MAXMOREGAZE: MAX-LOSS AND MORE-SALIENT GAZING FOR EFFICIENT VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient video understanding must preserve spatiotemporal features within a limited visual-token budget. Reducing the frame rate discards temporal structure, and lowering resolution removes small objects, text, and other local details. Reconstruction-guided selectors acquire multi-scale patches before full vision encoding, yet a frame-average reconstruction target conceals local failures and yields unstable stopping. We present **MaxMoreGaze**. **MaxGaze** replaces that scalar with a bias-corrected 7×7 reconstruction-error map: the same map ranks the next patch, constructs the training trajectory, and decides when to stop, namely when every spatial window meets the target. **MoreGaze** then uses gray-calibrated cross-scale attention from the frozen vision encoder to prune weak fine-scale patches in a training-free manner. Retained patches are serialized with explicit tile and frame structure. At matched patch budgets, MaxGaze improves local reconstruction and stopping. On seven video-understanding settings with identical frames and tiles, MaxMoreGaze improves the performance at a lower token cost than the previous method, and is stronger on detail-intensive high-resolution benchmarks.

1 INTRODUCTION

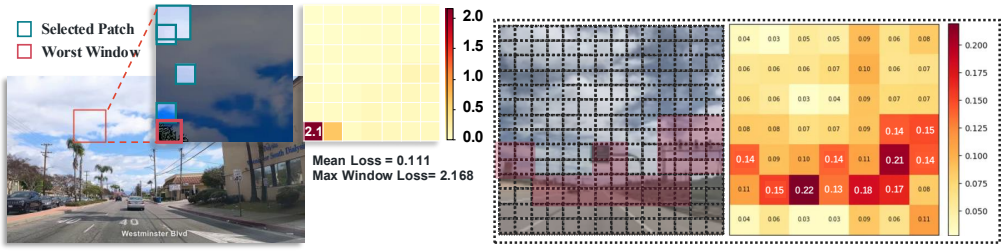
For video understanding tasks, MLLMs convert sampled sparse frames into visual tokens and concatenate them with text prompts for language-model reasoning. Both training and Inference bost grows quickly with video duration and resolution. Efficient video understanding must preserve spatiotemporal features within a limited visual-token budget. Reducing the frame rate discards temporal structure, and lowering resolution removes small objects, textual information, and other local details.

Prior work reduces this cost by sampling frames, merging tokens, or pruning after dense encoding (Xu et al., 2024a; Yu et al., 2024; Shen et al., 2024; Sun et al., 2025a; Park et al., 2026). A reconstruction-guided alternative Shi et al. (2026) selects multi-scale patches *before* full vision encoding. A lightweight autoregressive gazing controller chooses the next patch id from a multi-scale patch vocabulary given the recent frames and the prefix patch ids already taken. A frozen reconstruction model, typically a finetuned VideoMAE (Tong et al., 2022), scores each prefix during training; at inference only the controller runs. It ranks patches by how much they reduce *frame-average* reconstruction error, and it stops its sequence generation from a scalar estimate of that same mean.

Although content-adaptive, this shared mean-error objective weakens both patch ordering and stopping. A small, poorly reconstructed region contributes little to the frame average and can therefore remain unnoticed, as illustrated in Fig. 1(a). Meanwhile, predicting the full reconstruction state with one simple scalar is imprecise and vague, causing some frames to stop too early and others to collect too many unnecessary patches. The supervision itself is also biased: reconstruction error reflects not only unobserved content but also limitations of the learned decoder. As shown in Fig. 1(b), high error can remain in several complex patch regions even when all patches are available. An intrinsically difficult region may therefore offer only a small measured improvement, causing a mean-loss teacher to treat it as unrewarding and repeatedly prioritize easier regions.

Reconstruction quality is also not identical to usefulness for video understanding. Fine texture or background motion can lower reconstruction error while attracting little attention from the frozen vision encoder. These patches still occupy model compute. Moreover, directly flattening sparse patches from multiple tiles and frames obscures their spatial grouping and temporal order.

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.



(a) Mean loss hides Local Failures. (b) VideoMAE Bias

Figure 1: **Failure modes of frame-average reconstruction supervision.** (a) A low frame-average error hides a poorly reconstructed local window and leads to premature stopping. (b) VideoMAE retains high error even when all patches are available, showing that raw reconstruction error contains bias from the decoder.

We introduce **MaxMoreGaze**, which improves reconstruction-guided trajectories and stopping before removing weak semantic redundancy. **MaxGaze** replaces the scalar estimate with a 7×7 reconstruction map conditioned on the current frame and acquisition state. It prioritizes the worst remaining region and stops only when all regions meet the target. To reduce reconstruction-model bias, we correct every training map using the content revealed by selected patches and enforce non-increasing error along each prefix. A max-window teacher uses these corrected maps to construct patch trajectories, while high-loss-aware map regression discourages underestimated local failures. The resulting supervision improves patch ordering and stopping together. **MoreGaze** then removes semantically weak fine patches without additional training. It uses cross-scale attention already computed by the frozen multi-scale vision encoder and subtracts a gray-image response to calibrate positional bias. Coarse patches are preserved, while weak patches at the two finest scales are filtered. We serialize the retained features with explicit tile and frame boundaries and timestamps before MLLM inference. MaxGaze acquisition and MoreGaze scoring are independent, allowing both to run in parallel and merge before cross-frame encoding.

