OMNIZPD: TOKEN COMPRESSION VIA POLICY OPTIMIZATION AND SELF-DISTILLATION FOR OMNI-MODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (OmniLLMs) suffer from high inference costs due to dense audio-visual token redundancy. Existing token compression methods often discard reasoning-critical evidence, as sparse final-answer rewards fail to supervise intermediate thinking processes. We propose **OmniZPD**, an efficient compression framework integrating policy optimization and self-distillation. OmniZPD features a unified, ratio-conditioned selector that jointly evaluates audio-visual token importance based on queries, compression budgets, and local cross-modal contexts. Keeping the backbone frozen, we optimize the selector using a bounded reward that combines task correctness with self-distillation feedback, enforcing consistent reasoning predictions between full and compressed inputs. At inference, a single trained selector deterministically prunes tokens and flexibly adapts to varying retention budgets without retraining. Experiments across multiple omni-modal reasoning benchmarks show that OmniZPD significantly outperforms existing baselines, achieving a 2.79× prefill speedup while retaining 91.8% of the full-input accuracy with merely 20% of audio-visual tokens.

1 INTRODUCTION

Omni-modal large language models (OmniLLMs) have demonstrated strong capabilities in complex multimodal understanding and reasoning by jointly modeling audio, video, and text (Hurst et al., 2024; Li et al., 2024; Xu et al., 2025a; Ye et al., 2026). However, dense audio-visual inputs produce large numbers of tokens, substantially increasing context length and computational cost. Multimodal token compression reduces the input to the language decoder by selecting or merging redundant representations (Zhang et al., 2024; Yang et al., 2025a; Alvar et al., 2025; Shi et al., 2023; Ye et al., 2025), offering a practical path toward efficient omni-modal reasoning (Yang et al., 2025b). For models with explicit reasoning, compression must preserve the multimodal evidence needed for both intermediate reasoning and answer generation within a limited token budget (Sun et al., 2025).

Existing compression methods typically estimate token importance using attention(Chen et al., 2024), feature similarity(Bolya et al., 2022), or cross-modal relevance(Tao et al., 2026; Zhong et al., 2026), but these criteria often misalign with downstream reasoning needs(Yu et al., 2025; Zhang et al., 2025). Consequently, recent works have explored policy optimization to learn input selection directly from task feedback(Wang et al., 2026a; Lee et al., 2026). For instance, TSPO(Tang et al., 2026) optimizes video frame sampling via question-answering rewards, while SCORE (Wang et al., 2026b) learns visual token compression through generation-quality and sparsity signals. However, extending these unimodal paradigms to omni-modal reasoning introduces unique challenges due to audio-visual complementarity. Because audio-visual information density is highly asymmetric and temporally misaligned, applying independent pruning risks destroying the cross-modal context required to interpret the remaining tokens (e.g., a visual cue may become meaningless without its accompanying sound). This necessitates a unified selection policy that jointly evaluates token relevance and cross-modal dependencies under a shared compression budget.

Beyond the selection mechanism, the supervision signal determines what information compression preserves. Fig. 1 illustrates this challenge: the cat's contact with the cup and the accompanying sound provide complementary evidence for explaining why the cup falls. Rule-based selection may

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.

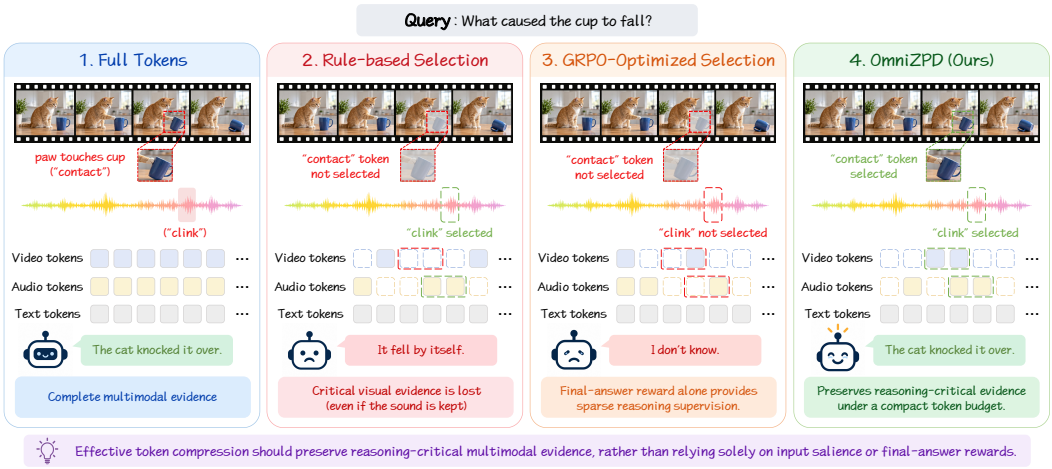

Figure 1: Motivation for OmniZPD. Token compression should preserve complementary audio-visual evidence needed for reasoning, beyond input importance or final-answer feedback alone.

discard these cues when predefined importance scores fail to capture their joint reasoning value. Furthermore, for OmniLLMs with explicit thinking, relying solely on final-answer rewards is insufficient. A compressed input might yield a correct answer through guessing or spurious cues, leaving the intermediate reasoning unsupported by genuine multimodal evidence. Conversely, an incorrect answer fails to indicate whether critical evidence was actually preserved. Consequently, final-answer correctness provides limited guidance on the faithfulness of the reasoning trajectory. This raises a central question: **How can intermediate reasoning guide audio-visual token selection?**

To address this question, we examine how token selection alters the model's predictions during reasoning, using full-input predictions as a reference. Inspired by on-policy distillation (OPD) (Agarwal et al., 2024) and self-distillation, we evaluate the same frozen OmniLLM under both full and compressed inputs, conditioned on identical reasoning prefixes generated by the compressed model. This comparison directly measures the impact of compression along the actual reasoning trajectory, yielding dense supervision signals beyond mere final-answer correctness. Furthermore, because our optimization objectives—spanning discrete answer correctness and generative distribution discrepancies—are non-differentiable, they preclude standard backpropagation. We therefore aggregate these discrepancies into a bounded reward and employ policy optimization to train the selector end-to-end for faithful evidence retention.

Based on this insight, we propose **OmniZPD**, a token compression framework that synergizes policy optimization and self-distillation for omni-modal reasoning. OmniZPD employs a lightweight, unified selector shared across audio and video modalities. Conditioned on textual queries, local cross-modal features, and a target retention ratio, this single policy jointly evaluates audio-visual token importance by modeling cross-modal dependencies across compression budgets. During training, we sample multiple compressed selections per input and prompt the frozen OmniLLM to generate corresponding reasoning trajectories. We then construct a bounded reward that integrates answer correctness, prediction confidence, and self-distillation feedback. By optimizing the selector via advantage-based policy gradients, OmniZPD learns to balance final answer quality with reasoning consistency, entirely without modifying the backbone OmniLLM.

Notably, the heavy computational overhead of policy sampling and self-distillation is strictly confined to the training phase. At inference, OmniZPD requires merely a single lightweight forward pass for deterministic token selection, introducing negligible latency. Furthermore, thanks to the ratio-conditioned design, a single trained selector can seamlessly adapt to varying retention budgets at deployment without retraining, offering flexible scaling across diverse hardware constraints. Extensive evaluations across multiple omni-modal reasoning benchmarks demonstrate the superiority of our approach: by retaining only 20% of the audio-visual tokens, OmniZPD achieves a 2.79× prefill speedup while preserving **91.8%** of the full-input performance *on average*.

