune the draft model's visual input, thereby improving drafting efficiency (Ji et al., 2025; Zhang et al., 2025b). Sparse-to-Dense (Zhang et al., 2025b) uses a sparse module to generate tokens with top-K attention in the draft phase, which are then verified in parallel using full attention. SpecVLM (Ji et al., 2025) selects tokens guided by target model attention and applies uniform pruning across all frames. However, existing video SD methods suffer from position bias in the target model's attention, limiting their effectiveness. HIPPO addresses this limitation through semantic-aware token preservation. We also discuss video-LLMs and long-context SD in the text domain, explaining why the latter are not directly applicable to video-LLMs in Appendix A.

2.2 Visual Token Reductions

Visual inputs exhibit significant spatiotemporal redundancy, incurring high computational and memory costs (Zou et al., 2025). To mitigate this, extensive research focuses on token compression and pruning in vision transformers (Dosovitskiy et al., 2021). In particular, video token reduction has received increasing attention due to the massive volume of tokens generated by video inputs. For instance, FastVID (Shen et al., 2025) partitions frames into segments and applies density-based token pruning within each segment. DyCoke (Tao et al., 2025) performs token merging across frames and reduces KV cache dynamically. Unlike single-model pruning approaches, HIPPO leverages signals from the target model to guide pruning for the draft model. Therefore, HIPPO is orthogonal to single-model pruning methods and can be combined with them for specific downstream tasks.

3 Preliminaries

3.1 Notations

Let F denote the number of video frames. Each token $\mathbf{v}_j \in \mathbb{R}^d$ is a visual embedding of dimension d. We use $\mathcal{M}_t$ and $\mathcal{M}_d$ to denote the target and draft models, respectively. During speculative decoding, the draft model generates γ candidate tokens with logits q_i, which the target model verifies using logits p_i. We define three scoring functions: s_{attn} for global attention-based relevance, s_{temp} for inter-frame temporal redundancy, and s_{spa} for intra-frame spatial redundancy, which are aggregated to produce the final score $s(\mathbf{v}_j)$.

3.2 Speculative Decoding

Let $\boldsymbol{x}$ denote an input sequence (prefix). A speculative decoding (SD) step consists of a drafting phase followed by a verification phase. In the drafting phase, the draft model $\mathcal{M}_d$ is invoked γ times to generate draft tokens $x_1, x_2, \ldots, x_\gamma$ by forward passes and sampling. We denote the output logit $\mathcal{M}_d\big(\boldsymbol{x} + [x_1, \ldots, x_i]\big)$ as q_i, and sample each draft token as $x_i \sim q_i$ for $i = 1, \ldots, \gamma$. In the verification phase, the prefix $\boldsymbol{x}$ and the γ draft tokens are sent to the target model $\mathcal{M}_t$ for verification. The model $\mathcal{M}_t$ takes $\boldsymbol{x} + [x_1, \ldots, x_\gamma]$ as input and outputs logits $p_0, p_1, \ldots, p_\gamma$. SD then sequentially verifies each x_i via speculative sampling (Leviathan et al., 2023), with acceptance probability

$$\alpha_i = \begin{cases} 1, & \text{if } p_i[x_i] \geq q_i[x_i], \\ \dfrac{p_i[x_i]}{q_i[x_i]}, & \text{if } p_i[x_i] < q_i[x_i]. \end{cases} \tag{1}$$

If SD rejects x_i, it resamples a token from $\text{norm}\big(\max(0, p_i - q_i)\big)$; otherwise, SD accepts all draft tokens and samples one additional token from p_γ. In this way, each SD yields theoretical lossless quality with improved decoding efficiency.

4 Motivated Experiments

4.1 Position Bias Issue of Visual Tokens

We analyze the attention score distribution of video-SALMONN2+ 72B (Tang et al., 2025) on the Video-MME benchmark, revealing a distinct position bias phenomenon. As shown in Figure 2, we select the top 10% video tokens by attention scores and investigate their spatial distribution within individual frames. We define tokens located within a normalized distance of 10% from the top or bottom edges as boundary tokens (constituting approximately 20% of the total), while the remaining 80% are classified as interior tokens. We observe that among the selected tokens in each frame, boundary tokens account for a disproportionate share—exceeding 44% and peaking at 57.8%. This suggests that relying solely on attention-based token selection is insufficient for effectively capturing visual semantics.

We further conduct a controlled experiment using video-SALMONN2+ 7B as the draft model and video-SALMONN2+ 72B as the target model. We evaluate the draft model's performance under three

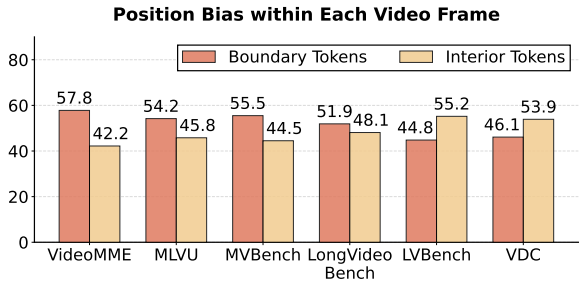


Figure 2: Statistics of per-frame positional bias.

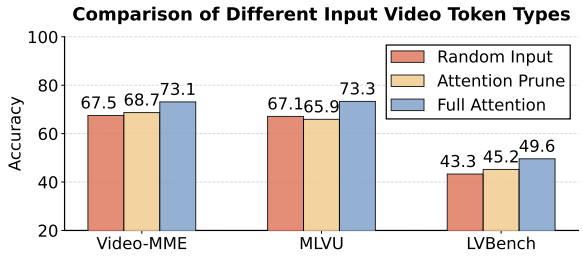


Figure 3: Performance comparison of three visual token inputs: random 10% sampling, target-model-guided top-10% attention selection, and full visual tokens.

input configurations: full visual tokens, top-10% tokens selected via target model attention scores, and a random 10% selection of visual tokens. We report accuracy on three video-QA benchmarks: Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2024b). As shown in Figure 3, both the attention-only and random strategies result in performance degradation compared to the full-token baseline. Notably, on MLVU, the random strategy outperforms attention-based pruning. This demonstrates the significance of preserving semantic visual tokens and suggests that attention-only selection can discard semantically important tokens due to position bias.

4.2 Non-negligible Draft Latency

We conduct a series of real-world experiments to demonstrate that draft model latency is a significant bottleneck in speculative decoding. As shown in Figure 4, both the drafting phase and verification phase consume substantial time. At each decoding step, the draft model remains idle during verification while the target model remains idle during drafting. This mutual waiting dilutes the speedup from accepted draft tokens, and the overhead from rejected draft tokens degrades performance.

For example, suppose the target model is three times slower than the draft model, i.e., $T_{\text{target}} = 3T_{\text{draft}}$. In each round, the draft model proposes 3

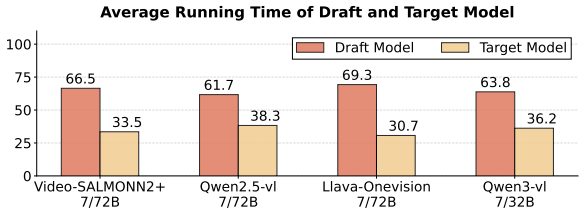


Figure 4: Statistics of average running time ratio of the draft and target model.

