STAC: Selective Spatiotemporal Aggregation and Compression for Video Reasoning Segmentation

Syed Ariff Syed Hesham[1,2], Yun Liu[3,4,5⋆], Guolei Sun[3,4,5], Jing Yang[6], Henghui Ding[7], Xue Geng[2], Xudong Jiang[1]

[1] School of EEE, Nanyang Technological University, Singapore
[2] Institute for Infocomm Research, A*STAR, Singapore
[3] VCIP, CS, Nankai University, Tianjin, China
[4] AAIS, Nankai University, Tianjin, China
[5] NKIARI, Shenzhen Futian, China
[6] Guizhou University, Guizhou, China
[7] Fudan University, Shanghai, China

Abstract. Video reasoning segmentation demands pixel-accurate object tracking across hundreds of frames under complex natural language queries, producing dense spatiotemporal tokens whose quadratic self-attention cost makes long-video processing prohibitive. Existing methods address this through token compression, yet typically operate on encoder features lacking temporal context, constraining selection before content redundancy can be reliably assessed. Informed compression requires contextual awareness, but acquiring that awareness at full resolution incurs the same quadratic cost compression aims to reduce. State-space models resolve this constraint, as their linear recurrence selectively conditions each token on temporal context at $\mathcal{O}(T)$ cost, producing representations where content redundancy becomes assessable. Building on this, **S**elective Spatio**T**emporal **A**ggregation and **C**ompression (**STAC**) enriches features via decoupled bidirectional spatial and causal temporal scanning, leveraging recurrence-derived redundancy for hierarchical compression with adaptive thresholds optimised with segmentation objective. STAC achieves 85% token reduction and 1.8× speedup while surpassing compression-free baselines on reasoning segmentation benchmarks in a zero-shot streaming-compatible setting. Code is available here.

Keywords: Video Reasoning Segmentation · Video Token Compression · Spatiotemporal Modeling · State-Space Models · Multimodal LLMs

1 Introduction

Large Language Models (LLMs) [1, 18, 34, 72] have transformed the field of artificial intelligence, demonstrating remarkable capabilities in reasoning and task completion. Building upon this foundation, Multimodal Large Language Models

⋆ Corresponding author: Yun Liu (liuyun@nankai.edu.cn)

Table 1: Comparison of architectural characteristics across video compression approaches for reasoning segmentation. Brackets (✓) indicate partial support.

Method	Temporal Integration	Task-Aware Compression	Hierarchical Understanding	Online Processing
Local Compression Methods				
FastV [11]	✗	✗	✗	✗
ATP-LLaVA [80]	✗	✓	✗	✗
TempMe [65]	✓	(✓)	✓	✗
Adaptive Compression Methods				
LongVU [66]	✓	✓	✗	✗
M-LLM [31]	✓	✓	✗	✗
TSPO [68]	✓	✓	✗	✗
Bidirectional SSM Methods				
VideoMamba [43]	✓	✗	✗	✗
BIMBA [32]	✓	(✓)	✗	✗
STORM [35]	✓	✗	✗	✗
STAC (Ours)	✓	✓	✓	✓

(MLLMs) have extended these capabilities to visual understanding, achieving impressive results on image comprehension [5, 13, 42, 47, 48] and video analysis [14, 45, 55, 70, 73, 83]. Among these capabilities, video reasoning segmentation [6, 41, 78] represents a particularly demanding challenge that requires models to interpret complex natural language queries, reason about spatiotemporal relationships, and produce pixel-accurate masks across extended sequences.

Most video MLLMs process videos by encoding each frame independently with pretrained image encoders [60, 82], concatenating the resulting tokens, and passing them to a language model [5, 47, 48]. This strategy proves effective for short clips spanning several seconds where token streams remain manageable and self-attention mechanisms can discover temporal relationships. However, this breaks down for longer videos due to quadratic attention complexity [50] and token explosion. For example, a 2-minute video at 2 FPS with a vision encoder [82] producing 729 tokens per frame generates 61K tokens, nearly twice the typical 32K context window [18, 34], making long-video processing impractical [12].

To address these challenges, most long-video methods adopt compression strategies that use different approaches to reduce the overall token sequence. Frame selection approaches [12, 40, 46, 66] target temporal redundancy but risk eliminating critical motion cues for tracking. An alternative strategy employs spatial compression [11, 67, 68, 80], applying pooling or attention-based aggregation to reduce the number of per-frame tokens. However, operating within isolated frames limits access to cross-frame dependencies. Recent work [31, 45, 66, 83] combines both strategies, though compression decisions typically rely on predetermined heuristics or frame-level features rather than learned selection informed by global spatiotemporal context.