Our main contributions are summarized as follows:

- We propose a novel supervision paradigm for omni-modal compression that leverages self-distillation along student-generated reasoning trajectories, providing dense feedback to preserve faithful intermediate reasoning beyond sparse final-answer rewards.
- We introduce **OmniZPD**, an RL-driven framework with a ratio-conditioned unified selector. It inherently captures audio-visual synergies and dynamically adapts to diverse computational constraints at inference without retraining.
- Extensive experiments across multiple omni-modal reasoning benchmarks demonstrate that OmniZPD establishes a new state-of-the-art performance-efficiency trade-off.

2 Related Work

Omni-Modal Models and Reasoning. Omni-modal large language models (OmniLLMs) jointly process audio, video, and text to support understanding across modalities. Representative models, including GPT-4o (Hurst et al., 2024), Baichuan-Omni (Li et al., 2024), and the Qwen-Omni series (Team, 2026), integrate complementary audio-visual evidence for multimodal tasks. Video captures appearance, motion, and spatial relationships, while audio conveys speech, environmental sounds, and vocal characteristics. Combining these signals supports understanding events and interactions that may be ambiguous within a single modality (Liu et al., 2026). Beyond multimodal perception, reasoning-enabled models such as Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025a) generate explicit reasoning traces before producing final responses (Wu et al., 2026; Kulkarni & Fazli, 2026). These models connect observations across modalities through intermediate reasoning steps, motivating the study of both answer accuracy and the evidence supporting the reasoning process. For input compression, preserving the final answer alone may not reflect whether intermediate predictions remain consistent with those under full inputs. OmniZPD studies this problem by learning audio-visual token selection guided by answer feedback and reasoning distribution consistency.

Multimodal Compression and Learned Selection. Token compression reduces multimodal inference costs by removing or merging redundant representations (Shen et al., 2024; Shao et al., 2026; Liu et al., 2025a;b). ToMe (Bolya et al., 2022) merges similar tokens, FastV (Chen et al., 2024) prunes visual tokens using attention scores, and DyCoke (Tao et al., 2025) exploits video redundancy. Recent methods, including OmniZip (Tao et al., 2026), EchoingPixels (Gong et al.), OmniFit (Wang et al.), OmniSIFT (Ding et al., 2026), DASH (Li & Huang, 2026), OmniRefine (Deng et al., 2026), and Omni-Prune (Zhong et al., 2026), extend compression to audio-visual inputs with varying redundancy and information requirements across modalities. A complementary direction learns selection from queries and task feedback. MLLM-VFS (Hu et al., 2025) and Frame-Voyager (Yu et al., 2025) learn query-conditioned frame selection, while TSPO (Tang et al., 2026) and related approaches (Du et al., 2026) optimize discrete sampling policies using downstream rewards. OmniZPD applies policy optimization to fine-grained audio-visual token selection through a shared selector. It combines answer feedback with predictive distribution consistency along student-generated reasoning trajectories, providing distinct training signals for selections with the same answer correctness.

On-Policy Distillation and Self-Distillation. Knowledge distillation transfers teacher behavior to a student (Gu et al., 2024; Feng et al., 2025). On-policy distillation evaluates teacher predictions on student-generated prefixes, aligning supervision with states encountered during student generation (Xu et al., 2025b; Cideron et al., 2025). For example, GKD (Agarwal et al., 2024) trains students on their own generated sequences using teacher distributional feedback. Self-distillation obtains supervision from the model itself without a separately trained teacher (Tan & Hong, 2026; Lu et al., 2025). In on-policy self-distillation, privileged information, such as ground-truth answers or additional context, can inform teacher predictions evaluated on student-generated prefixes (Zhang et al., 2026). This provides feedback on intermediate predictions beyond final-answer correctness. Existing methods primarily use such feedback to improve the student model (Kim et al., 2025; Jang et al., 2025). OmniZPD instead compares full-input and compressed-input predictions along student-generated trajectories, converting their distributional differences into a reward for audio-visual token selection while keeping the language model frozen.

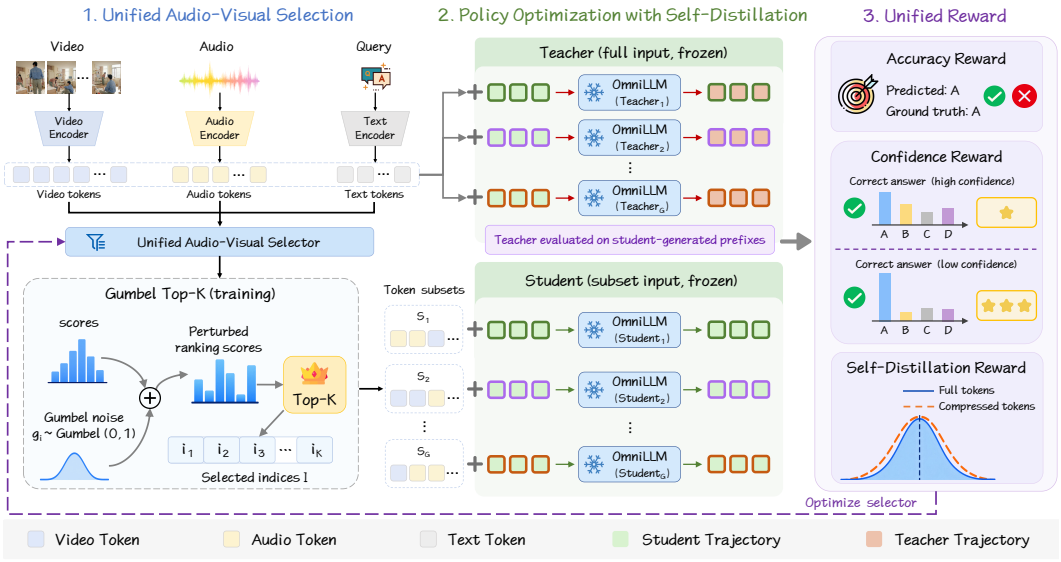

Figure 2: **Overview of OmniZPD.** Given video, audio, and a textual query, a query-conditioned audio-visual token selector samples multiple token subsets via Gumbel Top-K during training. Each subset is rolled out with a frozen OmniLLM, where the student uses the selected subset while the teacher uses the full input and is evaluated on student-generated reasoning prefixes. The resulting rollouts are evaluated using accuracy, confidence, and self-distillation rewards.

3 METHOD

3.1 OVERVIEW.

As shown in Fig. 2, OmniZPD combines a unified audio-visual selector with policy optimization and self-distillation for efficient omni-modal reasoning. Given audio-visual inputs, a textual query, and a retention ratio, the selector scores tokens using textual and temporal cross-modal context. It then selects tokens within modality-specific budgets while preserving the textual input. During training, we sample multiple token selections for the same input and budget, then roll out each selection with the frozen OmniLLM to obtain a reasoning trajectory and final answer. Answer rewards measure correctness and confidence, while self-distillation encourages consistent predictions during reasoning under full and compressed inputs. The selector is updated through policy gradients using relative rewards across candidates. The OmniLLM backbone remains frozen throughout training, and inference requires only a single deterministic token selection followed by response generation.

3.2 UNIFIED AUDIO-VISUAL SELECTOR

As shown in Fig. 3, we design a lightweight selector shared across audio and video to compress multimodal embeddings before they enter the language decoder.