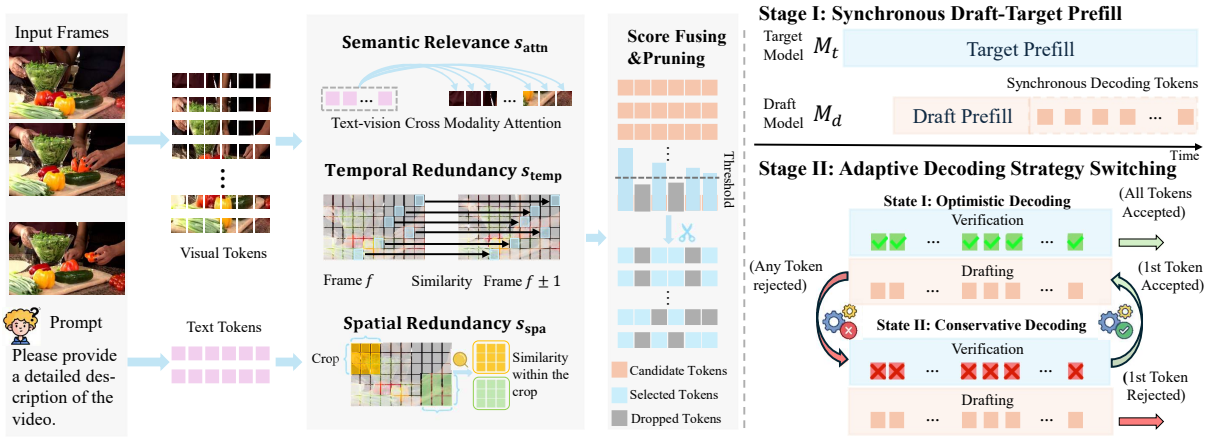
tokens. If all three tokens are accepted, the total computation corresponds to three forward passes of the draft model and one forward pass of the target model to obtain 4 correct tokens (3 draft tokens plus 1 token from the target model's verification). The effective speedup is $4T_{\text{target}}/(3T_{\text{draft}} + 1T_{\text{target}}) = 2$. However, if the first draft token is rejected, the same amount of computation (three draft forward passes and one target forward pass) yields only 1 correct token (from the target model's verification), and the effective speedup becomes $1T_{\text{target}}/(3T_{\text{draft}} + 1T_{\text{target}}) = 0.5$, which is significantly worse than vanilla AR decoding. Due to the serial execution of the draft and target models, together with the substantial draft model latency under long visual tokens, existing video-SD methods suffer from a fundamental bottleneck.

5 Method

We propose HIPPO, a **h**olist**i**c-aware **p**arallel s**p**eculative dec**o**ding framework that promotes video-LLMs to focus on informative visual content and overlapping the non-negligible draft latency. HIPPO integrates semantic-aware token preservation—which holistically scores tokens across global, temporal, and spatial dimensions to mitigate position bias—and video parallel speculative decoding—which overlaps draft generation with target verification via adaptive strategy switching. An overview of HIPPO is shown in Figure 5.

5.1 Semantic-aware Token Preservation

Consider a video of a news anchor. While the anchor's lip movements, gestures, and expressions exhibit temporal variations, the background and studio layout remain static across adjacent frames. From a spatial perspective, large intra-frame regions—such as the desk and walls—contain little information. This example illustrates the dual nature of video redundancy: **temporally**, static backgrounds repeat across frames; **spatially**, uniform regions within each frame carry minimal in-



(A) Semantic-aware Token Preservation **(B) Video Parallel Speculative Decoding**

Figure 5: An overview of HIPPO. (A) Semantic-aware Token Preservation: A holistic scoring mechanism integrates global semantic relevance, inter-frame temporal redundancy, and intra-frame spatial redundancy to select a compact subset of informative visual tokens, mitigating position bias. (B) Video Parallel Speculative Decoding: Synchronous Draft-Target Prefill generates initial draft tokens while the target model processes input, followed by an Adaptive Decoding Strategy that dynamically switches between optimistic post-verification and conservative pre-verification modes based on draft quality to maximize parallelism.

formation. An effective token preservation strategy should capture both dimensions to preserve semantic completeness under aggressive compression.

As demonstrated by motivated experiments in Section 4, relying solely on attention scores neglects crucial visual information due to position bias (Xiao et al., 2024). To address this, we propose a semantic-aware scoring mechanism that comprehensively evaluates each video token by integrating three orthogonal signals: **global semantic relevance**, **inter-frame temporal redundancy**, and **intra-frame spatial redundancy**. By scoring and ranking tokens holistically, we identify a compact subset that captures the essential visual semantics for the draft model.

Global Semantic Relevance. To identify tokens aligned with the user prompt, we compute a global semantic relevance score s_{attn} from language-to-video cross-attention. In our example, if the prompt is "What is the anchor reporting?", tokens corresponding to the anchor's face and on-screen graphics should receive high scores. Following prior work (Ji et al., 2025), we leverage KV-based reconstruction to avoid computing the full $O(L^2)$ attention matrix. Specifically, we extract visual keys $\mathbf{K}_{\text{vid}} \in \mathbb{R}^{N_v \times d_k}$ from the prefilled KV cache at minimal cost, obtain text queries $\mathbf{Q}_{\text{txt}} \in \mathbb{R}^{L_{\text{txt}} \times d_k}$, and compute the cross-modal attention matrix $\mathbf{S}_{\text{attn}} = \text{softmax}(\mathbf{Q}_{\text{txt}}\mathbf{K}_{\text{vid}}^T/\sqrt{d_k}) \in \mathbb{R}^{L_{\text{txt}} \times N_v}$. We aggregate weights across layers and heads:

$$s_{\text{attn}}(\mathbf{v}_j) = \frac{1}{N_L H L_{\text{txt}}} \sum_{\ell=1}^{N_L} \sum_{h=1}^{H} \sum_{i=1}^{L_{\text{txt}}} [\mathbf{S}_{\text{attn}}^{(\ell,h)}]_{i,j}, \quad (2)$$

where L_{txt} is the text sequence length, N_L is the number of layers, H is the number of attention heads, d_k is the key dimension, and ℓ, h index the layer and head respectively. This score effectively captures prompt-relevant content. However, it suffers from position bias, where boundary tokens receive inflated scores regardless of content.

Inter-frame Temporal Redundancy. To quantify temporal redundancy, we introduce a temporal score s_{temp} measuring cross-frame similarity at corresponding spatial locations. For example, tokens at static background positions exhibit high cross-frame similarity, while tokens at the anchor's moving lips show low similarity. For token $\mathbf{v}_j^f \in \mathbb{R}^d$ at position j in frame f, we compute its cosine similarity with tokens at the same spatial position in adjacent frames:

$$s'_{\text{temp}}(\mathbf{v}_j^f) = \frac{1}{|\mathcal{N}(f)|} \sum_{f' \in \mathcal{N}(f)} \frac{\mathbf{v}_j^{fT} \mathbf{v}_j^{f'}}{\|\mathbf{v}_j^f\| \|\mathbf{v}_j^{f'}\|}, \quad (3)$$

where $\mathcal{N}(f) = \{f-1, f+1\} \cap \{1, \ldots, F\}$ denotes adjacent frames. Since high similarity indicates redundancy (which we want to prune), we invert the score to prioritize temporal variation:

$$s_{\text{temp}}(\mathbf{v}_j^f) = 1 - s'_{\text{temp}}(\mathbf{v}_j^f). \quad (4)$$

Intra-frame Spatial Redundancy. We evaluate the visual complexity of each token within its local crop. For a given crop c containing M tokens, we compute the intra-crop similarity matrix by measuring pairwise relationships among all tokens. Given the L_2-normalized embeddings $\mathbf{E}_v^c \in \mathbb{R}^{M\times d}$ in crop c, we compute the similarity matrix $\mathbf{S}^c$ as:

$$\mathbf{S}^c = \mathbf{E}_v^c (\mathbf{E}_v^c)^\top. \tag{5}$$

We capture the spatial complexity of each token $\mathbf{v}_j^{f,c}$ by computing the variance of its similarity distribution within the crop:

$$s_{\text{spa}}(\mathbf{v}_j^{f,c}) = \frac{1}{M}\sum_{k=1}^{M}(\mathbf{S}_{j,k}^c - \mu_j^c)^2, \tag{6}$$

where $\mu_j^c = \frac{1}{M}\sum_{k=1}^{M}\mathbf{S}_{j,k}^c$ is the mean similarity of token j with other tokens in crop c, and M is the number of tokens per crop. High variance indicates that the token has diverse connections with others (e.g., anchor's gestures and expressions), suggesting high information content, while low variance suggests uniformity (e.g., desk and walls).