We evaluate MaxMoreGaze at the acquisition and downstream levels. At matched patch budgets, MaxGaze improves local reconstruction over the mean-error baseline. Moreover, MaxMoreGaze performs better than the previous method across 7 video-understanding benchmarks while using fewer visual tokens, and is substantially stronger on detail-intensive high-resolution benchmarks.

Our main contributions are:

- We show how frame-average supervision, scalar loss prediction, and reconstruction-model bias jointly weaken adaptive patch ordering and stopping.
- We introduce MaxGaze, which learns corrected spatial error maps to prioritize the worst remaining region and to supervise both trajectory construction and stopping.
- We introduce MoreGaze, a training-free, query-independent prune of fine-scale patches by encoder salience, after reconstruction-guided acquisition.
- We show matched-budget gains in local reconstruction and stopping, and token-saving accuracy gains versus baseline on seven controlled settings.

2 Related Work

Adaptive video sampling and token allocation. To fit long videos into a limited context, prior work selects frames by multi-rate sampling, query relevance, or inter-frame diversity (Xu et al., 2024a; Yu et al., 2024; Wang et al., 2024b; Sun et al., 2025b; Chen et al., 2026). Other methods compress the retained stream through similarity filtering, attention-based pruning, semantic merging, temporal pooling, or uncertainty-guided allocation (Shen et al., 2024; Sun et al., 2025a; Park et al., 2026; Li et al., 2026; Qi et al., 2026). Codec- and trajectory-based representations further exploit motion, residuals, and object tracks to avoid re-encoding unchanged RGB content (Sarkar et al., 2026; Tang et al., 2026; An et al., 2026; Hou et al., 2026; Yao et al., 2026; Zheng et al., 2026). These approaches decide where to spend a given budget, but usually at the frame, clip, or P-frame level. Static videos can still be over-encoded and dynamic videos under-represented. Query

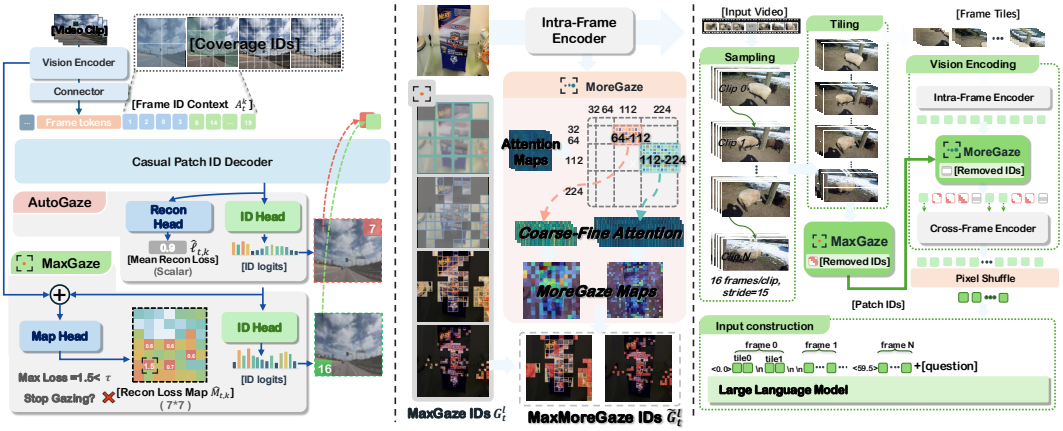


Figure 2: **Overall architecture of MaxMoreGaze.** (a) MaxGaze: Instead of predicting a scalar average reconstruction loss, we predict a dense reconstruction loss map. By incorporating early-frame visual features via residual connections, the model prioritizes regions with high reconstruction error when selecting subsequent patches. (b) MoreGaze: We extract attention maps from the intra-frame encoder and identify semantic regions ranging from coarse to fine granularity. By filtering out patches with negligible semantic contributions, MoreGaze improves the semantic density of the input sequence in a training-free manner. (c) Input Construction: We partition sampled video frames into independent clips and multiple tiles. MaxGaze selects multi-scale reconstruction-critical patches, while MoreGaze retains semantically dense patches to construct the final input sequence. Region segmentation cues and temporal awareness are explicitly integrated during this process.

scores, attention, and answer uncertainty also need not preserve the spatiotemporal content required to reconstruct the video.

Multi-granularity visual patching. High-resolution MLLMs partition images into aspect-ratio-aware tiles and assign more tokens to larger inputs (Xu et al., 2024b; Li et al., 2024; Chen et al., 2024b; Wang et al., 2024a; Wu et al., 2024). Multi-scale methods encode resized or tiled views at several resolutions and then compress features for the language model (Shi et al., 2024; Liu et al., 2024), while selective high-resolution methods keep a global view and spend extra computation on salient local regions (Shi et al., 2025; 2026). Most tiling schemes follow image geometry rather than content and densely encode every tile. Fixed scales also duplicate overlapping coarse and fine content without testing whether a region is already reconstructed well enough. AutoGaze is the closest content-adaptive selector, but its scalar average reconstruction loss can hide local failures, requires a manually chosen gazing ratio in practice, and may keep visually complex yet weakly semantic patches. We instead stop from the worst remaining local reconstruction error and then remove semantically weak fine patches, so granularity and token count follow the content.