Token Scoring. Let $X^m = \{x_j^m\}_{j=1}^{N_m}$ denote the N_m input embeddings of modality $m \in \{a, v\}$. A shared linear projection maps audio and video features to $z_j^m = W_x x_j^m \in \mathbb{R}^d$. We obtain the textual context $\bar{q}$ by mean-pooling embeddings at non-audio and non-video positions. To capture local cross-modal context c_b, we group tokens into B shared temporal bins using their normalized temporal positions $t_j^m \in [0, 1]$ and aggregate each nonempty bin:

$$c_b = g_{\text{ctx}} \left(\frac{1}{|\mathcal{I}_b|} \sum_{(m,j) \in \mathcal{I}_b} z_j^m \right), \tag{1}$$

where $\mathcal{I}_b$ contains the token indices in bin b, and g_{ctx} consists of LayerNorm, SiLU, and a linear layer. To distinguish token modalities and temporal positions, and to adapt scoring to the target compression budget, we combine the modality, time, and budget encodings into a conditioning vector

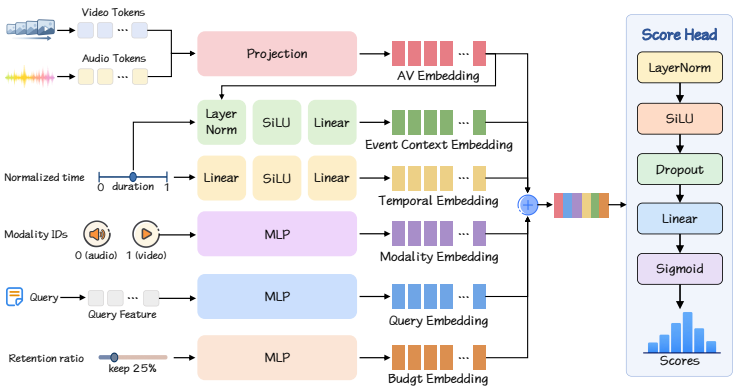

Figure 3: Architecture of the proposed OmniZPD selector. Multi-source token information is fused and fed into a lightweight score head to estimate token importance scores, followed by per-modality normalization and token selection.

$e_j^m \in \mathbb{R}^d$. Modality tags and compression budgets are encoded using learnable embeddings, while temporal positions are encoded by a two-layer MLP with SiLU activation. These three encodings are summed and fused with the token and context features to compute the score:

$$\begin{aligned} \tilde{s}_j^m &= \mathrm{Norm}_m\left(f_{\mathrm{score}}(h_j^m)\right), \\ h_j^m &= z_j^m + W_q \bar{q} + c_{b(j,m)} + e_j^m, \end{aligned} \tag{2}$$

where W_q projects the textual context into $\mathbb{R}^d$ and $b(j,m)$ identifies the token's temporal bin. The shared scoring head f_{score} consists of LayerNorm, SiLU, a linear layer, and sigmoid. Norm_m standardizes scores within modality m, yielding $\tilde{s}_j^m$ for subsequent token selection.

Token Selection. Using the aboved scores $\tilde{s}_j^m$, we select tokens within each modality's budget. Given a target retention ratio $r \in (0,1]$, we retain $K_m = \max\{1, \mathrm{round}(rN_m)\}$ tokens from modality m. During training, we use Gumbel Top-K sampling without replacement (Kool et al., 2019), adding random perturbations to the scores to explore different token combinations under a fixed budget:

$$\omega_m = \mathrm{TopK}\left(\{\tilde{s}_j^m/T + g_j^m\}_{j=1}^{N_m}, K_m\right), \quad g_j^m \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0,1), \tag{3}$$

where $T > 0$ is the sampling temperature and $\omega_m = (j_1^m, \ldots, j_{K_m}^m)$ contains the selected token indices in descending order of perturbed scores.

To optimize the selector through policy gradients, we compute the log-probability of the sampled selection. This sampling procedure follows a Plackett–Luce distribution (Plackett, 1975), giving the joint ordered selection $\omega = (\omega_a, \omega_v)$ the log-probability:

$$\log \pi_\theta(\omega \mid X, \bar{q}, r) = \sum_{m\in\{a,v\}} \sum_{k=1}^{K_m} \log \frac{\exp(\tilde{s}_{j_k^m}^m/T)}{\sum_{j\in\mathcal{R}_k^m} \exp(\tilde{s}_j^m/T)}, \tag{4}$$

where θ denotes the selector parameters, $X = \{X^a, X^v\}$, and $\mathcal{R}_k^m$ is the set of remaining candidates in modality m before the k-th selection. This log-probability is used in subsequent policy updates. The selected audio-visual tokens are passed to the language decoder in their original input order, with all text tokens retained.

3.3 Policy Optimization with Self-Distillation

Reward Design. We combine answer correctness, option confidence, and self-distillation feedback to evaluate how well the selected tokens preserve answer quality and reasoning consistency.

(1) Accuracy Reward. We compare the final answer extracted from the response with the ground-truth answer, assigning $R_i^{\mathrm{acc}} = +1$ for a correct answer and -1 otherwise. This reward encourages the selector to retain tokens that support correct answers.

(2) Confidence Reward. To distinguish rollouts with identical correctness across multiple-choice and open-ended benchmarks, we define a unified confidence reward:

$$R_i^{\text{conf}} = \begin{cases} \tanh\left(\log p_i(y^\star) - \log \sum_{y \neq y^\star} p_i(y)\right), & \text{multiple-choice} \\ \tanh\left(\frac{1}{|S^\star|} \sum_{t=1}^{|S^\star|} \log p_i(s_t^\star \mid s_{<t}^\star)\right), & \text{open-ended} \end{cases} \tag{5}$$

where $\mathcal{Y}$ is the set of candidate options, $y^\star \in \mathcal{Y}$ is the correct option, $S^\star = (s_1^\star, \ldots, s_{|S^\star|}^\star)$ is the ground-truth sequence of length $|S^\star|$, and $p_i(\cdot)$ denotes the predicted probability in the i-th rollout. The $\tanh$ function bounds the reward to $(-1, 1)$ for training stability. This reward encourages retaining tokens that elevate predictive confidence.

(3) Self-Distillation Reward. To assess support for intermediate reasoning, we use the same frozen OmniLLM as the teacher with full inputs and the student with compressed inputs. Both are conditioned on student-generated prefixes, measuring the effect of compression along the student's actual reasoning trajectory, even when the final answer is incorrect. We define a bounded reward using their average KL divergence:

$$R_i^{\text{sd}} = 2 \exp\left(-\frac{1}{\tau_{\text{sd}} L_i} \sum_{t=1}^{L_i} D_{\text{KL}}(p_{\text{T}}^{i,t} \| p_{\text{S}}^{i,t})\right) - 1, \tag{6}$$

where $p_{\text{T}}^{i,t}$ and $p_{\text{S}}^{i,t}$ are the next-token distributions under full and compressed inputs, respectively, given the same prefix $u_{i,<t}$. L_i is the number of reasoning tokens, and $\tau_{\text{sd}} > 0$ controls reward sensitivity. This reward encourages the selector to preserve multimodal evidence that keeps reasoning predictions close to those under full inputs, providing supervision beyond final-answer correctness.

The total reward is

$$R_i = \lambda_{\text{acc}} R_i^{\text{acc}} + \lambda_{\text{conf}} R_i^{\text{conf}} + \lambda_{\text{sd}} R_i^{\text{sd}}, \tag{7}$$

where each component lies in $[-1, 1]$ and the weights λ_{acc}, λ_{conf}, and λ_{sd} control their contributions.