Holistic Score Fusion. The three scores capture complementary aspects: s_{attn} identifies prompt-relevant semantics but suffers position bias; s_{temp} detects temporal change but ignores static importance; s_{spa} measures structural complexity but lacks global semantic context. For each frame f, we normalize each score type independently: $\hat{s}(\mathbf{v}_j) = (s(\mathbf{v}_j) - \mu_f)/\sigma_f$, where μ_f and σ_f are the mean and standard deviation of that score type within frame f. The final holistic score is:

$$s(\mathbf{v}_j) = \hat{s}_{\text{attn}}(\mathbf{v}_j) + \hat{s}_{\text{spa}}(\mathbf{v}_j) + \hat{s}_{\text{temp}}(\mathbf{v}_j). \tag{7}$$

We finally retain the top-k (by default 10%) visual tokens via $s(\mathbf{v}_j)$.

5.2 Video Parallel Speculative Decoding

While semantic-aware token preservation effectively reduces the visual input to the draft model, sequential dependencies in standard speculative decoding create mutual waiting: the draft model $\mathcal{M}_d$ must complete all γ tokens before the target model $\mathcal{M}_t$ can verify, and $\mathcal{M}_d$ idles during verification while $\mathcal{M}_t$ idles during drafting. Drawing upon parallelization paradigms from text-only LLMs (Liu et al., 2025), we propose a **Video Parallel Speculative Decoding (VPSD)** framework tailored for video-LLMs. As shown in Figure 5, VPSD transforms the generation pipeline by overlapping draft generation and target verification phases.

Synchronous Draft-Target Prefill. In video-LLMs, the target model's prefill latency is dominated by the computationally expensive vision encoder, significantly exceeding the draft model's prefill time. We exploit this asymmetry during the initial prefill stage: both models process the input through the vision encoder, but $\mathcal{M}_d$ continues autoregressive decoding until $\mathcal{M}_t$ completes its prefill. For instance, if $\mathcal{M}_t$ requires 0.8s for prefill while $\mathcal{M}_d$ needs only 0.2s, the draft model can generate approximately 12 tokens (assuming 50ms per token) before the target model is ready. It accumulates a larger initial buffer of candidate tokens, effectively hiding the target model's overhead.

Adaptive Decoding Strategy Switching. After the prefilling stage, we introduce an adaptive decoding mechanism that dynamically switches between two strategies based on the verification outcome of the previous step. **When draft quality is high** (i.e., the previous batch was fully accepted), the system enters an **optimistic decoding mode**: $\mathcal{M}_d$ speculatively generates the next batch $[x_{\gamma+1}, \ldots, x_{2\gamma}]$ concurrently with $\mathcal{M}_t$'s verification of the current batch $[x_1, \ldots, x_\gamma]$. If all tokens are accepted, the next batch is already available for immediate use; if any rejection occurs, it discards the speculative tokens at no additional cost since the computation was performed in parallel. **Conversely, when draft quality is low** (i.e., a rejection occurred), it switches to a **conservative decoding mode**: $\mathcal{M}_t$ verifies the first draft token x_1 in parallel with $\mathcal{M}_d$'s generation of the remaining tokens $[x_2, \ldots, x_\gamma]$. If x_1 is rejected, we immediately abort drafting and discard $[x_2, \ldots, x_\gamma]$, preventing wasted computation; if accepted, the remaining tokens proceed to standard verification. This dynamic switching implements an implicit draft length adaptation: aggressive when $\mathcal{M}_d$ produces reliable drafts, cautious when the quality is low. It ensures both models operate in parallel throughout generation, alleviating the mutual waiting issue.

6 Experiments

6.1 Experiment Setups

Tasks and Datasets. We evaluate HIPPO on four video-LLMs: video-SALMONN2+ (Tang et al., 2025), Qwen2.5-VL (Bai et al., 2025b), LLaVA-OneVision (Li et al., 2025a), and Qwen3-VL (Bai et al., 2025a). We use caption and understanding datasets, including Video-MME (Fu et al., 2025), VideoDetailCaption (VDC) (Chai

Table 1: Performance comparison across different backbone models. We report Mean Acceptance Tokens (MAT) and wall-time speedup relative to standard auto-regressive decoding. We **bold** the best results for each model.

Backbone Model	Method	VideoMME		VDC		MVBench		LongVideoBench		MLVU		LVbench	
		MAT	Spd.	MAT	Spd.	MAT	Spd.	MAT	Spd.	MAT	Spd.	MAT	Spd.
Video-SALMONN2+ 7&72B	Vanilla SD	2.90	1.36×	2.76	1.41×	2.71	1.33×	2.75	1.34×	3.30	1.53×	2.89	1.57×
	SD-tree	3.71	1.87×	3.64	1.73×	3.18	1.54×	3.11	1.45×	3.86	1.83×	3.29	1.72×
	SpecVLM	3.46	2.24×	3.42	2.01×	3.00	2.09×	3.34	2.15×	4.00	2.13×	3.42	2.28×
	HIPPO	**11.82**	**2.85×**	**6.97**	**2.31×**	**12.31**	**2.89×**	**10.06**	**2.74×**	**12.42**	**2.78×**	**12.12**	**2.64×**
Qwen2.5-VL 7&72B	Vanilla SD	2.70	1.48×	3.37	1.56×	2.69	1.61×	2.97	1.52×	3.39	1.64×	3.11	1.57×
	SD-tree	3.32	1.73×	3.64	1.72×	3.00	1.85×	3.31	1.72×	3.84	2.01×	3.53	1.83×
	SpecVLM	3.05	2.14×	3.42	2.11×	3.02	1.89×	3.29	2.12×	3.48	2.14×	3.49	2.10×
	HIPPO	**10.29**	**2.71×**	**6.29**	**2.31×**	**8.28**	**2.43×**	**8.44**	**2.40×**	**7.19**	**2.54×**	**9.25**	**2.44×**
LLaVA-OneVision 7&72B	Vanilla SD	4.01	1.88×	3.32	1.89×	3.28	1.82×	3.17	1.73×	2.87	1.64×	2.94	1.68×
	SD-tree	4.47	2.03×	3.68	2.08×	3.63	2.02×	3.57	2.08×	3.30	1.90×	3.13	1.93×
	SpecVLM	4.32	2.10×	3.48	2.68×	3.40	2.35×	3.33	2.37×	2.96	2.00×	3.32	2.37×
	HIPPO	**8.93**	**3.27×**	**8.05**	**3.31×**	**7.93**	**3.51×**	**8.97**	**3.25×**	**7.96**	**3.17×**	**8.39**	**3.16×**
Qwen3-VL 7&32B	Vanilla SD	3.39	1.35×	3.79	1.27×	3.84	1.32×	3.62	1.21×	3.67	1.23×	3.88	1.39×
	SD-tree	3.91	1.73×	4.25	1.63×	4.25	1.72×	4.51	1.69×	2.51	1.53×	4.51	1.88×
	SpecVLM	3.87	1.89×	4.01	2.05×	4.12	2.03×	4.13	1.96×	2.33	1.85×	4.20	2.11×
	HIPPO	**4.67**	**2.38×**	**4.91**	**2.35×**	**4.64**	**2.30×**	**5.91**	**2.24×**	**2.82**	**2.15×**	**5.39**	**2.32×**

et al., 2025), MVBench (Li et al., 2024a), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2024b).

Implementations and Metrics. We conduct experiments on four NVIDIA H200 (140GB) GPUs with batch size of 1. We employ 7B models as draft models and 32B or 72B models as target models. We use two evaluation metrics: (i) wall-time speedup relative to standard auto-regressive decoding, and (ii) mean accepted tokens (MAT). MAT accumulates accepted tokens across verification rounds, including consecutive rounds where all γ tokens are successfully verified. Output quality is not evaluated since our acceleration is lossless. More details are provided in Appendix B.