3 Method

3.1 Preliminaries: AutoGaze

Given a video clip $X = \{x_t\}_{t=0}^{T-1}$, AutoGaze (Shi et al., 2026) autoregressively selects multi-scale patch indices before expensive vision encoding. Let $\mathcal{P}_t$ be the patch vocabulary for frame t, $a_{t,k} \in \mathcal{P}_t$ its k-th selected patch, and $A_t^k = (a_{t,1}, \ldots, a_{t,k})$ the corresponding prefix. At state $s_{t,k} = (X_{\leq t}, A_{<t}, A_t^k)$, the ID head predicts $a_{t,k+1}$, while the loss head estimates the mean reconstruction error $\widehat{\ell}_{t,k}$ of the current frame under the causal context. Patch selection for frame t ends when $\widehat{\ell}_{t,k} \leq \tau$, where τ is the target reconstruction error threshold.

During training, a frozen VideoMAE (Tong et al., 2022) evaluates each selected prefix. The greedy teacher appends the candidate patch that maximally reduces the frame-average reconstruction error. The policy is initialized with next-token prediction (NTP) and refined with group-relative policy optimization (GRPO), while reconstruction-loss regression trains the stopping estimate. Thus, the

same scalar mean error supervises both patch ordering and stopping. As discussed in Fig. 1, this scalar hides local failures and its raw VideoMAE target contains decoder bias. MaxGaze therefore replaces it with a corrected spatial reconstruction target.

3.2 MAXGAZE

We maintain a 7×7 map $M_{t,k}$ whose entry j is the reconstruction error of spatial window j after prefix A_t^k. The controller predicts $\widehat{M}_{t,k}$ and acts on $\max_j \widehat{M}_{t,k,j}$ rather than on a frame mean.

Map head. A lightweight encoder produces frame features. They are pooled to 7×7 cells $v_{t,j}$. A map head combines each cell with the causal state $z_{t,k}$ of the ID decoder:

$$\widehat{M}_{t,k,j} = f_{\text{map}}([z_{t,k}; v_{t,j}]). \tag{1}$$

The ID head is unchanged. Stopping uses the predicted worst window: selection on frame t ends when $\max_j \widehat{M}_{t,k,j} \le \tau$.

Corrected target. Let $\widetilde{x}_{t,k}$ be the VideoMAE reconstruction of x_t from prefix A_t^k. Raw $|\widetilde{x}_{t,k} - x_t|$ still charges decoder error inside regions that have already been observed. We define a whole-patch paste $\mathsf{Paste}_{\text{wp}}(\widetilde{x}_{t,k}, x_t, A_t^k)$ as follows: for each selected patch $a \in A_t^k$, copy the scale-faithful observation of a from x_t onto the full spatial support of a in $\widetilde{x}_{t,k}$, but only if that copy has lower regional ℓ_1 error than the reconstruction on the same support. No individual pixel of x_t is written unless it belongs to a selected patch, so the operator is not a pixelwise oracle. A coarse patch keeps the error due to downsampling; a finest-scale patch can match x_t locally.

Let $\mathsf{Pool}_{7\times7}$ average-pool pixel ℓ_1 onto the 7×7 grid. The training map is the elementwise cumulative minimum

$$M_{t,k} = \min\big(M_{t,k-1},\ \mathsf{Pool}_{7\times7}\big|x_t - \mathsf{Paste}_{\text{wp}}(\widetilde{x}_{t,k}, x_t, A_t^k)\big|\big), \tag{2}$$

with $M_{t,0}$ the pooled error of the empty prefix. The $\min$ is elementwise over the 49 windows and removes non-monotonic decoder noise along a prefix. After this correction, a high value of $M_{t,k,j}$ means window j is still poorly explained by the observations so far, not that VideoMAE failed to decode an already pasted region.

Teacher and loss. At each prefix the teacher locates $j^\star = \arg\max_j M_{t,k,j}$, scores unselected patches that overlap $j^\star$, and appends the patch that most reduces that worst-window error, until $\max_j M_{t,k,j} \le \tau$. Every prefix yields a next-patch label and a 49-cell target map. Training is

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda_{\text{map}}\mathcal{L}_{\text{MAP}}, \qquad \lambda_{\text{map}} = 1, \tag{3}$$

where $\mathcal{L}_{\text{NTP}}$ is next-patch cross entropy and

$$\mathcal{L}_{\text{MAP}} = \sum_{j=1}^{49} w_{t,k,j}\left(\widehat{M}_{t,k,j} - M_{t,k,j}\right)^2. \tag{4}$$

We ignore the low loss regions using $\tau_h = 0.10$. The weight is

$$w_{t,k,j} = \left[M_{t,k,j} - \widehat{M}_{t,k,j}\right]_+ \mathbb{1}[M_{t,k,j} > \tau_h]. \tag{5}$$

High-loss cells are upweighted; underestimating them is upweighted further, so a degraded window is less likely to be predicted below the stopping threshold. The map is therefore used three times: by the teacher, by the ranking/stopping head, and as a regression target. This correction makes VideoMAE a more consistent *surrogate* for progressive observation.