These compression strategies, while effective, share a fundamental limitation. By operating on raw encoder features, they commit to token reduction before the model has understood the video's spatiotemporal structure. This creates a tension between compression and understanding, since acquiring contextual

awareness at full resolution is precisely the cost compression aims to avoid. We observe that state-space models (SSMs) [23,25–27] resolve this tension. Their linear recurrence selectively conditions each hidden state on current and past content at $\mathcal{O}(T)$ cost, such that already-absorbed tokens produce near-identical enriched representations. This provides a redundancy signal intrinsic to the model's dynamics, a capability unique to persistent state accumulation. Among SSMs, Mamba [22] introduces selective state-propagation mechanisms [24] that dynamically retain relevant information while filtering redundancy. Extensions to the video domain [10,43,51,57] confirm these properties for temporal modelling, but most rely on unified bidirectional scanning [32,35] that treats the spatiotemporal volume as a single block, hindering streaming deployment. A fundamental asymmetry underlies this limitation, as spatial relationships within a frame are non-causal with objects interacting with all neighbours simultaneously, whereas temporal evolution follows strict causal progression where future frames cannot influence past observations. This motivates decoupling spatial and temporal scanning to maintain both efficiency and streaming compatibility.

Building on this observation, **S**elective Spatio**T**emporal **A**ggregation and **C**ompression (**STAC**) exploits this SSM-derived redundancy signal for principled video token compression. Unlike existing SSM-based video methods that employ state-space model purely as a processing backbone, STAC enriches features through SSM before compressing them, ensuring redundancy decisions reflect tokens the recurrence has already absorbed rather than surface-level similarity. Specifically, our State-informed Spatiotemporal Aggregator (SSA) first enriches encoder features through bidirectional spatial and causal temporal scanning that respects each dimension's distinct structure, with the causal design independent of future observations to enable streaming-compatible deployment. Our Hierarchical State-adaptive Compression (HSC) module then leverages this enriched feature space for hierarchical temporal-then-spatial reduction, where adaptive thresholds respond to local content dynamics, compressing aggressively during static segments while preserving tokens at motion boundaries. Finally, task-grounded optimisation propagates segmentation gradients back into the compression policy, closing the loop so the downstream task directly determines which tokens are retained. This integrated design achieves $\sim$85% token reduction with $\sim$1.8$\times$ speedup while surpassing compression-free baselines on reasoning segmentation benchmarks in a zero-shot, streaming-compatible setting.

Our primary contributions include:

- **State-informed Spatiotemporal Aggregator (SSA)** that enriches encoder features through decoupled bidirectional spatial and causal temporal scanning, producing a semantically grounded feature space where content redundancy becomes discernible before compression, while supporting streaming deployment through its causal temporal design.
- **Hierarchical State-adaptive Compression (HSC)** that performs temporal -then-spatial token reduction in the recurrence-enriched feature space with adaptive thresholds responding to local content dynamics, compressing static regions aggressively while preserving motion boundaries.

– **Task-Grounded Differentiable Compression** that propagates segmentation gradients directly through discrete compression decisions via straight-through estimation, aligning token retention with downstream mask accuracy and yielding content-adaptive policies with zero-shot transferability to unseen reasoning benchmarks.

2 Related Work

2.1 Language-Guided Video Segmentation

Video reasoning segmentation requires interpreting complex natural-language queries to produce pixel-accurate masks while maintaining temporal consistency across extended sequences [16, 41, 77]. Traditional referring video object segmentation employs specialized transformers with explicit cross-modal fusion [9, 20, 38], with methods like ReferFormer [76] using deformable attention for multi-frame processing and MTTR [9] introducing end-to-end temporal architectures. However, these lack flexible reasoning for complex queries requiring world knowledge or multi-hop inference [46]. MLLM-based approaches integrate vision foundation models with large language models [39, 41, 48], with LISA [41] pioneering embedding-as-mask paradigm for reasoning-based segmentation and video extensions introducing temporal modeling [6, 62, 78]. VISA [78] employs hierarchical encoding with temporal propagation, VideoLISA [6] proposes sparse-dense sampling with one-token-seg-all, and GLUS [46] applies comprehensive global-local reasoning. Concurrently, VRS-HQ [21] introduces hierarchical frame-level and temporal tokens with dynamic aggregation and occlusion-aware keyframe selection via SAM2, while Sa2VA [81] unifies SAM2 with LLaVA into a shared token space where LLM-generated instruction tokens guide mask prediction across images and videos. While these approaches are effective offline, they process video tokens uniformly without explicit compression [6, 46, 78], treating all frames with equal computational weight and requiring access to complete sequences. This uniform treatment inherently limits scalability, prevents online deployment, and ignores the temporal redundancy that could potentially be exploited to enhance efficiency without sacrificing the fine-grained motion cues essential for reasoning.