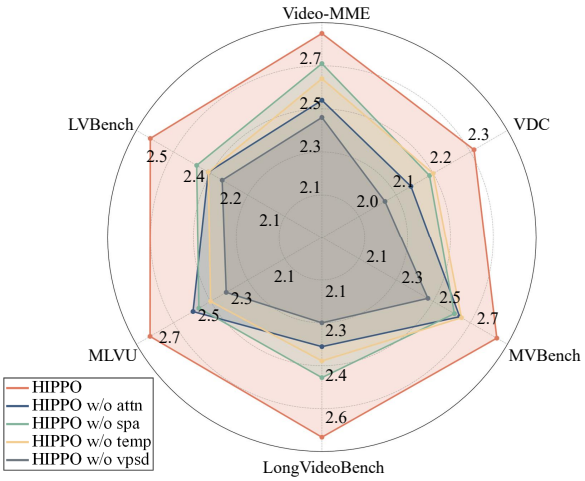


Figure 6: Ablation of HIPPO across six video benchmarks, using video-SALMONN2+ as the backbone.

6.2 Main Results

We conduct extensive experiments on the aforementioned benchmarks. As shown in Table 1, HIPPO consistently outperforms Vanilla SD, SD-tree, and SpecVLM across all backbone architectures and datasets, demonstrating robustness at varying model scales. Notably, HIPPO achieves an inference speedup of up to 3.51× compared to vanilla auto-regressive baselines. Moreover, HIPPO yields significantly higher MAT compared to existing baselines. Specifically, when utilizing video-SALMONN2+ as the backbone model on the MLVU dataset, HIPPO achieves a peak MAT of 12.42. This not only demonstrates the superior alignment between our pruned draft model and the target model but also validates that our parallel SD algorithm effectively mitigates the non-negligible inference overhead associated with draft models.

6.3 Ablation Study

To investigate the contribution of each component within HIPPO, we conduct an ablation study in Figure 6 using video-SALMONN2+. Results for the other three video-LLMs are in Appendix F. We denote HIPPO without global attention score as HIPPO $_{\text{w/o attn}}$, without temporal redundancy score as HIPPO $_{\text{w/o temp}}$, without spatial redundancy score as HIPPO $_{\text{w/o spa}}$, and without video parallel speculative decoding as HIPPO $_{\text{w/o vpsd}}$.

As shown in Figure 6, the absence of any compo-

nent within HIPPO results in performance degradation. Among the different scoring components, we observe that utilizing the target model's attention is consistently effective. Additionally, spatiotemporal-aware pruning contributes to improved performance. Specifically, on the MLVU dataset, removing temporal redundancy results in a marked decline in speedup. This underscores that reliance on target model attentions alone cannot effectively capture the essential information for complex video tasks, limiting the speedup ratios.

6.4 Case Study

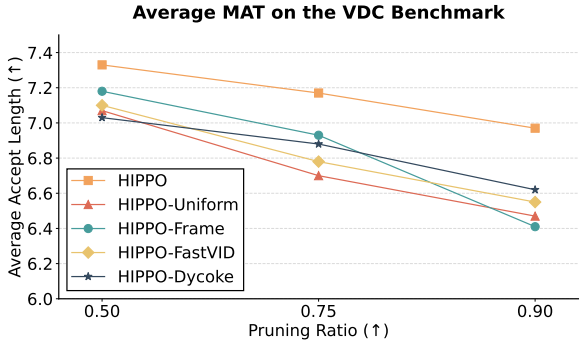


Figure 7: Comparison of different types of pruning methods under various pruning ratios.

Exploration of Diverse Pruning Criteria. Numerous studies investigate vision-centric pruning on standalone models to accelerate speed without significantly compromising accuracy (Shen et al., 2025; Tao et al., 2025). Following this paradigm, we adapt these methods for video-SALMONN2+ 7/72B to establish the following baselines:

(1) HIPPO-Random: randomly sampling top-k video tokens.

(2) HIPPO-Frame: full-frame dropping at regular temporal intervals.

(3) HIPPO-FastVID: frame-level pruning based on top-k similarity of consecutive frame transitions, following FastVID (Shen et al., 2025).

(4) HIPPO-DyCoke: token-level temporal merging adapted from DyCoke (Tao et al., 2025).

Figure 7 demonstrates that vision-centric pruning methods relying solely on a single model consistently underperform compared to HIPPO, which incorporates position-debiased guidance from the target model. We attribute this to two key factors: the draft model benefits from preserving the holistic temporal structure and distribution of the video, and the inclusion of target model signals provides essential guidance for achieving superior alignment. It is worth noting that single-model pruning techniques are not inherently incompatible with HIPPO. Exploring how to leverage intrinsic visual redundancy to refine pruning while integrating information from the target model is a promising direction for future research toward constructing better-aligned draft models.

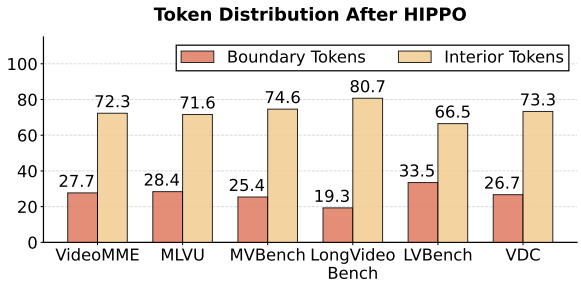


Figure 8: Per-frame token distributions after HIPPO.

Exploration of Token Distributions after HIPPO. As discussed in Section 4, pruning strategies that rely solely on the target model's attention often suffer from positional bias. To validate the effectiveness of HIPPO, we visualize token distributions after HIPPO, using video-SALMONN2+ as the backbone. As shown in Figure 8, tokens selected by HIPPO effectively mitigate this positional bias. Within each frame, HIPPO alleviates the tendency to disproportionately retain spatial boundary tokens and focuses on semantic subjects.

7 Conclusion

In this paper, we introduced HIPPO, a novel **h**olist**i**c-aware **p**arallel s**p**eculative dec**o**ding framework designed to accelerate video-LLM inference. HIPPO addresses the efficiency bottlenecks of existing methods through two key innovations: (i) a semantic-aware token preservation strategy that integrates **global attention scores** with **local visual semantics**, thereby maintaining high alignment between the draft and target models at high pruning ratios; (ii) a video parallel speculative decoding algorithm that decouples draft generation from verification, effectively eliminating the mutual waiting overhead. Experiments on four video-LLMs across six datasets demonstrate the effectiveness of our approach, achieving up to $3.51\times$ speedup compared to vanilla auto-regressive decoding.

8 Limitations

We consider a few limitations and future directions. (i) Specialized draft model training. Our current evaluation employs existing smaller models from the same family as draft models to accelerate the target model. Training lightweight, target-specific draft models with better alignment represents a promising direction for further improving speculative decoding efficiency. Our method is naturally compatible with such specialized draft models, requiring only one-time pruning of the draft model's KV cache during the prefill stage. As training lightweight video-LLMs remains an emerging area, future work can explore draft model optimization techniques—such as knowledge distillation or architecture specialization—to further enhance speculative decoding efficiency. (ii) Applicability scope. Our method is primarily effective in memory bandwidth scenarios as our adaptive pruning strategy directly targets memory I/O overhead reduction, extending the approach to high-throughput processing requires further investigation. (iii) Scalability to large batch sizes. Our evaluation focuses on latency-critical scenarios with batch size 1, the primary use case for real-time interaction. For high-throughput scenarios with large batches, GPU compute saturation may occur, and maintaining diverse dynamic tree structures per request could introduce non-trivial overhead.

9 Ethical Considerations

The data we collect in specialized domains is publicly available and viewable online. The data owners have indicated that the data can be used for scientific research or have not indicated that the data cannot be used for scientific research, and our collection process is also in compliance with regulations. Moreover, there is no unique identification of individuals or offensive content in these data.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL technical report. *arXiv preprint, arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint, arXiv:2502.13923*.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your ViT but faster. In *Proc. ICLR*, Kigali.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Proc. ICML*, Vienna.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. 2025. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *Proc. ICLR*, Singapore.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint, arXiv:2302.01318*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024. ShareGPT4Video: Improving video understanding and generation with better captions. In *Proc. EMNLP*, Vancouver.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*.

Cunxiao Du, Jing Jiang, Yuanchen Xu, Jiawei Wu, Sicheng Yu, Yongqi Li, Shenggui Li, Kai Xu, Liqiang Nie, Zhaopeng Tu, and Yang You. 2024. GliDe with a CaPE: A low-Hassle method to accelerate speculative decoding. In *Proc. ICML*, Vienna.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. 2024. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. *arXiv preprint, arXiv:2407.11550*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proc. CVPR*, Nashville.

Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. 2024. Break the sequential dependency of LLM inference using lookahead decoding. In *Proc. ICML*, Vienna.

Mukul Gagrani, Raghavv Goel, Wonseok Jeon, Junyoung Park, Mingu Lee, and Christopher Lott. 2024. On speculative decoding for multimodal large language models. *arXiv preprint, arXiv:2404.08856*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *Proc. ICLR*, Singapore.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. LLaVa-UHD: An LMM perceiving any aspect ratio and high-resolution images. In *Proc. ECCV*, Milan.

Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. 2024. ZipVL: Efficient large vision-language models with dynamic token sparsification and KV cache compression. *arXiv preprint, arXiv:2410.08584*.

Xiaohu Huang, Hao Zhou, and Kai Han. 2025. PruneVid: Visual token pruning for efficient video large language models. In *Proc. ACL*, Vienna.

Yicheng Ji, Jun Zhang, Heming Xia, Jinpeng Chen, Lidan Shou, Gang Chen, and Huan Li. 2025. SpecVLM: Enhancing speculative decoding of video LLMs via verifier-guided token pruning. *arXiv preprint, arXiv:2508.16201*.

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

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. 2024b. MARVEL: Multidimensional abstraction and reasoning through visual evaluation and learning. In *Proc. NeurIPS*, Vancouver.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proc. ICML*, Honolulu.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024a. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proc. CVPR*, Seattle.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024b. LLaMa-VID: An image is worth 2 tokens in large language models. In *Proc. ECCV*, Milan.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2025b. EAGLE-3: Scaling up inference acceleration of large language models via training-time test. *arXiv preprint, arXiv:2503.01840*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVa: Learning united visual representation by alignment before projection. In *Proc. EMNLP*, Miami.

Luxi Lin, Zhihang Lin, Zhanpeng Zeng, and Rongrong Ji. 2025. Speculative decoding reimagined for multimodal large language models. *arXiv preprint, arXiv:2505.14260*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proc. CVPR*, Seattle.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proc. NeurIPS*, New Orleans.

Tianyu Liu, Yun Li, Qitan Lv, Kai Liu, Jianchen Zhu, Winston Hu, and Xiao Sun. 2025. PEARL: Parallel speculative decoding with adaptive draft length. In *Proc. ICLR*, Singapore.

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

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

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. 2025. DeCo: Frequency-decoupled pixel diffusion for end-to-end image generation. *arXiv preprint, arXiv:2511.19365*.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Rae Ying Yee Wong, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. 2023. SpecInfer: Accelerating generative LLM serving with speculative inference and token tree verification. *arXiv preprint, arXiv:2305.09781*.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Proc. NeurIPS*, Vancouver.

Ranajoy Sadhukhan, Jian Chen, Zhuoming Chen, Vashisth Tiwari, Ruihang Lai, Jinyuan Shi, Ian En-Hsu Yen, Avner May, Tianqi Chen, and Beidi Chen. 2025. MagicDec: Breaking the latency-throughput tradeoff for long context generation with speculative decoding. In *Proc. ICLR*, Singapore.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint, arXiv:2403.15388*.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. 2025. FastVID: Dynamic density pruning for fast video large language models. *arXiv preprint, arXiv:2503.11187*.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. 2025a. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proc. COLING*, Abu Dhabi.

Mingbo Song, Heming Xia, Jun Zhang, Chak Tou Leong, Qiancheng Xu, Wenjie Li, and Sujian Li. 2025b. KNN-SSD: Enabling dynamic self-speculative decoding via nearest neighbor layer set optimization. *arXiv preprint, arXiv:2505.16162*.

Fengyuan Sun, Leqi Shen, Hui Chen, Sicheng Zhao, Jungong Han, and Guiguang Ding. 2025. AdaTP: Attention-debiased token pruning for video large language models. *arXiv preprint, arXiv:2505.20100*.

Hanshi Sun, Zhuoming Chen, Xinyu Yang, Yuandong Tian, and Beidi Chen. 2024. TriForce: Lossless acceleration of long sequence generation with hierarchical speculative decoding. *arXiv preprint, arXiv:2404.11912*.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix X. Yu. 2023. SpecTr: Fast speculative decoding via optimal transport. In *Proc. NeurIPS*, New Orleans.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-SALMONN 2: Captioning-enhanced audio-visual large language models. *arXiv preprint, arXiv:2511.21631*.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. DyCoke: Dynamic compression of tokens for fast video large language models. In *Proc. CVPR*, Nashville.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint, arXiv:2312.11805*.

Llama Team. 2024. The Llama 3 herd of models. *arXiv preprint, arXiv:2307.09288*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? Exploring the visual shortcomings of multimodal llms. In *Proc. CVPR*, Seattle.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMa: Open and efficient foundation language models. *arXiv preprint, arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint, arXiv:2307.09288*.

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

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024b. LVBench: An extreme long video understanding benchmark. *arXiv preprint, arXiv:2406.08035*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proc. EMNLP*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Proc. NeurIPS*, Vancouver.

Heming Xia, Yongqi Li, Jun Zhang, Cunxiao Du, and Wenjie Li. 2025. SWIFT: On-the-fly self-speculative decoding for LLM inference acceleration. In *Proc. ICLR*, Singapore.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proc. ICLR*, Vienna.

Zhinan Xie, Peisong Wang, Shuang Qiu, and Jian Cheng. 2025. HiViS: Hiding visual tokens from the drafter for speculative decoding in vision-language models. *arXiv preprint, arXiv:2509.23928*.

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

Penghui Yang, Cunxiao Du, Fengzhuo Zhang, Haonan Wang, Tianyu Pang, Chao Du, and Bo An. 2025. LongSpec: Long-context speculative decoding with efficient drafting and verification. *arXiv preprint, arXiv:2502.17421*.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G. Hauptmann, Yonatan Bisk, and Yiming Yang. 2025a. Direct preference optimization of video large multimodal models from language model reward. In *Proc. ACL*, Albuquerque.

Xuan Zhang, Cunxiao Du, Sicheng Yu, Jiawei Wu, Fengzhuo Zhang, Wei Gao, and Qian Liu. 2025b. Sparse-to-dense: A free lunch for lossless acceleration of video understanding in LLMs. In *Proc. ACL*, Vienna.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: Benchmarking multi-task long video understanding. In *Proc. CVPR*, Nashville.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2025. Don't just chase "highlighted tokens" in MLLMs: Revisiting visual holistic context retention. *arXiv preprint, arXiv:2510.02912*.

A More Discussion of Related Work

In Section 2, we briefly discuss the recent achievements of speculative decoding and visual token reduction methods. In this section, we provide a more detailed discussion on existing multi-modal LLMs and their computational challenges, various speculative decoding methods for LLMs, and visual token reduction methods.

A.1 MLLMs and Computational Challenges

Recent MLLMs typically adhere to a paradigm that bridges multi-modal encoders with LLMs via specialized modality adapters (Team, 2023; Liu et al., 2023, 2024a). Built upon open-source foundations such as the LLaMA series (Touvron et al., 2023a,b; Team, 2024), state-of-the-art MLLMs (Team, 2023; OpenAI, 2023) have demonstrated remarkable adaptability across diverse visual understanding tasks, significantly enhancing their ability to interpret complex real-world environments. In the video domain, video-LLaVA (Lin et al., 2024) aligns image and video adapters to learn unified representations. ShareGPT4Video (Chen et al., 2024) utilizes GPT-4 to generate dense video captions for data quality improvement, while LLaVA-Hound (Zhang et al., 2025a) introduces Direct Preference Optimization (DPO) to refine video comprehension capabilities. Furthermore, LLaVA-OneVision (Li et al., 2025a) achieves robust performance in single-image, multi-image, and video scenarios, facilitating effective transfer learning from static images to dynamic video content. Despite these advancements, the integration of extensive visual tokens introduces substantial computational overhead. While visual perception is improved, existing MLLMs continue to suffer from certain perceptual limitations (Tong et al., 2024; Jiang et al., 2024b). Strategies to mitigate these deficiencies by increasing the resolution of input images or videos (Luo et al., 2025; Guo et al., 2024) inadvertently exacerbate computational costs. For instance, LLaVA-OneVision (Li et al., 2025a) encodes a single video frame into 196 tokens; consequently, processing a two-minute video at 60 FPS yields over one million tokens. Such massive visual inputs become a significant portion of the LLM's context window. In this work, we conduct experiments on these representative models to empirically validate the applicability and efficiency of our HIPPO.