3.3 MOREGAZE

MaxGaze can retain fine patches that help reconstruction (texture, background motion) yet receive little attention from the frozen vision encoder. **MoreGaze** prunes after encoding, using only that encoder, with no question token and no extra trained parameters.

Let G_t^l be the patches selected at scale l in frame t, and let $A_{t,h}^{l^- \to l}$ be last-layer attention from the adjacent coarser scale l^- to l for head h. The encoder-salience map is

$$q_t^l = \mathcal{N}_t \left[\left(\frac{1}{H N_{l^-}} \sum_{h=1}^{H} \sum_{i=1}^{N_{l^-}} A_{t,h,i,:}^{l^- \to l} - b^l \right)_+ \right], \tag{6}$$

where b^l is the same attention computed on a uniform gray image, following the observation that uninformative visual tokens can attract systematic attention (Kang et al., 2025), and $\mathcal{N}_t$ is per-frame min–max normalization. Gray calibration is an implementation choice for positional bias; the score is encoder salience, not a query-conditioned evidence map. We filter only the two finest scales $\mathcal{F}$:

$$\widetilde{G}_t^l = \begin{cases} \{p \in G_t^l \mid q_{t,p}^l \geq \gamma\}, & l \in \mathcal{F}, \\ G_t^l, & l \notin \mathcal{F}. \end{cases} \tag{7}$$

The default $\gamma = 0.10$ is selected on HLVid (Table 4) and then held fixed on the other benchmarks.

4 Model

Previous method (Shi et al., 2026) selects visual patch sequences under a fixed token budget and supplements them with thumbnail tokens for global context. However, its flattened visual sequence does not explicitly encode patch geometry across tiles or the real temporal intervals between video frames, and it provides no explicit delimiters between frame-level patch groups. With the calibrated maximum-loss stopping criterion, MaxGaze removes the need for a manually specified gazing ratio and continues selecting until local reconstruction errors are sufficiently low, allowing more complete visual details to be retained. We therefore reorganize the selected patches, insert an explicit timestamp before each frame, and propagate their geometry and time through the visual pipeline.

4.1 Gaze-Guided Vision Encoding

As shown in Fig. 2(c), video frames are grouped into 16-frame clips with stride 15. The shared boundary frame is encoded in both neighboring clips to preserve temporal context, but appears only once in the final language-model input. Each video frame is divided into several tiles. MaxGaze selects multi-scale patch IDs from each tile.

During intra-frame encoding, each complete tile is first encoded densely, allowing every selected patch to acquire full-tile context. Then we extract the features selected by MaxGaze and further apply MoreGaze to remove tokens with weak semantic relevance. The selected features are then gathered and processed by the cross-frame encoder. We regroup patches from all tiles by frame, apply pixel shuffle (Liu et al., 2024), and project the resulting tokens into the language model.

4.2 Input Construction for the LLM

For each frame, we concatenate the patch features from all its tiles, arranging them in a left-to-right, top-to-bottom order and separating adjacent tiles with newline characters (`\n`). Frame-level representations are then serialized temporally, with pairs of newline characters (`\n\n`) denoting the boundary between consecutive frames. To align the visual stream with the textual timeline, we insert a textual timestamp (e.g., `<3.5>`) before each frame at fixed 0.5-second intervals.

5 Experiments

5.1 Implementation details

For MaxGaze, we initialize it from AutoGaze and retain its four-scale patch vocabulary. We optimize it on our 27k corrected sequences using cosine learning-rate decay from 10^{-3} to 10^{-5}. We set $\lambda_{\text{map}} = 1$, identify high-loss windows using a corrected reconstruction loss of 0.10, and use $\tau_{max} = 0.15$ as the default maximum-loss stopping threshold.

We finetune NVILA-8B-VideoLiu et al. (2024) using LLaVA-Video-178K (Zhang et al., 2024) to maintain consistency with AutoGaze. Each frame is adaptively tiled into 392×392 regions

Table 1: Controlled accuracy–token comparison under two visual-input protocols. Tok. is the mean number of visual tokens entering the LLM. The last row is versus thumbnail-free AutoGaze. *64 thumbnails:* public NVILA-HD with its original thumbnail packing and no fine-tuning to our format; AutoGaze and MaxMoreGaze differ only in the gaze sequence. *No thumbnails:* MaxMoreGaze is evaluated using an MLLM instruction-tuned to our packed visual format, while AutoGaze is evaluated using its origin NVILA-HD.

Method	Thumbnail	VideoMME (w/o sub)		VideoMME (w/ sub)		MLVU		HLVid		MVBench		NExT-QA		LongVideoBench	
		Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓
AutoGaze	64	65.30	12.40K	70.33	12.40K	**67.20**	12.82K	46.64	26.36K	**68.70**	23.53K	**81.46**	12.06K	59.98	12.62K
MaxMoreGaze	64	**65.93**	11.82K	**70.75**	11.82K	67.00	11.53K	**48.30**	23.36K	**68.70**	20.53K	81.44	10.56K	**60.65**	11.48K
AutoGaze	-	60.59	5.27K	62.07	5.27K	61.04	3.54K	50.70	42.03K	59.65	2.86K	76.33	4.85K	55.35	7.42K
MaxMoreGaze	-	63.48	**4.71K**	65.33	**4.71K**	67.02	**2.42K**	**54.48**	**22.55K**	63.05	**2.69K**	**81.47**	**3.03K**	59.54	**5.51K**
Ours vs. AutoGaze (0 thumbnail)	–	**+2.89**	−10.6%	**+3.26**	−10.6%	**+5.98**	−31.6%	**+3.78**	−46.3%	**+3.40**	−5.9%	**+5.14**	−37.5%	**+4.19**	−25.7%

Table 2: Performance on long and high-resolution video benchmarks. Public token budgets are reported maximum or nominal values, while ours are measured final visual-token counts.