2.2 Efficient Video Understanding

Conventional video understanding methods have extensively explored efficient spatiotemporal modeling to reduce redundant computation while preserving motion-sensitive representations [3, 4, 19, 28, 37, 74]. This concern becomes more pronounced in modern token-based video reasoning systems: video MLLMs generate hundreds of tokens per frame with the use of pretrained encoders [48, 60, 82], rapidly exceeding context limits [18, 34]. *Token compression* mitigates spatial redundancy via learned selection [56, 61] or attention-based pruning [8, 11, 80]. DynamicViT [61] hierarchically prunes tokens, ToMe [8] merges via bipartite matching, FastV [11] performs early pruning, and ATP-LLaVA [80] introduces adaptive layer-wise thresholds. Video extensions exploit temporal redundancy [35, 65,

67], with TempMe [65] merging cross-frame tokens and STORM [35] applying Mamba layers. *Frame selection* reduces temporal dimensionality through uniform subsampling [12, 46] or adaptive scoring [31, 40, 66, 68]. LongVU [66] leverages similarity-based filtering, while M-LLM [31] and TSPO [68] employ spatial pooling or reinforcement learning. However, these methods often rely on static ratios which are unsuited for varying complexity or heuristics lacking supervision, and universally require offline access to complete video sequences.

2.3 State-Space Models for Visual Understanding

State-space models building on classical theory [36] provide efficient sequence processing through structured representations [23,25–27]. Mamba [22] introduces selective scan mechanisms that achieve linear complexity via input-dependent state propagation, retaining salient features while filtering redundancy [24]. Vision applications adapt these principles through directional scanning [49, 58, 86], with Vision mamba [86] employing bidirectional scanning and VMamba [49] introducing cross-scan mechanisms. Video extensions apply SSM across spatiotemporal volumes [10, 43, 51, 57, 79], with VideoMamba [43] adopting bidirectional scanning and VideoMambaPro [51] addressing historical state decay through masked backward computation. Recent work integrates mamba into video MLLMs [32,35,84], with BIMBA [32] introducing bi-directional spatiotemporal token selection with interleaved visual queries. However, existing architectures apply a unified bidirectional scan across complete sequences [32, 35, 43, 57], treating spatial and temporal dimensions uniformly despite temporal evolution being strictly causal unlike symmetric spatial relationships. This requires full video access, preventing incremental frame-by-frame computation essential for online processing [32, 65, 75]. Furthermore, prior work focuses on end-to-end processing [32, 35] rather than leveraging selective conditioning to enrich intermediate representations before task-specific selection in dense prediction.

Our work addresses these gaps by compressing visual tokens *upstream* through selective state conditioning, architecturally decoupling bidirectional spatial scanning from causal temporal scanning, and propagating segmentation gradients through compression decisions via task-grounded optimisation to enable efficient, streaming-compatible reasoning segmentation.

3 Methodology

3.1 Architecture Overview

Video reasoning segmentation requires discriminating temporally redundant content from semantically critical motion cues across long sequences. Since compression applied to raw encoder features lacks this temporal awareness, STAC places selective state-space conditioning before compression, ensuring token reduction operates on features where redundancy reflects integrated temporal context rather than raw appearance alone (Fig. 1). Given video $\mathbf{V} \in \mathbb{R}^{T\times 3\times H\times W}$

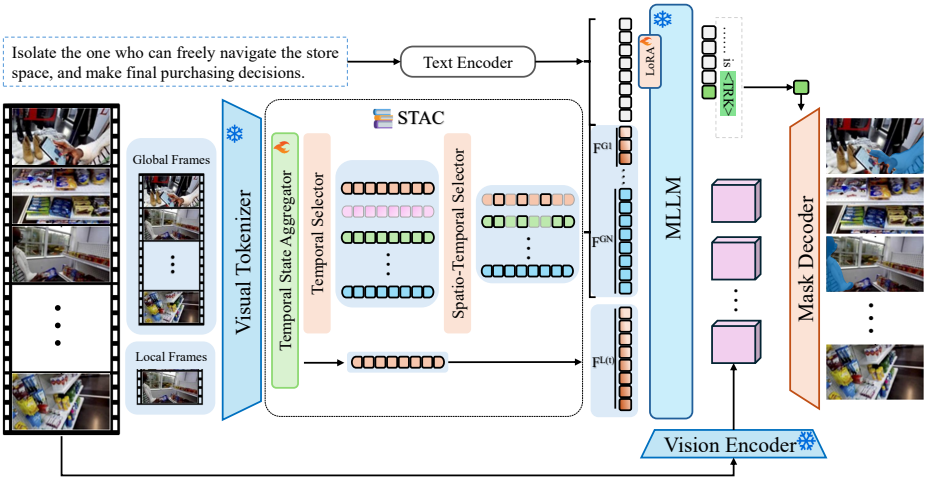

Fig. 1: Our architecture introduces a generic spatiotemporal summarizer, which allows for temporal selection and spatiotemporal compression to computation while improving overall performance of the model.