A.2 Speculative Decoding for LLMs

Pioneering approaches in speculative decoding (Leviathan et al., 2023; Chen et al., 2023) typically employ an independent, lightweight draft model to propose short token sequences, which are subsequently verified in parallel by the target model. To ensure reliable speculation, these early methods often utilize existing smaller LLMs as drafters. Alternatively, self-speculative frameworks (Xia et al., 2025; Song et al., 2025b) leverage intermediate layers of the target model itself to generate predictions, thereby eliminating the need for auxiliary models. Contemporary research has largely focused on enhancing draft model efficiency; for instance, recent work (Du et al., 2024) introduces specially trained drafters with reduced latency to maximize the speedup ratio. Several studies have also adapted speculative decoding methods from the text domain to the single-image domain (Lin et al., 2025; Gagrani et al., 2024; Xie et al., 2025).

Beyond standard decoding, several SD methods address the specific challenges of long-context text inference (Yang et al., 2025; Sun et al., 2024; Sadhukhan et al., 2025). These approaches aim to mitigate the Key-Value (KV) cache bottleneck inherent in conventional SD by integrating KV cache sparsification techniques (Feng et al., 2024), often utilizing StreamingLLM-style cache management (Xiao et al., 2024) to constrain the draft model's memory budget. However, the absence of visual modality awareness precludes these methods from operating effectively on video tokens. Unlike discrete linguistic tokens, video tokens exhibit substantial redundancy and high temporal correlation. Consequently, the naive application of StreamingLLM or sliding window mechanisms restricts the draft model's attention to a narrow fraction of uncompressed visual content, causing it to fail in capturing the visual semantics of the video.

A.3 Visual Token Reductions

In MLLMs, visual redundancy identification facilitates the distillation of visual tokens with high informativeness for faster inference. There are two main research directions: (i) Vision-centric strategies analyze the image's structure and feature distribution to discard less relevant visual tokens (Li et al., 2024b; Wang et al., 2024a; Sun et al., 2025). Approaches such as LLaMA-VID (Li et al., 2024b) and DeCo (Ma et al., 2025) address

this by modifying model architectures with additional training; however, this inevitably escalates computational expenses. While ToMe (Bolya et al., 2023) achieves token reduction without retraining, it risks disrupting early-stage cross-modal interactions (Xing et al., 2024). LLaVA-PruMerge (Shang et al., 2024) selectively retains salient tokens while merging less critical ones based on key similarity. (ii) Instruction-centric strategies typically use cross-modal attention analysis or gradient accumulation to identify redundant tokens (Liu et al., 2024b; Song et al., 2025a). Tokens with low attention or negligible gradient impact are deemed redundant (He et al., 2024). Building on this, some studies explore learned importance scoring and retain the highlighted visual tokens (Jiang et al., 2024a). Distinct from these single-model pruning approaches, HIPPO focuses on leveraging signals derived from the target model to guide draft model pruning. Consequently, our HIPPO is orthogonal to existing techniques; the pruned draft model can be seamlessly combined with these methods to adapt to specific downstream tasks.

B More Details of Experiment Setups

We present more details of datasets and experiment setups in this section.

Details of Baselines. We implement and evaluate three inference acceleration baselines to benchmark our approach: **(i) Vanilla Speculative Decoding (SD)**: we employ a compact video-LLM as the draft model to generate preliminary token sequences, which are subsequently verified by the target model. This foundational approach establishes the basic speculative decoding framework for video-language tasks, where the draft model produces candidate continuations in an auto-regressive manner, and the target model validates these proposals in parallel. **(ii) Speculative Decoding with Tree Attention (SD-tree)**: Building upon vanilla SD, this method constructs a draft tree structure by proposing multiple candidate tokens at each position, thereby exploring diverse continuation paths simultaneously. The tree-based representation allows the target model to verify multiple speculative branches in parallel, potentially increasing the acceptance rate and overall throughput compared to sequential speculation. **(iii) SpecVLM**: this approach introduces an attention-guided token selection mechanism, where candidate tokens are strategically chosen based on the target model's attention patterns from previous decoding steps. By leveraging these attention weights as guidance signals, SpecVLM prunes redundant or low-probability candidates early in the speculation process, focusing computational resources on more promising token sequences and reducing unnecessary verification overhead.

Notably, we avoid introducing additional training procedures for the draft models in our baseline implementations. This design decision is motivated by the observation that the primary computational bottleneck in video-LLM inference stems from the management and retrieval of large key-value caches rather than the forward pass through model parameters. Therefore, our baselines focus on optimizing the decoding strategy and speculation efficiency while maintaining the original model weights.

Details of Experimental Setups. We conduct experiments on four NVIDIA H200 (140GB) GPUs. We employ 7B models as draft models and 32B or 72B models as target models. We perform all evaluations with a batch size of 1 to follow the standard settings and simulate real-world latency-critical inference scenarios. Given each prompt, the target model is employed to generate 256 tokens following greedy decoding. When video tokens are pruned, we remove the corresponding video features based on the pruning ratio r. During evaluation, the default pruning ratio r is set to 90%. The value of γ is set to 5, 5, 7, and 3 for video-SALMONN2+ 7B/72B, Qwen2.5-VL 7B/72B, LLaVA-OneVision 7B/72B, and Qwen3-VL 7B/32B, respectively. The crop size M is set to 5×5 for all models and benchmarks. To ensure a fair comparison, all methods are evaluated under an identical hardware environment using model parallelism. For the AR baseline, the target model spans four GPUs. For the all SD-based methods we apply the standard protocol by deploying both the draft and target models across the four GPUs.

For LLaVA-OneVision, we uniformly sample 64 and 128 frames to generate a 196×64 and 196×128 video token input, respectively. For Qwen series and video-SALMONN2+, we adjust the FPS to generate input of comparable length. For video-SALMONN2+, we only apply pruning to the visual components. The audio tokens are excluded from the scoring process and are retained in their entirety as input. Our entire experimental framework is implemented using `PyTorch` (Paszke et al., 2019) version `2.8.0`, a widely-adopted deep learning

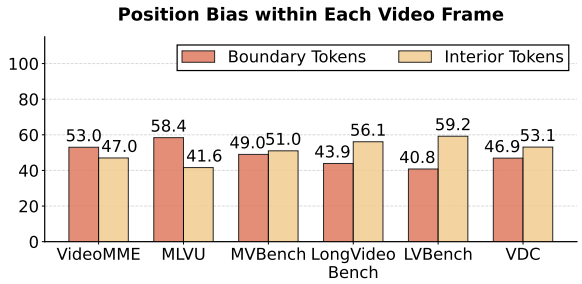


Figure 9: Statistics of per-frame positional bias using Qwen2.5-VL as the backbone model.

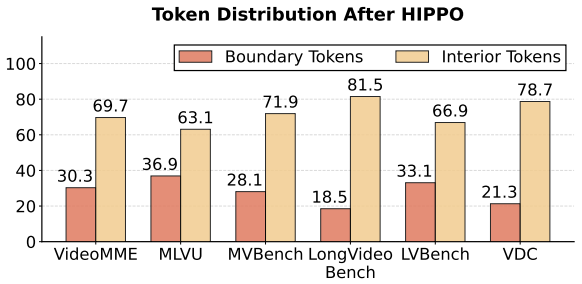


Figure 10: Per-frame token distributions after HIPPO using Qwen2.5-VL as the backbone model.