Policy Update. We use the rewards above to optimize the selection policy π_θ. For the same input and retention ratio, we sample $G > 1$ selection sequences $\{\omega_i\}_{i=1}^G$ and compute their rewards R_i. We obtain leave-one-out advantages using the mean reward of the other selections as a baseline (Ahmadian et al., 2024):

$$A_i = R_i - \frac{1}{G-1} \sum_{j \neq i} R_j. \tag{8}$$

We then combine A_i with the selection log-probability $\log \pi_\theta(\omega_i \mid X, \bar{q}, r)$ defined above to form the policy gradient loss:

$$\mathcal{L}_{\text{policy}} = -\frac{1}{G} \sum_{i=1}^{G} \text{sg}(A_i) \frac{\log \pi_\theta(\omega_i \mid X, \bar{q}, r)}{K_a + K_v}, \tag{9}$$

where sg denotes stop-gradient and $K_a + K_v$ is the total number of retained audio-visual tokens used to normalize the log-probability. Gradients flow through the Plackett–Luce log-probability to the token scores, encouraging selections with rewards above the baseline and discouraging those below it. Only the selector is updated; the OmniLLM remains frozen.

3.4 Training and Inference

Training. We train the selector in two stages. First, we initialize it through supervised learning using selection targets produced by a rule-based method. Second, we perform policy optimization using the rewards and leave-one-out advantages, progressively introducing lower retention ratios and reducing the sampling temperature. The OmniLLM remains frozen throughout training.

Inference. At inference time, the same trained selector supports different retention ratios without retraining. Given a target ratio, it performs a single deterministic top-K_m selection within each modality. Selected tokens retain their original order, and all text tokens are preserved. No teacher evaluation or policy sampling is required.

Table 1: Comparison with audio-visual token compression methods on omni-modal reasoning benchmarks. Avg. denotes the average performance retention relative to the Full Tokens model across all evaluated benchmarks. Bold numbers indicate the best results, while underlined numbers indicate the second-best results under each token retention ratio.

Method	Retained	DailyOmni	WorldSense	AVUTBench	VideoMME	OmniVideoBench	Avg.
Full Tokens	100%	75.65	53.67	72.67	74.85	40.63	100.0%
Random	10%	54.38	41.38	52.70	59.44	31.34	75.4%
Uniform-T	10%	55.64	39.86	53.29	59.23	33.45	76.1%
OmniZip	10%	60.42	44.06	57.97	64.03	35.75	82.6%
DASH	10%	58.07	43.72	58.28	63.94	34.36	81.4%
OmniRefine	10%	59.61	44.49	58.16	66.71	36.36	83.6%
OmniPrune	10%	61.58	43.22	56.65	62.29	34.24	81.3%
OmniZPD	10%	**62.73**	**46.27**	**59.65**	**67.10**	**37.01**	**85.9%**
Random	20%	61.57	45.03	59.32	63.47	33.94	82.9%
Uniform-T	20%	60.60	46.36	59.76	62.48	35.78	83.5%
OmniZip	20%	66.15	48.53	62.20	69.04	37.65	89.3%
DASH	20%	65.55	49.03	62.92	68.32	35.87	88.7%
OmniRefine	20%	62.18	48.28	60.99	68.14	38.52	87.6%
OmniPrune	20%	64.29	47.21	60.33	67.52	35.18	86.5%
OmniZPD	20%	**67.78**	**50.35**	**64.07**	**70.33**	**38.96**	**91.8%**
Random	30%	64.93	48.33	65.41	67.99	34.18	88.5%
Uniform-T	30%	66.75	48.78	65.47	67.69	36.84	90.0%
OmniZip	30%	70.45	50.18	69.75	70.63	38.14	94.2%
DASH	30%	69.61	50.99	68.01	69.86	37.77	93.3%
OmniRefine	30%	67.79	51.47	66.37	70.93	39.55	93.3%
OmniPrune	30%	66.72	48.80	67.23	69.31	37.81	91.3%
OmniZPD	30%	**72.02**	**52.36**	**70.13**	**71.89**	**40.37**	**96.6%**

4 EXPERIMENTS

4.1 EVALUATION SETUPS AND IMPLEMENTATION DETAILS

Benchmarks. We evaluate OmniZPD on five representative omni-modal reasoning benchmarks, including DailyOmni (Zhou et al., 2025), WorldSense (Hong et al., 2026), AVUTBench (Yang et al., 2025c), VideoMME (Fu et al., 2025), and OmniVideoBench (Li et al., 2026). These benchmarks cover diverse audio-visual understanding tasks, including video understanding, audio-visual joint reasoning, and complex multimodal question answering, providing a comprehensive evaluation of omni-modal understanding and reasoning capabilities.

Comparison Methods. We compare OmniZPD with existing audio-visual token compression methods, including Random-Global, Uniform-Temporal, OmniZip (Tao et al., 2026), DASH (Li & Huang, 2026), OmniRefine (Deng et al., 2026), and Omni-Prune (Zhong et al., 2026). Random-Global and Uniform-Temporal are adopted as basic token selection baselines to evaluate random selection and temporal sampling strategies. OmniZip, DASH, OmniRefine, and Omni-Prune represent recent token importance estimation-based compression approaches. All methods are evaluated using the same Qwen3-Omni-30B-A3B-Thinking backbone under identical token retention ratios for fair comparison.

Implementation Details. We use Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025a) as the frozen backbone and train a 2.68M-parameter audio-visual selector. Cold-start uses 500 OmniVideo examples (Cai et al., 2026), with token selection targets generated by voting-based aggregation under a fixed token budget over four rule-based importance estimators. The rules and voting procedure are detailed in the appendix. Policy optimization uses 5000 OmniVideo examples and $G = 4$ selections per input. We use at most 16 frames per video. All experiments run on 8 NVIDIA H800 GPUs in bfloat16. We use AdamW (Loshchilov & Hutter, 2017) with learning rates of 1×10^{-4} for supervised initialization and 3×10^{-5} for policy optimization, and train the selector for one epoch.

4.2 Main Results

Main Results. Table 1 compares OmniZPD with representative audio-visual token compression methods across five benchmarks at retention ratios of 10%, 20%, and 30%. OmniZPD achieves the highest accuracy among compression methods on every benchmark at all evaluated budgets, showing consistent advantages across tasks and compression levels. With only 10% of audio-visual tokens retained, OmniZPD preserves 85.9% of the full-input model's mean accuracy, compared with 83.6% for OmniRefine. The improvements also extend across individual benchmarks. Compared with the best baseline on each benchmark, accuracy increases from 44.49% to 46.27% on WorldSense and from 58.28% to 59.65% on AVUTBench. These results show that OmniZPD remains effective when the available token budget is severely constrained. At 20% and 30% retention, OmniZPD preserves 91.8% and 96.6% of the full-input model's mean accuracy, exceeding OmniZip by 2.5 and 2.4 percentage points, respectively. On OmniVideoBench, OmniZPD reaches 40.37% accuracy at 30% retention, only 0.26 percentage points below the full-input result of 40.63%. These results demonstrate a favorable trade-off between token budget and accuracy, retaining 96.6% of full-input mean accuracy while removing 70% of audio-visual tokens.

4.3 Ablation Study

Table 2 presents a progressive ablation of cold-start initialization and reward components at 20% audio-visual token retention. Cold-start initialization improves accuracy over random selection by 1.97, 2.16, and 1.92 percentage points on DailyOmni, WorldSense, and AVUTBench, respectively. Policy optimization with the accuracy reward (Acc) provides additional gains of 2.13, 1.76, and 1.29 percentage points, showing the benefit of downstream task feedback beyond supervised initialization. Adding the confidence reward (Conf) further raises accuracy to 66.93%, 49.50%, and 63.41% on DailyOmni, WorldSense, and AVUTBench, respectively. With the self-distillation reward (SD), the full OmniZPD configuration achieves the best results of 67.78%, 50.35%, and 64.07%, respectively. These consistent improvements show that self-distillation along reasoning trajectories provides useful supervision for token selection beyond answer correctness and confidence.