with frames $\{\mathbf{I}_1, \ldots, \mathbf{I}_T\}$ and text query q, we generate segmentation masks $\mathbb{M} = \{m_1, \ldots, m_T\}$ where $m_t \in \{0,1\}^{H \times W}$. Following standard protocols [5,47], we encode frames through frozen CLIP ViT-L/14 [60] to produce $N = 256$ tokens per frame with features $\mathbf{F}_e \in \mathbb{R}^{T \times N \times d_e}$ where $d_e = 1024$. The framework realises this through three cascaded stages, each respecting the distinct causal structure of spatial and temporal dimensions:

1. **Stage 1: Decoupled Spatiotemporal Scanning.** We transform independent frame tokens into globally-aware representations $\mathbf{F}_{\text{ST}} \in \mathbb{R}^{T \times N \times d_e}$ via a decoupled scanning architecture built on cascaded Mamba modules [15, 22]. Diverging from architectures that treat spatiotemporal volumes as unified isotropic blocks, we systematically *decouple* dimensions via bidirectional spatial scanning for semantic completeness and strictly causal temporal scanning for motion evolution. This separation guarantees $\mathcal{O}(TNd_e)$ linear complexity and prevents future-frame leakage to enable incremental inference.
2. **Stage 2: Asymmetric Causal Compression.** To circumvent latency bottlenecks in offline clustering, we implement an Asymmetric "Predict-then-Compress" policy where the current frame $\mathbf{F}_{\text{ST}}[t]$ is processed at full resolution for the immediate timestep, while all preceding frames are *subsequently* compressed into a compact history $\mathbf{F}_c$ ($T' \ll T$) serving as temporal memory.
3. **Stage 3: Task-Objective Optimization.** We generate the ⟨TRK⟩ token for mask decoding [63] by projecting the compressed features $\mathbf{F}_c$ to dimension $D = 4096$ while treating compression as a differentiable component. Instead of relying on generic heuristic reconstruction metrics, we train end-to-end to propagate segmentation gradients directly through discrete compression decisions via straight-through estimation [7], aligning the compression policy specifically to the segmentation objective.

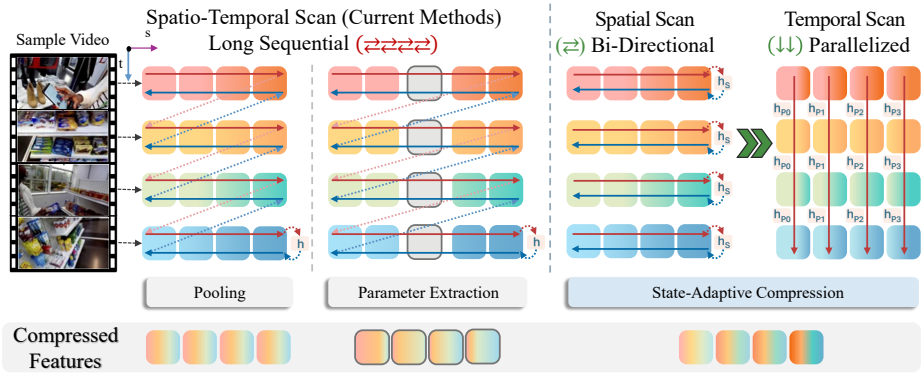

Fig. 2: Comparison of scanning strategies. *Current methods (left)* flatten videos into long T×H×W sequences requiring bidirectional processing over the entire video, and use fixed pooling [35, 43] or learnable parameter extraction [32] for compression. *Our method (right)* decouples bidirectional spatial scanning from causal temporal scanning (spatially parallelized), enabling streaming-compatible content-adaptive compression via our HSC module while achieving superior token reduction.

3.2 State-informed Spatiotemporal Aggregator

Standard visual encoders produce features $\mathbf{F}_e$ as isolated spatial patches devoid of global or temporal context. Inter-frame similarity on these representations captures surface-level visual resemblance rather than semantic redundancy, fundamentally limiting the quality of any compression decision derived from them. The proposed State-informed Spatiotemporal Aggregator (SSA) resolves this deficiency through a dual-stage architecture that enriches each token with spatiotemporal context *before* compression. SSA realises this in two cascaded stages, integrating spatial context within each frame using bidirectional scanning, then propagating temporal context causally across frames to produce a feature space in which content redundancy becomes apparent.

Spatial Aggregation. Within a single frame, spatial relationships are non-causal with pixels interacting with all neighbors simultaneously. We capture this global context without imposing artificial ordering by applying bidirectional selective state scanning [86] independently across all T frames. By averaging forward ($\mathcal{M}^b_{\rightarrow}$) and backward ($\mathcal{M}^b_{\leftarrow}$) state scans via

$$\mathbf{F}_{\text{spatial}}[t] = \frac{1}{2}\left(\mathcal{M}^b_{\rightarrow}(\mathbf{F}_e[t]) + \mathcal{M}^b_{\leftarrow}(\mathbf{F}_e[t])\right), \tag{1}$$

we effectively eliminate directional bias. This operation ensures the resulting features encode precise object boundaries and rich within-frame semantics which are essential prerequisites for pixel-level segmentation.