Method	Input setting	Visual tokens	Accuracy
HLVid			
LLaVA-OV-8B An et al. (2025)	32F, 384px	6.3K max	41.1
LongVILA-7B Chen et al. (2024a)	2048F, 384px	∼401K nominal	41.4
VideoLLaMA3-7B Zhang et al. (2025)	180F, 384px	16K max	38.8
VideoChat-Flash Li et al. (2025)	10K F, 448px	160K nominal	46.6
Qwen2.5-VL-7B Wang et al. (2024a)	48F, 896px	24.6K max	48.1
NVILA-8B-Video Liu et al. (2024)	256F, 448px	8.2K max	42.5
LiteFrame Kim et al. (2026)	48F, 2688 px	N/R	54.1
AutoGaze(baseline) Shi et al. (2026)	256F, 3584px	72.0K max	52.6
MaxMoreGaze(Ours)	1 fps, 3584 px	26.17K mean	**66.79**
HRVideoBench			
VISTA-Mantis	24F, up to 980px	1.54K max	51.0
AutoGaze(baseline)	64F, dynamic HD	30.0K max	52.3
MaxMoreGaze(Ours))	64F, dynamic HD	13.49K mean / 30K max	**60.5**

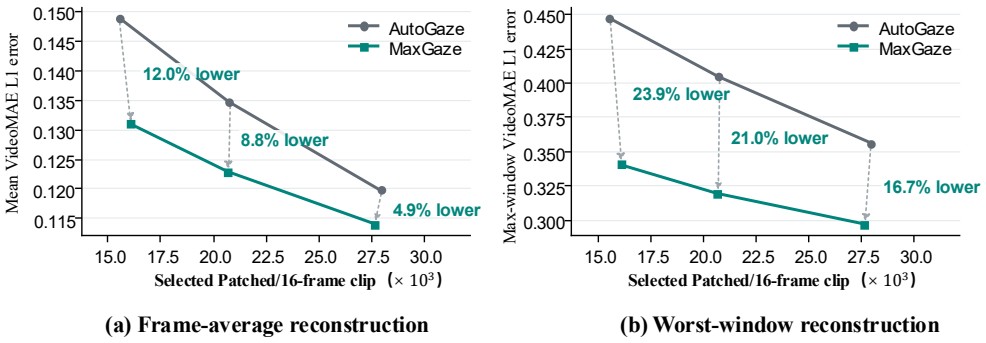
without an additional thumbnail. The dense intra-frame encoder is the first 22 layers of SigLIP2-SO400M (Tschannen et al., 2025). We keep it frozen, while its last five layers process the selected sparse cross-frame tokens. We use a TokenShuffle factor of 9 and train with a learning rate of 5×10^{-6} and a batch size of 512.

MoreGaze filters only the two finest patch scales using the attention maps of the last intra-frame encoder layer. We use gray-image calibration and set $\gamma = 0.10$ for both scales by default. Unless stated otherwise, video-question answering uses greedy decoding with temperature zero and the lmms-eval prompts and scorers. Visual-token counts are measured after TokenShuffle, immediately before the language model.

5.2 Main Results

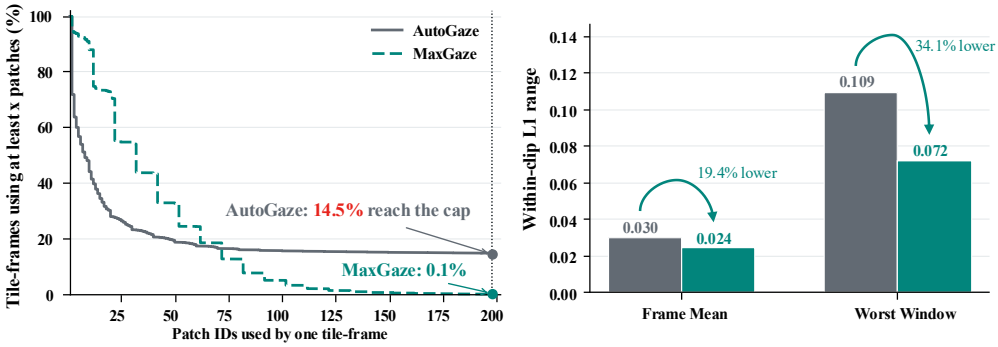
Controlled comparison with AutoGaze. Table 1 compares AutoGaze and MaxMoreGaze under two matched protocols: with 64 thumbnails and with none. With 64 thumbnails, MaxMoreGaze uses fewer visual tokens on all seven settings (-4.7% to -12.8%). Accuracy is higher on VideoMME, HLVid, and LongVideoBench, tied on MVBench-20, and within 0.20 points on MLVU and NExT-QA. Without thumbnails, MaxMoreGaze is more accurate on all seven settings ($+2.89$ to $+5.14$ points; mean $+4.09$) and uses fewer visual tokens on all seven (-5.9% to -46.3%). The accuracy gain is therefore not obtained by increasing the visual context.