Table 2: Cumulative reward ablation at 20% token retention. Acc, Conf, and SD denote accuracy, confidence, and self-distillation rewards, respectively.

Configuration	DailyOmni	WorldSense	AVUTBench
Random Selection	61.57	45.03	59.32
Cold Start	63.54	47.19	61.24
+ Acc Reward	65.67	48.95	62.53
+ Conf Reward	66.93	49.50	63.41
+ SD Reward	**67.78**	**50.35**	**64.07**

4.4 Efficiency Analysis

Table 3 compares input length, prefill latency, and accuracy on OmniVideoBench. OmniZPD reduces audio-visual tokens from 10792 to 2159 and total prompt length from 10932 to 2299 tokens. Prefill latency decreases from 1508.5 ms to 540.3 ms, yielding a 2.79× prefill speedup over full inputs. At the evaluated operating points, OmniZPD achieves higher accuracy and lower prefill latency than all reported compression baselines. Compared with OmniZip, it uses 32.5% fewer audio-visual tokens and reduces prefill latency by 12.4%, while improving accuracy from 38.14% to 38.96%. These results demonstrate that OmniZPD preserves task performance more effectively with a smaller token budget.

Table 3: Efficiency comparison on OmniVideoBench.

Method	AV Tokens ↓	Effective Prompt ↓	Prefill Time (ms) ↓	Performance (%) ↑
Full Tokens	10792	10932	1508.5	40.63
Random	3237	3378	626.3	34.18
Uniform-T	3237	3378	620.6	36.84
OmniZip	3200	3341	616.6	<u>38.14</u>
DASH	3327	3468	631.9	37.77
OmniPrune	3238	3378	618.0	37.81
OmniZPD	**2159**	**2299**	**540.3**	**38.96**

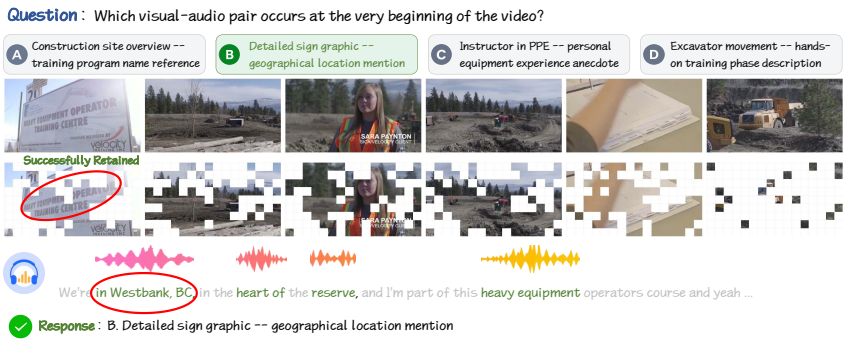

Figure 4: Qualitative visualization of OmniZPD token selection. White patches indicate discarded visual tokens, and gray transcript text marks discarded audio segments. OmniZPD retained query-relevant and complementary audio-visual evidence and removed redundant information.

4.5 Generalization Performance Analysis

To evaluate cross-architecture generalizability, we train the OmniZPD selector on MiniCPM-o 4.5(Cui et al., 2026). As shown in Table4, at a 30% audio-visual token retention rate, OmniZPD consistently outperforms all baselines across three benchmarks. Notably, it surpasses the strongest competitor, OmniZip, by absolute margins of +1.62% (73.61% vs. 71.99%) on DailyOmni and +1.55% (69.49% vs. 67.94%) on AVUTBench. Furthermore, on WorldSense, OmniZPD achieves 51.05%, narrowing the performance gap to the uncompressed full-token baseline (51.94%) to less than 1%. These results confirm that our self-distillation reward is model-agnostic, highlighting OmniZPD as a universally applicable optimization pipeline for diverse multimodal large language models.

Table 4: Comparison on MiniCPM-o 4.5 at 30% audio-visual token retention. Full Tokens is the uncompressed reference. Bold and underlined values mark the best and second-best results among compression methods.

Method	Retained	DailyOmni	WorldSense	AVUTBench
Full Tokens	100%	77.53	51.94	71.72
Random	30%	67.35	47.01	64.59
OmniZip	30%	<u>71.99</u>	48.83	<u>67.94</u>
DASH	30%	71.10	49.24	66.80
OmniRefine	30%	69.72	<u>50.35</u>	65.77
OmniZPD	30%	**73.61**	**51.05**	**69.49**

4.6 Qualitative Analysis

Fig. 4 illustrates OmniZPD's token selection for a question about the beginning of a video. OmniZPD retains visual patches containing the signboard ("HEAVY EQUIPMENT OPERATOR TRAINING CENTRE") while discarding many background patches. In the audio stream, it preserves the location mention ("in Westbank, BC") while removing less relevant audio tokens. Together, the retained sign and spoken location provide complementary evidence for option **B**. This case demonstrates that OmniZPD effectively suppresses modality-wise redundancy while preserving reasoning-critical evidence to guarantee reliable prediction. More visualizations in Appendix.

5 Conclusion

In conclusion, we present **OmniZPD**, an RL-driven token compression framework designed to accelerate Thinking OmniLLMs without sacrificing reasoning fidelity. Unlike heuristic pruning methods, OmniZPD employs a unified, ratio-conditioned selector optimized via a combination of task rewards and self-distillation feedback. This design not only dynamically balances audio-visual token budgets across modalities but also enforces consistent intermediate reasoning trajectories between full and compressed inputs. Empirical evaluations across multiple omni-modal benchmarks confirm the superiority of our approach. Notably, OmniZPD delivers a substantial 2.79× prefill speedup and retains 91.8% of the original accuracy using only 20% of the audio-visual tokens, offering a practical and scalable solution for efficient omni-modal reasoning.

AI USE STATEMENT

Generative AI tools were used only as auxiliary tools for literature search and summarization, language editing, figure preparation, reference organization, and selected software-development tasks. The research questions, core methodology, model architecture, reward design, experimental setup, analysis, and conclusions were developed by the authors. AI-assisted outputs were manually reviewed and verified before inclusion in the manuscript. The authors take full responsibility for all content, results, and scientific claims presented in this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. Omnivideo-100k: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026.

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 *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Geoffrey Cideron, Andrea Agostinelli, Johan Ferret, Sertan Girgin, Romuald Elie, Olivier Bachem, Sarah Perrin, and Alexandre Ramé. Diversity-rewarded cfg distillation. In *International Conference on Learning Representations*, volume 2025, pp. 66277–66299, 2025.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Yuchen Deng, Zidang Cai, Hai-Tao Zheng, Jie Wang, Feidiao Yang, and Yuxing Han. Omnirefine: Alignment-aware cooperative compression for efficient omnimodal large language models. *arXiv preprint arXiv:2605.12056*, 2026.

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. Appo: Attention-guided perception policy optimization for video reasoning. *arXiv preprint arXiv:2602.23823*, 2026.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4178–4188. IEEE, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Chao Gong, Depeng Wang, Zhipeng Wei, Ya Guo, Huijia Zhu, and Jingjing Chen. Echoingpixels: Aliasing-resistant joint token reduction for audio-visual llms. In *Forty-third International Conference on Machine Learning*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712. IEEE, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jinseong Jang, Chunfei Ma, and Byeongwon Lee. Vl2lite: Task-specific knowledge distillation from large vision-language models to lightweight networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30073–30083. IEEE, 2025.

Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. Cosmos: Cross-modality self-distillation for vision language pre-training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14690–14700. IEEE, 2025.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7912–7922, 2026.

Hosu Lee, Junho Kim, Hyunjun Kim, and Yong Man Ro. Refocus: Reinforcement-guided frame optimization for contextual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8291–8302, 2026.