Temporal Aggregation. Conversely, temporal evolution adheres to strict physical causality. We respect this constraint to enable online streaming by applying causal selective state scanning $\mathcal{M}_{\text{causal}}$ [15] efficiently parallelized across

N spatially independent locations. The mechanism accumulates temporal context through a recurrent update

$$\mathbf{F}_{\mathrm{ST}}[n] = \mathcal{M}_{\mathrm{causal}}(\mathbf{F}_{\mathrm{spatial}}[:, n]), \tag{2}$$

where the causal selective state-space mechanism for location n accumulates temporal context through

$$\mathbf{h}_{n,t} = \bar{\mathbf{A}}_t \mathbf{h}_{n,t-1} + \bar{\mathbf{B}}_t \mathbf{F}_{\mathrm{spatial}}[t, n]. \tag{3}$$

The input-dependent matrices $\bar{\mathbf{A}}_t, \bar{\mathbf{B}}_t \in \mathbb{R}^{d_e \times d_e}$ enable selective propagation of salient motion patterns while filtering redundancy. This causal structure supports incremental inference, as each incoming frame updates hidden states $\mathbf{h}_{n,t}$ without recomputing past representations, diverging from unified bidirectional methods [32,43,57] that require full video access and preclude streaming deployment. The resulting enriched features $\mathbf{F}_{\mathrm{ST}}$ form the basis for the compression decisions that follow.

3.3 Hierarchical State-adaptive Compression

The proposed Hierarchical State-adaptive Compression (HSC) module performs hierarchical token reduction in the enriched feature space $\mathbf{F}_{\mathrm{ST}}$, where the selective state conditioning of SSA (Sec. 3.2) has already integrated temporal context into each token. Within this space, consecutive frames whose content the recurrence has represented produce near-identical outputs, revealing redundancy that raw encoder features without such temporal integration cannot expose. HSC exploits this redundancy through a temporal-then-spatial hierarchy (Tab. 3), first identifying and merging redundant frames before addressing within-frame spatial reduction, as premature spatial pooling would destroy the inter-frame differences this redundancy signal depends on.

Adaptive Temporal Compression. Adjacent frames in video sequences exhibit significant temporal redundancy [40], yet motion information resides precisely in the inter-frame differences that uniform downsampling discards. We identify redundancy via efficient $\mathcal{O}(T)$ sequential similarity. For each frame t, we compute frame-level representations via spatial averaging and measure similarity to the preceding frame:

$$\mathbf{g}_t = \frac{1}{N} \sum_{n=1}^{N} \mathbf{F}_{\mathrm{ST}}[t, n], \quad s_t = \frac{\langle \mathbf{g}_t, \mathbf{g}_{t-1} \rangle}{\|\mathbf{g}_t\|_2 \|\mathbf{g}_{t-1}\|_2}. \tag{4}$$

To circumvent the rigidity of fixed hyperparameters and ensure streaming compatibility, we avoid global statistics by estimating the distribution online via exponential moving averages (EMA). Let μ_t and σ_t^2 denote the running mean and variance at step t with momentum $\alpha = 0.1$:

$$\mu_t = \alpha s_t + (1 - \alpha)\mu_{t-1}, \quad \sigma_t^2 = \alpha(s_t - \mu_t)^2 + (1 - \alpha)\sigma_{t-1}^2. \tag{5}$$

The adaptive threshold is defined as $\tau_t = \mu_t + k\sigma_t$, where parameter k is learned via a lightweight MLP. Consecutive frames satisfying $s_t > \tau_t$ are merged via averaging, producing compressed history $\mathbf{F}_{\text{temp}}$ with typical reduction of 60%-70% driven purely taking interframe content dynamics into account.

Adaptive Spatial Compression. While temporal compression reduces frame count, spatial redundancy persists in background regions. We address this by measuring cross-frame coherence c_n across the temporally-compressed sequence for each spatial location n across the temporally-compressed sequence:

$$c_n = \frac{1}{T' - 1} \sum_{t=2}^{T'} \frac{\langle \mathbf{F}_{\text{temp}}[t, n], \mathbf{F}_{\text{temp}}[t-1, n] \rangle}{\|\mathbf{F}_{\text{temp}}[t, n]\|_2 \|\mathbf{F}_{\text{temp}}[t-1, n]\|_2}. \tag{6}$$

High coherence indicates temporally static patches whose representations across frames are suitable for merging, while low coherence highlights dynamic foreground regions requiring preservation. We compute adaptive thresholds using the statistic formulation in Eq. (5) applied per spatial location with local statistics $(\mu_c^{(n)}, \sigma_c^{(n)})$. Locations exceeding the threshold have their temporal tokens merged via averaging, while dynamic locations retain full temporal resolution. This process achieves 40%-50% further reduction concentrated on static regions, yielding compressed features $\mathbf{F}_c$ with $|\mathbf{F}_c| \ll T' \times N$ total tokens and a combined overall reduction of $\sim$85%.