Long and high-resolution video understanding. Table 2 reports a second protocol that does not match the frame/tile budgets in Table 1. On HLVid, MaxMoreGaze reaches **66.79** under a 1 fps, 3584 px setting, compared with 54.48 in the controlled table; the two numbers are not interchangeable. This setting outperforms the strongest listed prior method by **12.69** points and the MLLM with AutoGaze by **14.19** points. On HRVideoBench, it reaches **60.5**, improving over AutoGaze by **8.2** points. Reported token counts for our method are the final visual tokens passed to the LLM; several public baselines report only nominal or maximum budgets.



(a) Frame-average reconstruction (b) Worst-window reconstruction

Figure 3: Full-clip reconstruction error at matched budgets. (a) Frame-average VideoMAE L1: MaxGaze is 4.9–12.0% lower than AutoGaze. (b) Worst-window VideoMAE L1: MaxGaze is 16.7–23.9% lower.



(a) Per-tile budget tail (b) Variation across 16 frames

Figure 4: Stopping reliability under matched average budgets. (a) Survival function of per-tile acquisition length. With the same mean budgets (39 patches/tile), 14.5% of AutoGaze tile-frames reach the cap, compared with only 0.1% for MaxGaze. (b) Within-video variation of the terminal reconstruction loss across 16 frames.

5.3 Ablation Study

Evaluation protocol. We evaluate the HLVid videos using a frozen VideoMAE. In the full-clip test, each method reconstructs 16 frames with its own causal history, and we pair operating points with similar budget. We report both frame-average L1 reconstruction loss and the maximum L1 reconstruction loss over a 7×7 loss map. We also evaluate the 2k shared-context frame trajectories and measure the patches and max-loss AUC required to reach $\tau_{\max} \in \{0.30, 0.25, 0.20, 0.15\}$.

Full-clip reconstruction at matched compute. Figure 3 shows that, at matched budgets, MaxGaze reduces mean reconstruction error by **4.9–12.0%** and worst-window error by **16.7–23.9%**. The worst-window improvement is significant at all three operating points.

Stopping reliability. Average patch consumption alone does not reveal whether a policy stops reliably. Figure 4(a) reports the survival function of acquisition length over 55k tile-frame rollouts. At patch count x, the vertical axis denotes the percentage of tile-frames that use at least x patches. Although AutoGaze and MaxGaze have nearly identical mean budgets of 38.9 and 38.5 patches per tile, their distributions differ substantially. AutoGaze has a median/95th-percentile acquisition length of $7/198$, and 14.5% of its rollouts reach the configured 198-ID cap. In contrast, MaxGaze

Table 3: MaxGaze efficiency at corrected VideoMAE map quality targets. A lower max-loss AUC means the model fixes the worst-reconstructed regions faster with fewer patches.

Target	Patches ↓		Patch Δ	Pre-target max-loss AUC ↓		AUC Δ
$\tau_{\max}$	AutoGaze	MaxGaze		AutoGaze	MaxGaze	
.30	106.48	**93.57**	−12.12%	.2158	**.2086**	−3.36%
.25	132.75	**110.88**	−16.48%	.2428	**.2263**	−6.79%
.20	162.34	**134.17**	−17.35%	.2680	**.2462**	−8.12%
.15	190.54	**164.66**	−13.58%	.2865	**.2661**	−7.14%

Table 4: MoreGaze ablation on the full HLVid benchmark. The same threshold γ is applied to the 224- and 112-scale branches. Visual tokens are counted after TokenShuffle.

Gray base	γ	Accuracy (%)	Mean tokens	Reduction (%)
–	–	50.7	53,108	–
✓	0.05	**51.5**	31,654	40.4
✓	0.10	**51.9**	**25,492**	**52.0**
✓	0.15	50.0	21,899	58.7
✓	0.20	48.50	19,685	62.9
×	0.15	44.8	18,057	66.0

has a median/95th percentile of $31/101$, with only 0.1% reaching the cap. AutoGaze therefore stops many simple tiles prematurely while concentrating computation in a saturated tail.

Temporal stability. We further measure stopping stability using the within-video range of the terminal reconstruction loss across 16 frames at the same budget. As shown in Figure 4(b), MaxGaze reduces the frame-mean loss range from 0.030 to 0.024, a **19.4%** reduction. The worst-window range decreases from 0.109 to 0.072, a **34.1%** reduction.

Efficiency at strict local-quality targets. Table 3 shows that MaxGaze reaches the four local-quality targets with **12.12–17.35%** fewer patches. At the strict $0.20/0.15$ targets, it also reduces pre-target max-loss AUC by **8.12%/7.14%**.

5.3.1 Effectiveness of MoreGaze

We ablate the semantic-filter threshold and gray-image baseline on HLVid using frozen NVILA-8B-HD-Video. As shown in Table 4, $\gamma = 0.10$ achieves the best trade-off: it improves accuracy from 50.7% to 51.9% while reducing visual tokens by 52.0%. Increasing the threshold further yields stronger compression but gradually reduces accuracy. At the same threshold of 0.15, removing the gray baseline decreases accuracy from 50.0% to 44.8%.