Bingzhou Li and Tao Huang. Dash: Dynamic audio-driven semantic chunking for efficient omnimodal token compression. *arXiv preprint arXiv:2603.15685*, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pp. 138214–138236, 2026.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024.

Naiming Liu, Zhiheng Wu, Shuning Wang, Tie Zhang, Bowen Liu, and Tong Wang. Ovibench: Benchmarking online video question answering under interruption. *arXiv preprint arXiv:2608.22279*, 2026.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025a.

Zhihang Liu, Chen-Wei Xie, Pandeng Li, Liming Zhao, Longxiang Tang, Yun Zheng, Chuanbin Liu, and Hongtao Xie. Hybrid-level instruction injection for video token compression in multi-modal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8568–8578. IEEE, 2025b.

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

Zhicong Lu, Changyuan Tian, Peiguang Li, Li Jin, Sirui Wang, Wei Jia, Ying Shen, and Guangluan Xu. Piper: Benchmarking and prompting event reasoning boundary of llms via debiasing-distillation enhanced tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28591–28613, 2025.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *Advances in Neural Information Processing Systems*, 38:135547–135570, 2026.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Dachuan Shi, Chaofan Tao, Anyi Rao, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers. *arXiv preprint arXiv:2305.17455*, 2023.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. *arXiv preprint arXiv:2502.11775*, 2025.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models. *arXiv preprint arXiv:2605.17497*, 2026.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. Tspo: Temporal sampling policy optimization for long-form video language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9368–9376, 2026.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001. IEEE, 2025.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17682–17692, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Hengyi Wang, Weiying Xie, Hui Jiang, Yaotao Wei, Kai Jiang, Mingxiang Cao, Chenhe Hao, and Leyuan Fang. Top-rl: Task-optimized progressive token pruning with reinforcement learning for vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15824–15832, 2026a.

Shida Wang, YongXiang Hua, Zhou Tao, Haoyu Cao, and Linli Xu. Dynamic token compression for efficient video understanding through reinforcement learning. *arXiv preprint arXiv:2603.26365*, 2026b.

Zining Wang, Zhihang Yuan, Yingjie Zhai, Wenshuo Li, Han Shu, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Omnifit: Bridging modalities via layer-adaptive token compression for omni-modal large language models. In *Forty-third International Conference on Machine Learning*.

Zhiheng Wu, Tong Wang, Shuning Wang, Naiming Liu, and Yumeng Zhang. See further, think deeper: Advancing vlm's reasoning ability with low-level visual cues and reflection. *arXiv preprint arXiv:2604.24339*, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025a.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025b.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19803–19813. IEEE, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025b.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6580–6598, 2025c.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24972–24982. IEEE, 2025.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pp. 84154–84179, 2025.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025.

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

Yiming Zhong, Chang Nie, and Caifeng Shan. Omni-prune: Query-aware unified token pruning for efficient omnimodal large language models. *arXiv preprint arXiv:2607.23445*, 2026.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Cold-Start Rule-Based Initialization

Before policy optimization, we initialize the token selector using pseudo-labels generated by a rule-based procedure. Specifically, the selection results of four existing audio-visual token compression methods, OmniZip, DASH, OmniRefine, and OmniPrune, are aggregated to construct token-level selection targets for cold-start training.

For modality $m \in \{a, v\}$, where a and v denote audio and video, respectively, we denote the candidate tokens by $X^m = \{x_j^m\}_{j=1}^{N_m}$, with N_m being the number of tokens in modality m. Given a target retention ratio r, the corresponding token budget is

$$K_m = \max\{1, \min(N_m, \mathrm{round}(rN_m))\}, \tag{10}$$

which specifies the number of tokens retained from modality m.

Each baseline method b is evaluated under the same K_m-token budget. For the j-th token of modality m, its binary keep indicator is defined as

$$y_{j,b}^m \in \{0, 1\}, \tag{11}$$

where $y_{j,b}^m = 1$ indicates that method b retains the token and $y_{j,b}^m = 0$ otherwise. For methods that natively use an adaptive token budget, the selection result is adjusted to the common K_m-token budget according to the method-specific token-importance ordering before pseudo-label construction. This ensures that all methods contribute their selections under the same compression budget.

The number of baseline methods retaining token j is computed as

$$c_j^m = \sum_{b \in \mathcal{B}} y_{j,b}^m, \qquad \mathcal{B} = \{\text{OmniZip}, \text{DASH}, \text{OmniRefine}, \text{OmniPrune}\}. \tag{12}$$

The resulting vote count satisfies $c_j^m \in \{0, 1, 2, 3, 4\}$, with larger values indicating stronger agreement among the baseline compression methods that token j should be retained.

To obtain a deterministic ordering when multiple tokens receive the same vote count, we use an auxiliary tie-breaking score q_j^m. When continuous token-importance scores are available, the scores produced by each baseline are first normalized and then summed to obtain q_j^m. If such scores are unavailable, we set $q_j^m = 0$ and use the original token index j as the final deterministic tie breaker.

All candidate tokens in modality m are ranked using the lexicographic key $(c_j^m, q_j^m, -j)$, and the top K_m tokens form the cold-start pseudo-label set

$$S_m^{\mathrm{cold}} = \mathrm{TopK}_{j \in \{1,\ldots,N_m\}}\big(c_j^m, q_j^m, -j; K_m\big), \qquad |S_m^{\mathrm{cold}}| = K_m. \tag{13}$$

This procedure aggregates the selections of multiple existing compression methods into a unified supervision signal under a fixed token budget.

The resulting pseudo-labels are used to initialize the OmniZPD selector through supervised training. Within each modality, tokens in S_m^{cold} are treated as positive samples, while the remaining tokens are treated as negative samples. During training, positive-negative token pairs are randomly sampled and optimized with a pairwise ranking objective:

$$\mathcal{L}_{\mathrm{cold}} = \mathrm{mean}\left[\max\left(0,\, \delta - \tilde{s}_p^m + \tilde{s}_n^m\right)\right], \tag{14}$$

where $\tilde{s}_p^m$ and $\tilde{s}_n^m$ denote the selector scores of a positive and a negative token, respectively, and δ is the ranking margin. This objective encourages the selector to assign higher importance scores to tokens selected by the pseudo-labeler, providing a meaningful initialization before policy optimization. During cold-start training, only the selector parameters are updated, while the OmniLLM backbone remains frozen.

All text tokens are preserved throughout cold-start initialization and are not subject to compression. The selected audio and video tokens are restored to their original temporal order before being passed to the frozen OmniLLM. This stage provides a stable initialization for the selector and avoids starting policy optimization from a completely random token-selection policy.

B Policy Optimization with Self-Distillation

Algorithm 1 summarizes the complete training procedure of OmniZPD, including token-subset sampling, reward computation, and RLOO-based policy optimization.