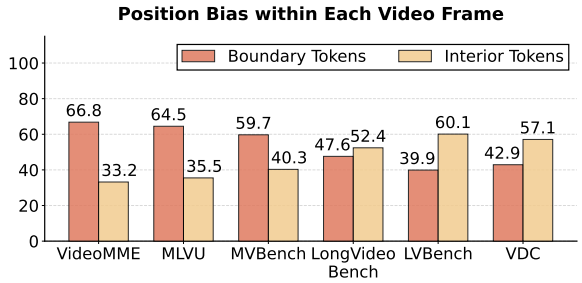


Figure 11: Statistics of per-frame positional bias using LLaVA-OneVision as the backbone model.

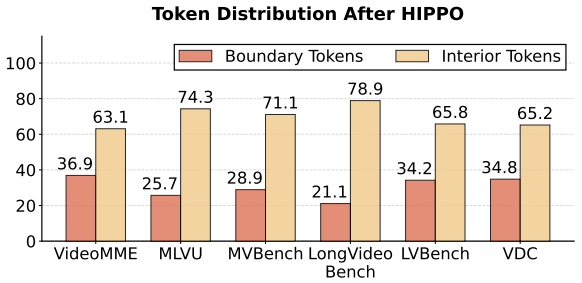


Figure 12: Per-frame token distributions after HIPPO using LLaVA-OneVision as the backbone model.

Table 2: Inference time breakdown (s) of video-SALMONN2+ 7B/72B pair. "-" indicates that vanilla AR does not require a draft model or pruning. "overlapped" indicates that due to our parallelized HIPPO, the execution time of the draft model is overlapped by the concurrent execution of the target model.

Operation	Vanilla	HIPPO
Target Model Prefilling	33.71	33.71
Target Model Decoding	131.13	25.11
Draft Model Prefilling	-	2.79 (overlapped)
Draft Model Decoding	-	23.77 (overlapped)
Video Token Pruning	-	0.11 (overlapped)
Total Latency	164.84	58.82

framework that provides comprehensive support for GPU-accelerated tensor operations and automatic differentiation. We leverage the `Transformers` (Wolf et al., 2020) library version `4.57.0` from Hugging Face. The codebase is compiled and executed with CUDA toolkit version `12.8`. More details for the best performance of each task and benchmark can be seen within our code.

C Inference Time Breakdown

Beyond accelerating the decoding stage of target video-LLMs, HIPPO introduces minimal overhead during the pruning process. We randomly sample 10 videos in Video-MME, employ a video-SALMONN2+ 7B/72B model pair for evaluation, and report average wall-time of each operation. As shown in 2, through video token pruning (only 0.11s for the whole decoding process), the prefill length of the draft model is substantially reduced. Furthermore, since HIPPO parallelizes the prefill and decoding processes of both the draft and target model, the execution times of these operations overlap, thereby further reducing the overall latency.

D More Comparison with Parallel Speculative Methods

As discussed in the main text, within the textual domain, PEARL (Liu et al., 2025) employs a parallelization strategy for draft and target models to enable adaptive draft lengths and eliminate mutual waiting latency. Building upon this, we further optimize the framework for the video domain. Specifically, observing that the target model's prefill latency in video-LLMs is dominated by the computationally expensive vision encoder—thereby significantly exceeding the draft model's prefill time—we propose the Synchronous Draft-Target Prefill mechanism. To comprehensively evaluate the effectiveness of our HIPPO, we compare it against several baselines: (i) PEARL-Video, a direct adaptation of the textual PEARL to the video domain; (ii) PEARL w/ Syn-Prefill, which incor-

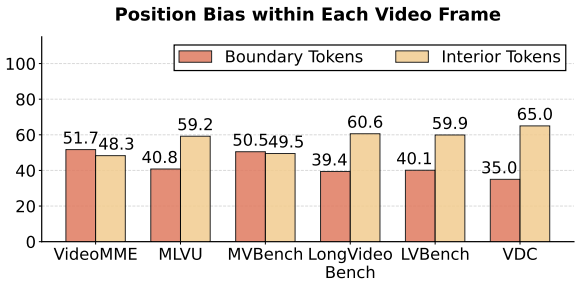


Figure 13: Statistics of per-frame positional bias using Qwen3-VL as the backbone model.

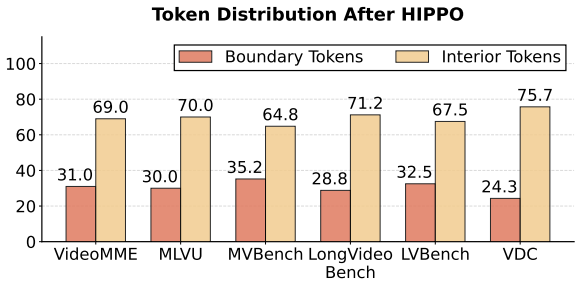


Figure 14: Per-frame token distributions after HIPPO using Qwen3-VL as the backbone model.

Table 3: Performance comparison with parallel speculative decoding methods, using video-SALMONN2+ 7B/72B as the backbone model pair. We report wall-time speedup relative to standard auto-regressive decoding. We **bold** the best results for each benchmark.

Method	**VideoMME** Spd.	**VDC** Spd.	**MVBench** Spd.	**LongVideoBench** Spd.	**MLVU** Spd.	**LVbench** Spd.
PEARL-video	2.24×	1.87×	2.45×	2.18×	2.20×	2.21×
PEARL w/ Syn-Prefill	2.46×	2.03×	2.64×	2.35×	2.30×	2.33×
PEARL-SpecVLM	2.52×	2.06×	2.69×	2.47×	2.46×	2.44×
HIPPO	**2.85**×	**2.31**×	**2.89**×	**2.74**×	**2.78**×	**2.64**×

porates our proposed synchronous draft-target prefill into the PEARL framework; and (iii) PEARL-SpecVLM, which integrates SpecVLM's attention-based draft selection mechanism within the parallel speculative decoding framework. Using video-SALMONN2+ 7B/72B as the backbone model pair, we present the results in Table 3. We observe that our HIPPO consistently outperforms competing parallel SD methods. This indicates that HIPPO effectively preserves visual semantic tokens, thereby improving draft quality. Furthermore, PEARL w/ Syn-Prefill achieves consistent improvements over vanilla PEARL-video across all benchmarks. This demonstrates that employing synchronous draft-target prefill in the video domain allows for the accumulation of a larger initial buffer of candidate tokens, effectively utilizing the target model's prefilling stage.

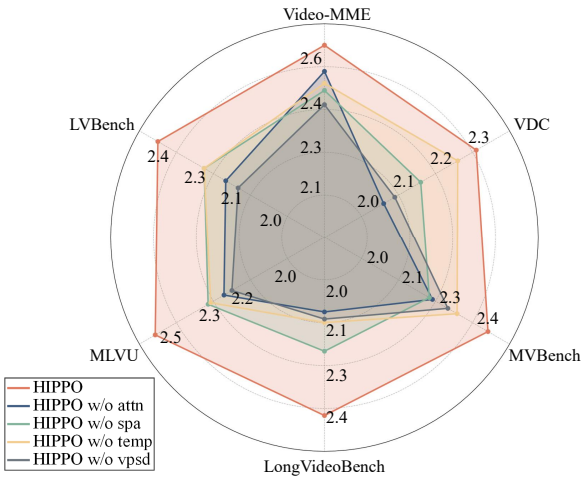


Figure 15: Ablation of HIPPO across six video benchmarks, using Qwen2.5-VL as the backbone.

E More Motivated Experiments

In Section 4, we present comprehensive position bias statistics utilizing video-SALMONN2+ as the backbone model. To establish the generality and broader applicability of our observations, we conduct extensive empirical analyses across a diverse range of video-LLMs with varying architectures and capacities. These systematic experiments substantiate that the identified inefficiencies associated with position bias are not merely model-specific phenomena, but rather represent fundamental and inherent challenges endemic to attention-based pruning mechanisms in existing video-LLM SD approaches.