5.4 Qualitative results

MaxGaze As shown in Figure 5 (left), MaxGaze focuses more rapidly on detailed bridge structures under equal budgets. Figure 5 (right) shows higher priority on high-loss regions across budgets.

MoreGaze Figure 6 shows MoreGaze filtering. After gray-base correction, retained patches emphasize semantic elements such as road signs and building logos, while dropping low-information structure.

6 Limitations

Our experiments are subject to two practical constraints. First, due to limited training data and computational resources, the current training set does not provide substantial performance gains from large-scale pretraining. Consequently, our adaptation primarily focuses on aligning the model with

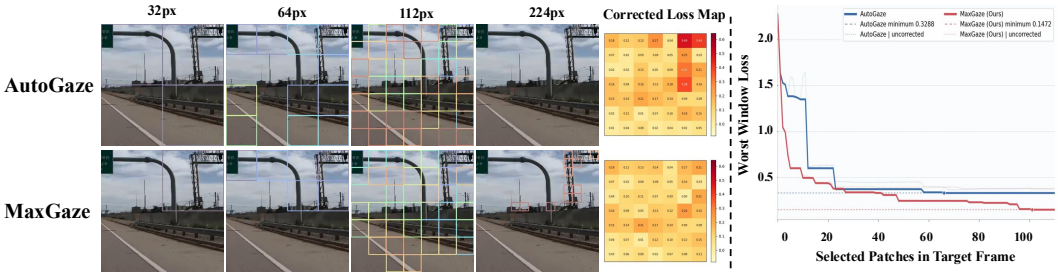


Figure 5: Qualitative results of MaxGaze. (Left) Under an identical context and fixed patch budget, our method captures finer details on the right-side bridge structure (latest selected patches in red boxes). (Right) Across patch budgets, the policy assigns higher selection priority to high-reconstruction-error regions (warmer colors).

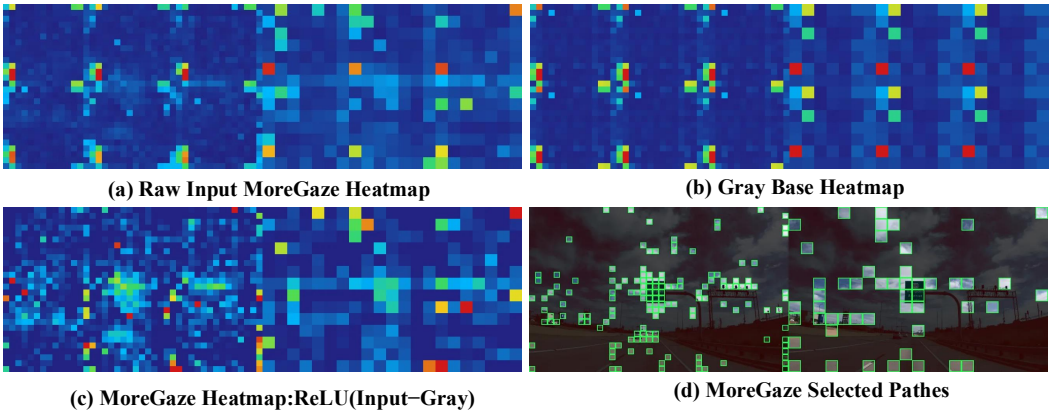


Figure 6: Qualitative results of MoreGaze. (a) Raw attention heatmap from the intra-frame encoder. (b) Gray-base heatmap from a uniform gray input. (c) Calibrated heatmap after correcting attention sinks. (d) Thresholded patches retained for downstream processing.

our proposed input construction pipeline rather than pursuing end-to-end performance improvements across all benchmarks. Second, existing evaluation protocols are mostly designed for low-resolution, short-clip scenarios. As a result, many baselines fail to benefit from longer temporal contexts or higher spatial resolutions. Moreover, there is a notable scarcity of established benchmarks for high-definition video understanding. Under these constraints, our method achieves significant accuracy improvements on two HD video understanding benchmarks, HLVid and HRVideoBench, while on other datasets, it primarily demonstrates its strength in reducing inference tokens and improving computational efficiency.

7 Conclusion

We presented **MaxMoreGaze**, a two-stage framework for more reliable and token-efficient adaptive video understanding. MaxGaze addresses the limitations of frame-average reconstruction supervision by predicting corrected spatial reconstruction-error maps, prioritizing the worst-reconstructed regions, and stopping according to the maximum remaining local error. This design improves both patch ordering and stopping reliability, reducing mean reconstruction error by 4.9–12.0% and worst-window error by 16.7–23.9% under matched patch budgets. MoreGaze complements this acquisition stage by reusing gray-calibrated attention signals from the frozen vision encoder to remove semantically weak fine-grained patches without introducing additional learned parameters. The retained tokens are further organized with explicit spatial and temporal structure for downstream reasoning. Because MaxGaze acquisition and MoreGaze semantic scoring are decoupled, their computation can be overlapped without extending the critical inference path.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing manuscript text. We have not used generative AI tools for synthetic data, theoretical models, mathematical claims and proofs, hypotheses, methodology or experiment design, implementing methods, translation, dataset cleaning, qualitative analysis, or interpreting results; these tasks were performed by the authors. We reviewed all AI-assisted wording and take responsibility for the final content of this work, including text, claims, and reported results.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards Next-Generation Perceptual Intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Jingfeng Chen, Jiawen Qian, Wendi Deng, Yinuo Guo, Jiaqi Yu, Sicong Leng, Raghuveer Thirukovalluru, and Bhuwan Dhingra. LDDR: Linear-DPP-Based Dynamic-Resolution Frame Sampling for Video MLLMs, 2026. URL https://arxiv.org/abs/2605.11477.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos, 2024a. URL https://arxiv.org/abs/2408.10188.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites, 2024b. URL https://arxiv.org/abs/2404.16821.