Algorithm 1 OmniZPD Policy Optimization with Self-Distillation

Require: Training set $\mathcal{D}$, frozen OmniLLM $\mathcal{M}$, initialized selector π_θ, group size G
Ensure: Optimized selector π_θ
for each training step s **do**
Sample an example $(X, q, y^\star) \sim \mathcal{D}$
$r \leftarrow \mathrm{RetentionSchedule}(s), \quad T \leftarrow \mathrm{TemperatureSchedule}(s)$

Rollout and Reward Computation
for $i = 1, \ldots, G$ **do**
Sample ordered token selections $\omega_i = (\omega_i^a, \omega_i^v)$ using modality-wise Gumbel Top-K.
$X_i^c \leftarrow \mathrm{Compress}(X, \omega_i)$
$(u_i, \hat{y}_i) \leftarrow \mathcal{M}(X_i^c, q)$

Answer correctness reward
$$R_i^{\mathrm{acc}} \leftarrow \begin{cases} +1, & \hat{y}_i = y^\star, \\ -1, & \text{otherwise} \end{cases}$$

Answer confidence reward
$$\Delta_i \leftarrow \log p_i(y^\star) - \log \sum_{y \neq y^\star} p_i(y)$$
$$R_i^{\mathrm{conf}} \leftarrow \tanh(\Delta_i)$$

Self-distillation reward
Extract the reasoning prefix $u_i = (u_{i,1}, \ldots, u_{i,L_i})$ before the final-answer marker
for $t = 1, \ldots, L_i$ **do**
$p_{i,t}^T \leftarrow \mathcal{M}(\cdot \mid X, q, u_{i,<t})$ ▷ full input
$p_{i,t}^S \leftarrow \mathcal{M}(\cdot \mid X_i^c, q, u_{i,<t})$ ▷ compressed input
end for
$$D_i \leftarrow \frac{1}{L_i} \sum_{t=1}^{L_i} D_{\mathrm{KL}}\left(p_{i,t}^T \,\|\, p_{i,t}^S\right)$$
$$R_i^{\mathrm{sd}} \leftarrow 2\exp(-D_i/\tau_{\mathrm{sd}}) - 1$$

Unified reward
$$R_i \leftarrow \lambda_{\mathrm{acc}} R_i^{\mathrm{acc}} + \lambda_{\mathrm{conf}} R_i^{\mathrm{conf}} + \lambda_{\mathrm{sd}} R_i^{\mathrm{sd}}$$
end for

RLOO Policy Update
for $i = 1, \ldots, G$ **do**
$$A_i \leftarrow R_i - \frac{1}{G-1} \sum_{j \neq i} R_j$$
Replay the sampled action ω_i and compute $\log \pi_\theta(\omega_i \mid X, q, r)$
end for
$$\mathcal{L}_{\mathrm{policy}} \leftarrow -\frac{1}{G} \sum_{i=1}^{G} \mathrm{sg}(A_i) \frac{\log \pi_\theta(\omega_i \mid X, q, r)}{K_a + K_v}$$
$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{policy}}$
end for
return π_θ

C Additional Generalization Results on MiniCPM-o 4.5

We further evaluate the cross-backbone generalization of OmniZPD on MiniCPM-o 4.5 at 10% and 20% audio-visual token retention ratios. As shown in Table 5, OmniZPD consistently achieves the best performance among all compressed methods on DailyOmni, WorldSense, and AVUTBench at both retention ratios. The advantage is preserved under both aggressive and moderate compression, indicating that the learned token selection strategy remains effective across different token budgets after adaptation to a different omni-modal backbone. These results further support the robustness of OmniZPD in cross-backbone settings. Bold and underlined values denote the best and second-best compressed results, respectively.

Table 5: Comparison on MiniCPM-o 4.5 under different audio-visual token retention ratios.

Method	Retained	DailyOmni	WorldSense	AVUTBench
Random	10%	55.35	40.21	52.14
OmniZip	10%	61.04	43.52	56.70
DASH	10%	59.61	42.89	57.49
OmniRefine	10%	60.46	43.70	57.02
OmniZPD	10%	**64.08**	**44.98**	**58.62**
Random	20%	62.38	43.78	59.05
OmniZip	20%	66.72	47.62	61.64
DASH	20%	65.96	48.15	61.81
OmniRefine	20%	63.64	47.79	60.21
OmniZPD	20%	**67.97**	**49.01**	**63.27**

D Unified Audio-Visual Selector Algorithm

We provide the complete procedure of the unified audio-visual selector in Algorithm 2. The selector jointly incorporates token content, query semantics, modality identity, temporal context, and the target retention ratio to estimate token importance. Audio and video tokens are then selected independently under modality-specific budgets, using stochastic Gumbel Top-K sampling during training and deterministic Top-K selection at inference.

Algorithm 2 Unified Audio-Visual Token Selection

Require: X^a, X^v: audio/video tokens; $\bar{q}$: textual context; t: normalized temporal positions; r: retention ratio
Ensure: Selected tokens $\mathcal{S}$; normalized scores $\tilde{s}$

$X \leftarrow \mathrm{Concat}(X^a, X^v)$
Record modality m_j, normalized time t_j, and original index of each $x_j \in X$

Stage I: Context-Aware Token Scoring
for each token $x_j \in X$ **do**
  $z_j \leftarrow \mathrm{TokenProj}(x_j), \quad e_j^m \leftarrow \mathrm{ModalityEmbed}(m_j), \quad e_j^t \leftarrow \mathrm{TimeProj}(t_j)$
end for
$q' \leftarrow \mathrm{QueryProj}(\bar{q}), \qquad e^r \leftarrow \mathrm{BudgetEmbed}(r)$
Partition tokens into shared temporal bins $\{\mathcal{I}_b\}_{b=1}^{B}$ according to t_j
for each non-empty temporal bin b **do**
  $c_b \leftarrow \mathrm{ContextProj}\Big(|\mathcal{I}_b|^{-1} \sum_{j \in \mathcal{I}_b} z_j\Big)$
end for
for each token $x_j \in X$ **do**
  $h_j \leftarrow z_j + q' + c_{b(j)} + e_j^m + e_j^t + e^r$
  $s_j \leftarrow \sigma(\mathrm{ScoreHead}(h_j))$
end for
$\tilde{s}^{\,m} \leftarrow \mathrm{Norm}_m(s^m), \quad m \in \{a, v\}$

Stage II: Modality-Wise Token Selection
for $m \in \{a, v\}$ **do**
  $K_m \leftarrow \max\{1, \mathrm{round}(rN_m)\}$
  if training **then**
    $g_j^m \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(0,1); \quad \hat{s}_j^m \leftarrow \tilde{s}_j^m / T + g_j^m$
    $\omega_m \leftarrow \mathrm{TopK}(\hat{s}^m, K_m)$
    Compute $\log \pi_m(\omega_m)$ via the Plackett–Luce likelihood
  else
    $\omega_m \leftarrow \mathrm{TopK}(\tilde{s}^m, K_m)$
  end if
end for
$\mathcal{I} \leftarrow \mathrm{Concat}(\omega_a, \omega_v)$
$\mathcal{S} \leftarrow \mathrm{RestoreOriginalOrder}(\mathcal{I})$
return $\mathcal{S}, \tilde{s}$

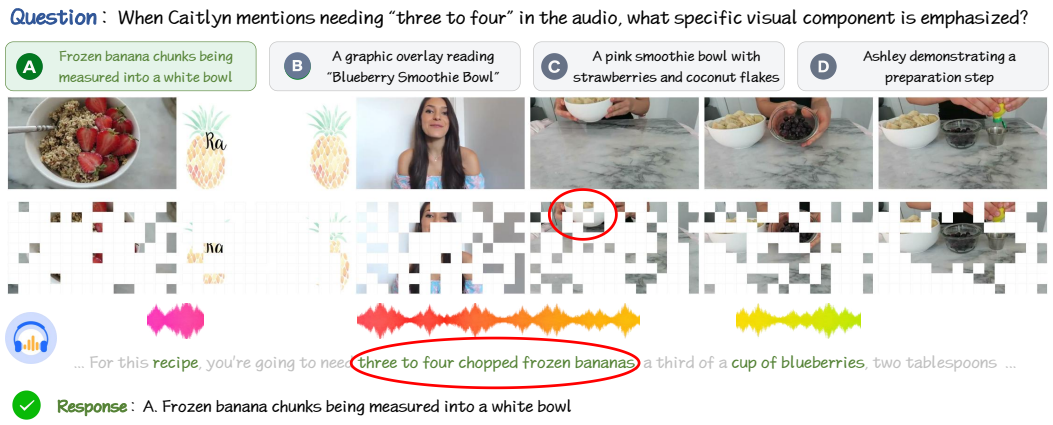

Figure 5: Cross-modal evidence preservation guided by audio cues. When the question asks which visual content corresponds to the phrase "three to four" in the audio, OmniZPD preserves the key speech segment "three to four chopped frozen bananas" together with the temporally aligned visual regions containing the frozen bananas and the white bowl. Although substantial irrelevant audio-visual tokens are removed, the retained cross-modal evidence remains sufficient to support the correct answer A.