As illustrated in Figures 9, 11, and 13, when employing solely the target model's attention scores to extract highlighted visual tokens across different video-LLM architectures, a pronounced position bias consistently emerges across all evaluated models. Conversely, as shown in Figures 10, 12, and 14, we observe that tokens selected by our proposed HIPPO effectively mitigate this positional

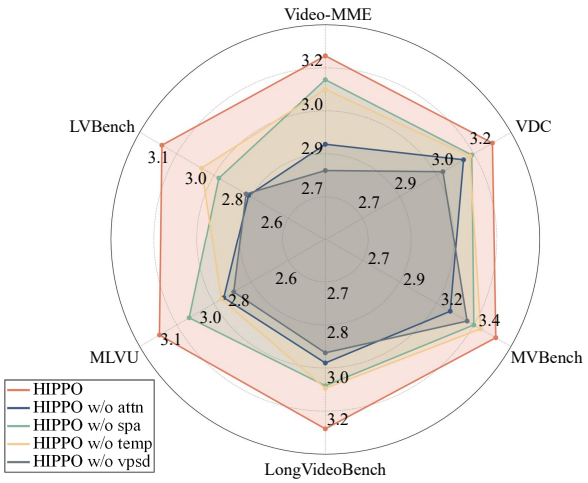


Figure 16: Ablation of HIPPO across six video benchmarks, using LLaVA-OneVision as the backbone.

bias phenomenon. HIPPO preserves semantically salient tokens distributed throughout frames, rather than exhibiting the tendency to disproportionately favor tokens located at spatial boundaries, thereby achieving a balanced token selection strategy that captures the true informational content of the video.

F More Results of Ablation Study

In Section 6.3, we conduct comprehensive ablation studies on both the semantic-aware token selection mechanism and the video parallel speculative decoding framework, initially employing video-SALMONN2+ as the backbone model. To rigorously validate the generalization and consistency of our proposed modules across diverse video-LLMs, we extend these ablation experiments to encompass three additional popular video-LLMs: Qwen2.5-VL, LLaVA-OneVision, and Qwen3-VL. This broader experimental scope enables us to establish the universal efficacy of our approach beyond a single model implementation.

As illustrated in Figures 15, 16, and 17, our ablation results consistently demonstrate that the removal of any individual component within the HIPPO framework leads to measurable performance degradation across all evaluated backbone models, thereby validating the necessity and synergistic contribution of each module. When examining different pruning strategies in isolation, we observe that leveraging the spatial and temporal redundancy yields consistently effective results across all architectures. This empirical finding provides evidence that reliance solely on target model attention signals—without explicit vision-centric modeling—proves insufficient for effectively capturing the hierarchical and temporally-evolving semantic information essential for complex, long-form video understanding tasks. These results underscore the critical importance of our integrated spatiotemporal-aware pruning strategy in achieving both computational efficiency and semantic fidelity.

G More Results of Case Study

In Section 6.4, we evaluate traditional vision-centric single-model pruning methods using video-SALMONN2+ as the backbone model on the VideoDetailedCaption dataset. Our preliminary findings reveal that pruning methods relying exclusively on signals from a single model consistently underperform compared to our proposed HIPPO approach, which strategically incorporates position-debiased guidance derived from the target model. To comprehensively assess the generalization of these observations, we extend our comparative analysis to different video-LLMs.

As illustrated in Figures 18, 19, and 20, our extended experimental results corroborate the initial findings across all evaluated datasets. Specifically, single-model pruning strategies consistently exhibit inferior performance when contrasted with our proposed framework that integrates position-debiased guidance signals from the target model. We attribute this substantial and consistent performance gap to two fundamental factors. First, the draft model derives significant benefits from preserving the holistic temporal structure and distributional characteristics inherent in the original video sequence—properties that single-model attention-based pruning often fails to maintain due to its susceptibility to position bias. Second, the explicit incorporation of target model signals provides crucial cross-model guidance that enables superior alignment with the target model's internal representation space and decision-making process, thereby facilitating more effective speculative decoding.

It is important to emphasize that single-model pruning techniques are not inherently incompatible with our HIPPO. Consequently, investigating principled methods to synergistically combine these two paradigms—specifically, leveraging vision-centric redundancy signals to refine initial pruning decisions while simultaneously integrating target model guidance to ensure distributional alignment—represents a particularly promising avenue for future research. Such hybrid strategies hold

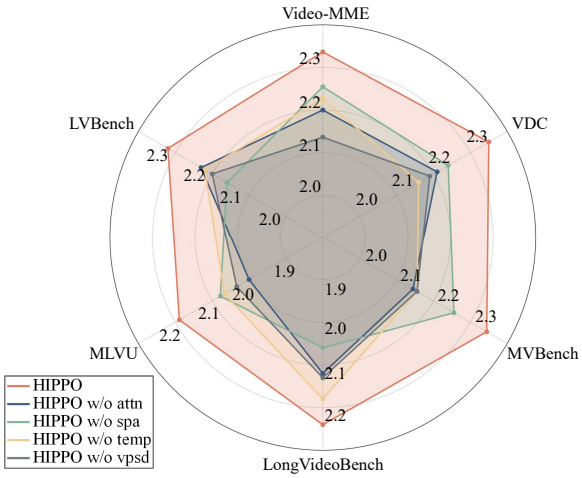


Figure 17: Ablation of HIPPO across six video benchmarks, using Qwen3-VL as the backbone.

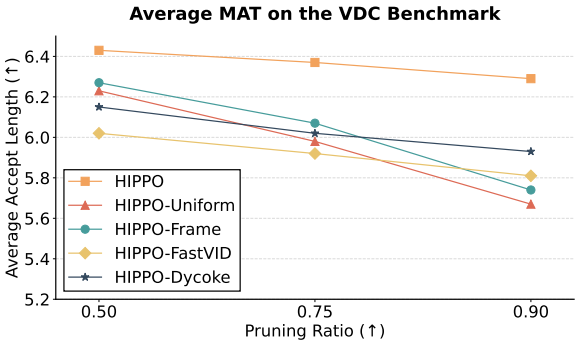


Figure 18: Comparison of different types of pruning methods under various pruning ratios using Qwen2.5-VL as the backbone model.

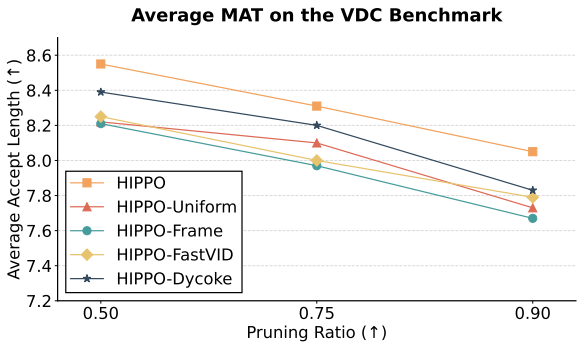


Figure 19: Comparison of different types of pruning methods under various pruning ratios using LLaVA-OneVision as the backbone model.

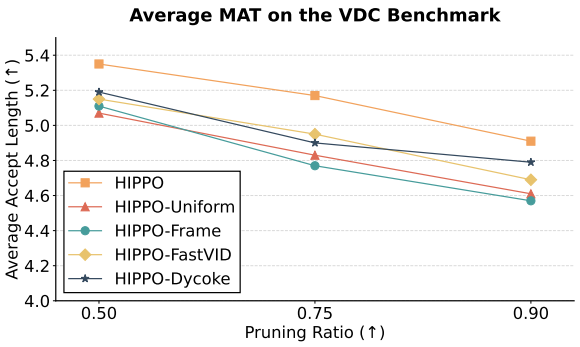


Figure 20: Comparison of different types of pruning methods under various pruning ratios using Qwen3-VL as the backbone model.

substantial potential for constructing draft models that are simultaneously more computationally efficient and better aligned with target model semantics, thereby pushing the efficiency-accuracy frontier of video-LLM speculative decoding.

H LLM Usage

We used a large language model (LLM)–based writing assistant for grammar and wording improvements on draft text. The LLM did not generate research ideas, claims, proofs, algorithms, code, figures, or analyses, and it did not have access to any non-public data. During rationale generation, we use LLMs to transfer reasoning path into rationales. All edits suggested by the LLM were manually reviewed and either accepted or rewritten by the authors, who take full responsibility for the final content. The LLM is not an author of this paper.