3.4 Task-Grounded Differentiable Compression

Standard compression metrics optimised for reconstruction may not guarantee preservation of segmentation-critical boundaries. The proposed framework instead integrates compression as a differentiable component within the segmentation pipeline, so that task gradients directly shape token retention through a joint objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{comp}}, \tag{7}$$

where $\mathcal{L}_{\text{seg}}$ comprises standard cross-entropy and Dice loss [63] with $\lambda = 0.001$ balances the two objectives. The compression regularization term $\mathcal{L}_{\text{comp}}$ adapts to video complexity:

$$\mathcal{L}_{\text{comp}} = \frac{1}{B} \sum_{i=1}^{B} (1 - \hat{C}_i) \cdot \frac{c_i}{\sqrt{T_i}}. \tag{8}$$

Here B denotes batch size, c_i counts retained tokens, T_i denotes original token count, and $\hat{C}_i = \frac{1}{2}[(1 - \mu_i) + \sigma_i]$ estimates complexity based on feature variance to reduce penalties for dynamic videos, and the square-root normalization $1/\sqrt{T_i}$ ensures sublinear scaling for long sequences. Such scaling incurs lower per-token penalties, preventing over-aggressive compression that would eliminate critical motion information in extended sequences while still incentivizing efficiency.

Alignment between compression and segmentation emerges through the learned thresholds in Sec. 3.3, which segmentation gradients adjust via the straight-through estimator [7]. The threshold evolution (Eq. (5)) stabilises these updates

despite discrete selection, allowing training to converge on a content-adaptive policy that compresses aggressively during static segments while preserving tokens at motion boundaries. This content-adaptive policy completes the design loop in which decoupled scanning (Sec. 3.2) enriches features with spatiotemporal context, HSC translates the resulting redundancy into compression decisions, and the task objective calibrates those decisions to segmentation accuracy. As demonstrated in Fig. 4, static videos sustain compression exceeding 90% while action sequences retain 30%–50% of tokens for boundary precision.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets. We evaluate on five benchmarks spanning two categories distinguished by query complexity. *Referring benchmarks* require localizing objects through linguistic descriptions. Refer-YouTube-VOS [77] provides 3,978 videos with 15,458 expressions emphasizing appearance-based attributes and spatial relationships, while MeViS [16] comprises 2,006 videos with 28,570 expressions where targets must be disambiguated through motion patterns rather than static features. Ref-DAVIS17 [38] extends DAVIS-2017 [59] with 90 videos to evaluate performance under occlusion and appearance variation. *Reasoning benchmarks* require multi-step inference integrating world knowledge. ReVOS [78] provides 1,042 videos with 35,074 instruction-mask pairs demanding contextual inference beyond direct visual matching, while ReasonVOS [6] contains 91 videos with 458 samples specifically evaluating temporal understanding and causal reasoning across extended contexts. We report region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ on official validation splits.

Implementation. Our architecture employs pretrained frozen CLIP ViT-L/14 [60] at 224×224 resolution as the vision encoder, producing 256 tokens per frame with 1024-dimensional features. Features undergo spatiotemporal enrichment through our dual-stage SSA module maintaining 1024-dimensional hidden states, followed by HSC module for hierarchical compression via learned adaptive thresholds as described in Sec. 3. Compressed tokens are projected to 4096 dimensions and processed by LLaVA-7B [48, 71] with LoRA fine-tuning [30] applied exclusively to attention projection matrices at rank 8 and α=16, while base parameters remain frozen. SAM2 [63] then decodes the language-aligned embeddings into segmentation masks. We train exclusively on referring segmentation datasets— MeViS [16] and Refer-YouTube-VOS [77]—withholding all reasoning segmentation data to evaluate zero-shot transfer on Ref-DAVIS17 [38], ReVOS [78], and ReasonVOS [6]. Training uses AdamW [2] optimizer with β_1=0.9, β_2=0.999, weight decay 0.05, learning rate 3×10^{-4}, and 100-step linear warmup. Each iteration processes batch size 2 with context frames sampled dynamically between 2-32 frames per video. Training completes 3,000 iterations in approximately 30 hours on 2× NVIDIA A40 GPUs.

Baselines. We compare against two categories isolating compression strategies. *Compression-free baselines* include specialized referring architectures like

Table 2: Comparison with state-of-the-art methods on referring and reasoning video object segmentation benchmarks. We report $\mathcal{J}\&\mathcal{F}$, $\mathcal{J}$, and $\mathcal{F}$ scores. Notably, STAC achieves competitive or superior performance across all benchmarks while operating on approximately **∼15% visual tokens** through adaptive compression, compared to other methods that process complete token sequences.