Haowen Hou, Zhen Huang, Zheming Liang, Qingyi Si, Chenglin Li, Shuai Dong, Kele Shao, Ruilin Li, Dianyi Wang, Nan Duan, and Jiaqi Wang. AdaCodec: A Predictive Visual Code for Video MLLMs, 2026. URL https://arxiv.org/abs/2606.02569.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models, 2025. URL https://arxiv.org/abs/2503.03321.

Jihwan Kim, Nikhil Parthasarathy, Danfeng Qin, Junhwa Hur, Deqing Sun, Bohyung Han, Ming-Hsuan Yang, and Boqing Gong. Liteframe: Efficient vision encoders unlock frame scaling in video llms. *arXiv preprint arXiv:2605.17260*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling, 2025. URL https://arxiv.org/abs/2501.00574.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, Xinyu Chen, Haoran Chen, Zizheng Huang, Jun Zhang, Kun Ouyang, Lin Sui, Ziang Yan, Yicheng Xu, Chenting Wang, Yinan He, Hongjie Zhang, Yi Wang, Yu Qiao, Yali Wang, Ziwei Liu, Kai Chen, and Limin Wang. VideoChat3: Fully Open Video MLLM for Efficient and Generalist Video Understanding, 2026. URL https://arxiv.org/abs/2607.14935.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024. URL https://arxiv.org/abs/2412.04468.

Minyoung Park, Taehun Kong, and Sangjun Ahn. DynaTok: Temporally Adaptive and Positional Bias-Aware Token Compression for Video-LLMs, 2026. URL https://arxiv.org/abs/2605.19322.

Haozhe Qi, Kevin Qu, Mahdi Rad, Rui Wang, Alexander Mathis, and Marc Pollefeys. AdaptToken: Entropy-based Adaptive Token Selection for MLLM Long Video Understanding, 2026. URL https://arxiv.org/abs/2603.28696.

Sayan Deb Sarkar, Rémi Pautrat, Ondrej Miksik, Marc Pollefeys, Iro Armeni, Mahdi Rad, and Mihai Dusmanu. CoPE-VideoLM: Leveraging Codec Primitives For Efficient Video Language Modeling, 2026. URL https://arxiv.org/abs/2602.13191.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding, 2024. URL https://arxiv.org/abs/2410.17434.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When Do We Not Need Larger Vision Models?, 2024. URL https://arxiv.org/abs/2403.13043.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and Hongxu Yin. Scaling Vision Pre-Training to 4K Resolution, 2025. URL https://arxiv.org/abs/2503.19903.

Baifeng Shi, Stephanie Fu, Long Lian, Hanrong Ye, David Eigen, Aaron Reite, Boyi Li, Jan Kautz, Song Han, David M. Chan, Pavlo Molchanov, Trevor Darrell, and Hongxu Yin. Attend before attention: Efficient and scalable video understanding via autoregressive gazing, 2026. URL https://arxiv.org/abs/2603.12254.

Boyuan Sun, Jiaxing Zhao, Xihan Wei, and Qibin Hou. LLaVA-Scissor: Token Compression with Semantic Connected Components for Video LLMs, 2025a. URL https://arxiv.org/abs/2506.21862.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A Training-free Approach for List-wise Frame Selection in Video-LLMs, 2025b. URL https://arxiv.org/abs/2501.02885.

Feilong Tang, Xiang An, Yunyao Yan, Yin Xie, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Chunyuan Li, Shikun Feng, Changrui Chen, Huajie Tan, Ming Hu, Manyuan Zhang, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. OneVision-Encoder: Codec-Aligned Sparsity as a Foundational Principle for Multimodal Intelligence, 2026. URL https://arxiv.org/abs/2602.08683.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. URL https://arxiv.org/abs/2203.12602.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos, 2024b. URL https://arxiv.org/abs/2405.19209.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models, 2024a. URL https://arxiv.org/abs/2407.15841.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images, 2024b. URL https://arxiv.org/abs/2403.11703.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-Time Vision-Language Interaction Intelligence, 2026. URL https://arxiv.org/abs/2606.14777.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-Voyager: Learning to Query Frames for Video Large Language Models, 2024. URL https://arxiv.org/abs/2410.03226.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. 2025. URL https://arxiv.org/abs/2501.13106.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. URL https://arxiv.org/abs/2410.02713.

Chenhao Zheng, Jieyu Zhang, Jianing Zhang, Weikai Huang, Ashutosh Kumar, Quan Kong, Oncel Tuzel, Chun-Liang Li, and Ranjay Krishna. TrajTok: Learning Trajectory Tokens enables better Video Understanding, 2026. URL https://arxiv.org/abs/2602.22779.