E Additional Qualitative Visualizations

Figure 5 illustrates how OmniZPD preserves query-relevant cross-modal evidence. In this example, the question requires the model to associate the phrase "three to four" in the audio with the corresponding visual content. OmniZPD retains the key speech segment "three to four chopped frozen bananas" together with the temporally aligned visual regions containing the frozen bananas and the white bowl. Although a large number of unrelated audio and visual tokens are removed, the retained evidence still provides a clear cross-modal correspondence and is sufficient to support the correct answer. This case shows that OmniZPD can preserve complementary information across modalities according to the query, rather than relying only on saliency within a single modality.

Figure 6 demonstrates the ability of OmniZPD to preserve temporally relevant visual evidence. Unlike a static recognition task, this question requires identifying the first visual event that appears after a specified textual cue disappears. OmniZPD concentrates the available token budget around the relevant temporal position and preserves the major visual regions containing the apartment buildings, while removing substantial redundant content from neighboring frames. The retained visual evidence is sufficient to support the correct answer. This example suggests that OmniZPD can prioritize information with temporal discriminative value when the query depends on event ordering.

Figure 7 presents a complementary failure case caused by incomplete input evidence. The question asks which audio event coincides with the appearance of cartoon sharks and therefore requires audio information for reliable prediction. However, no usable audio evidence is provided in the original sample, meaning that even the full-token input lacks the information required to answer the question reliably. Although the final prediction is incorrect, OmniZPD still preserves informative visual regions containing the main characters, their interactions, and the surrounding scene. This indicates that the failure mainly arises from missing source-modality information rather than from discarding critical visual evidence during token compression.

Overall, these three cases illustrate OmniZPD from complementary perspectives, including cross-modal evidence alignment, temporal event localization, and failure under incomplete multimodal input. The first two cases show that OmniZPD can retain compact but task-relevant evidence under a limited token budget, while the third highlights a practical limitation when the original input itself lacks the modality required by the question. Together, these examples provide a more complete view of the token-selection behavior of OmniZPD in both successful and failure cases.

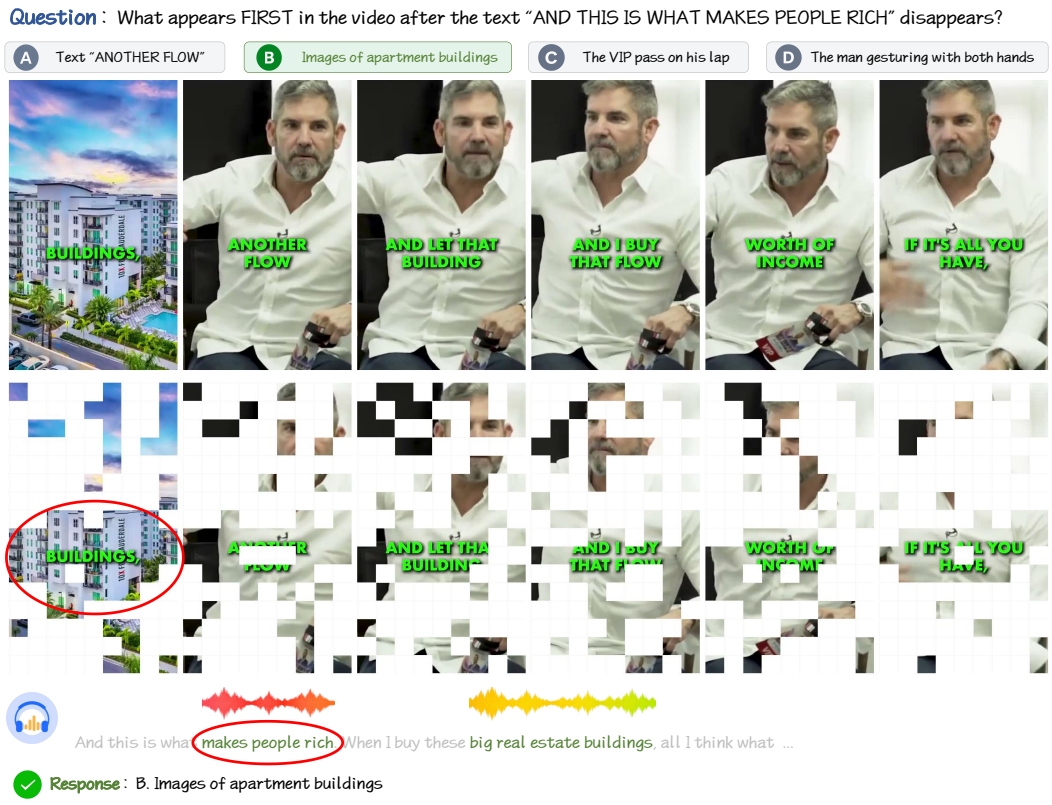

Figure 6: Preservation of key visual evidence for temporal reasoning. For the temporal-order question asking what appears first after the specified text disappears, OmniZPD concentrates the limited token budget on the relevant time point, preserving the main visual regions containing the apartment buildings while removing substantial redundant content from neighboring frames. The retained information is sufficient to support the correct answer B.

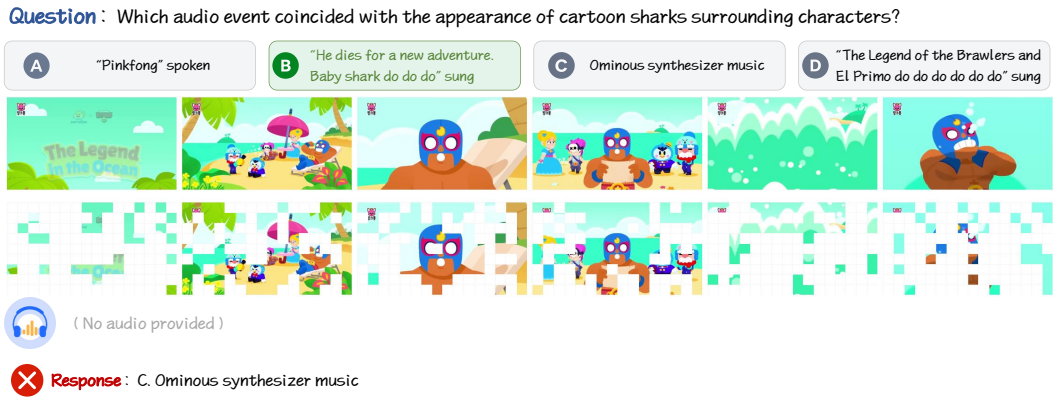

Figure 7: Failure case under incomplete multimodal evidence. The question requires identifying the audio event associated with the appearance of cartoon sharks, while the original sample provides no usable audio evidence. As a result, even the full-token input is insufficient for reliable prediction. Despite the incorrect answer, OmniZPD still preserves visually informative regions containing the main characters and scene context, indicating that the failure is primarily caused by missing source information rather than the loss of critical visual evidence during compression.