Models	Ref-DAVIS17			MeViS			Ref-YouTube			ReasonVOS			ReVOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Referring Video Object Segmentation Methods															
URVOS [64]	51.6	47.3	56.0	27.8	25.7	29.9	47.1	45.3	49.2	-	-	-	-	-	-
RIS [17]	54.3	-	-	29.3	27.8	30.8	49.4	48.2	50.6	-	-	-	-	-	-
TempCD [69]	-	-	-	-	-	-	65.8	63.6	68.0	-	-	-	-	-	-
DsHmp [29]	64.9	61.7	68.1	46.4	43.0	49.8	67.1	65.0	69.1	-	-	-	-	-	-
OnlineRefer [75]	64.8	61.6	67.7	-	-	-	63.5	61.6	65.5	38.7	34.6	42.9	-	-	-
SOC [52]	-	-	-	-	-	-	67.3	65.3	69.3	35.9	33.3	38.5	-	-	-
SgMg [53]	63.3	60.6	66.0	-	-	-	65.7	63.9	67.4	36.2	33.7	38.7	-	-	-
MTTR [9]	-	-	-	30.0	28.8	31.2	55.3	54.0	56.6	31.1	29.1	33.1	25.5	25.1	25.9
ReferFormer [76]	61.1	58.1	64.1	31.0	29.8	32.2	62.9	61.3	64.6	32.9	30.2	35.6	28.1	26.2	29.9
LMPM [16]	61.1	58.1	64.1	31.0	29.8	32.2	62.9	61.3	64.6	32.9	30.2	35.6	28.1	26.2	29.9
Multimodal LLM-based Reasoning Methods															
VideoGLaMM [54]	69.5	65.6	73.3	45.2	42.1	48.2	-	-	-	-	-	-	-	-	-
LISA-7B [41]	64.8	62.2	67.3	37.2	35.1	39.4	53.9	53.4	54.3	31.1	29.1	33.1	-	-	-
LISA-13B [41]	66.0	63.2	68.8	37.9	35.8	40.0	54.4	54.0	54.8	-	-	-	-	-	-
LLaMA-VID [44]+LMPM	-	-	-	-	-	-	-	-	-	-	-	-	26.1	20.9	31.4
VideoLISA-3.8B (1-Tok) [6]	67.7	63.8	71.5	42.3	39.4	45.2	61.7	60.2	63.3	45.1	43.1	47.1	-	-	-
VideoLISA-3.8B (Post) [6]	68.8	64.9	72.7	44.4	41.3	47.6	63.7	61.7	65.7	47.5	45.1	49.9	-	-	-
TrackGPT-7B [85]	63.2	59.4	67.0	40.1	37.6	42.6	56.4	55.3	57.4	-	-	-	43.6	41.8	45.5
TrackGPT-13B [85]	66.5	62.7	70.4	41.2	39.2	43.1	59.5	58.1	60.8	-	-	-	45.0	43.2	46.8
VISA-Chat-7B [78]	69.4	66.3	72.5	43.5	40.7	46.3	61.5	59.8	63.2	-	-	-	46.1	43.9	48.2
VISA-Chat-13B [78]	70.4	67.0	73.8	44.5	41.8	47.1	63.0	61.4	64.7	-	-	-	47.5	45.3	49.7
GLUS^S [46]	71.9	68.3	75.4	**50.3**	**47.5**	**53.0**	64.8	63.3	66.4	49.8	47.2	52.5	**49.5**	**46.9**	**52.0**
STAC (Ours)	**73.6**	**70.1**	**77.2**	49.9	47.2	52.5	**65.2**	**63.5**	**66.9**	**52.3**	**49.5**	**55.0**	48.4	45.8	51.3

ReferFormer [76], OnlineRefer [75], and TempCD [69], alongside MLLM-based methods LISA [41], VISA [78], VideoLISA [6], and GLUS [46], which process full token sequences without explicit reduction. *Compression-based baselines* include 3D pooling with uniform spatiotemporal reduction, attention-based selection replacing Mamba modules with transformer layers and pruning via attention scores, Perceiver [33] using learned cross-attention queries for aggregation, and BIMBA [32] employing bidirectional state-space compression. All baselines use official implementations with hyperparameters aligned to our configuration and equivalent reduction ratios for fair comparison.

4.2 Main Results

Tab. 2 compares STAC against state-of-the-art methods on referring and reasoning video object segmentation. Despite reducing visual tokens by 85%, STAC achieves strong results on benchmarks emphasizing spatial localization and temporal reasoning. On Ref-DAVIS17 [38] and Ref-YouTube [77], our method surpasses all baseline compression-free approaches, demonstrating that context-aware selection enhances appearance-based referring segmentation even under

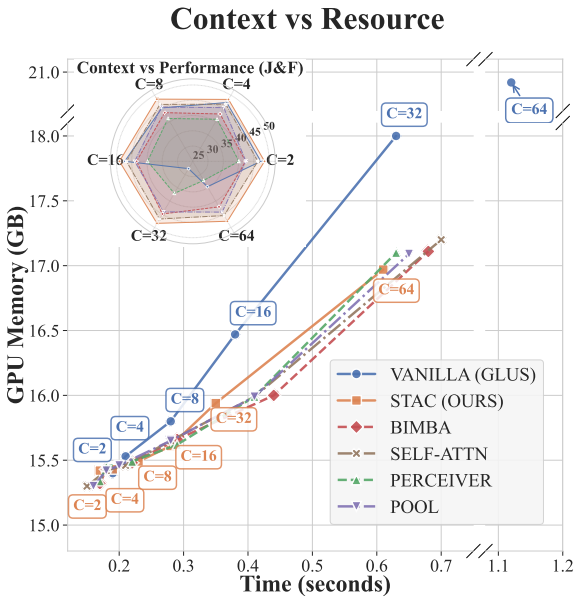

Fig. 3: Performance evaluation across compression methods. STAC maintains efficiency (main) and consistent performance across contexts (inset).

occlusion. Most notably, despite training exclusively on referring segmentation data, STAC transfers effectively to reasoning benchmarks, outperforming the strongest compression-free baseline on ReasonVOS [6] in a zero-shot setting. This suggests that informed compression does not merely preserve reasoning ability but may actively improve it by forcing the model to retain only motion-critical content, filtering static redundancy that uncompressed approaches process.

On motion-centric benchmarks such as MeViS [16] and ReVOS [78], methods with full bidirectional temporal access achieve marginally stronger results. Yet STAC approaches comparable performance while compressing 85% of visual tokens through selective state conditioning upstream before MLLM ingestion, with a causal temporal design that permits frame-by-frame online processing using only past observations. Furthermore, this upstream enrichment-before-compression strategy addresses the quadratic attention bottleneck before it reaches the MLLM, demonstrating that selective state conditioning captures sufficient temporal context from causal processing alone. This capability makes STAC distinctive among current approaches in combining competitive segmentation accuracy with substantial token compression and real-time deployability.

4.3 Ablation Studies

All ablations follow the training recipe from Sec. 4.1, with 3K iteration training on referring data with context fixed at 32 frames. Validation is done on the held out ReasonVOS to analyze the zero-shot performance.

Table 3: Evaluation of different directional combinations of spatial and temporal scanning with pruning and merging strategies.

Spatial Scan	Temporal Scan		Prune	Merge	ReasonVOS		
	Uni-Dir	Bi-Dir			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✗	✗	✗	✗	✗	48.1	45.4	50.0
✓	✗	✗	✗	✓	44.1	41.0	47.1
✓	✗	✓	✗	✓	42.7	39.4	46.1
✗	✓	✗	✗	✓	45.2	42.2	48.3
✓	✓	✗	✗	✓	**49.6**	**47.2**	**52.0**
✓	✓	✗	✓	✗	47.9	45.4	50.4

Table 4: Comparison of scanning and compression ordering (spatial vs. temporal-first).

Scan		Comp.		ReasonVOS		
Sp. 1st	Tmp. 1st	Sp. 1st	Tmp. 1st	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
	✓	✓		31.3	27.5	35.2
	✓		✓	29.6	25.4	33.9
✓		✓		45.7	43.4	48.0
✓			✓	**49.6**	**47.2**	**52.0**

Compression strategy comparison. Fig. 3 evaluates the compression approaches across context lengths from C=2 to 64 frames under standardized batch processing. The main plot illustrates the memory-latency trade-off, where VANILLA [46] method exhibits quadratic scaling while compression methods maintain manageable footprints. The inset spider plot reveals performance that STAC consistently achieves the highest $\mathcal{J}\&\mathcal{F}$ scores with monotonic improvement as context increases, validating that our causal temporal design preserves long-range dependencies through selective state propagation. Perceiver [33] saturates early due to its fixed query set, while Pooling plateaus from uniform averaging that discards motion information. BIMBA [32] remains competitive at short contexts but requires full video access, limiting streaming deployment. Crucially, in online scenarios, STAC maintains a constant inference latency of ∼200ms per frame regardless of sequence length, a capability inaccessible to bidirectional baselines that exhibit linear or quadratic growth.

Spatiotemporal enhancement architecture. Tab. 3 validates our asymmetric scanning strategy, where bidirectional spatial scanning captures symmetric within-frame relationships while unidirectional temporal scanning enables streaming compatibility. Compared to without any enhancement or compression (48.1 $\mathcal{J}\&\mathcal{F}$), our full architecture reaches higher performance (49.6), confirming that globally-aware features enable superior token selection. Bidirectional scanning on both dimensions yield lower performance while(42.7) at the same time eliminates streaming capacity and introduces future information leakage, while unidirectional on spatial and temporal scan (44.1) preserves causality but compromises spatial accuracy. Temporal-only unidirectional method (45.2) outperforms bidirectional variants yet underperforms our dual-stage design. Comparing aggregation strategies, merging (49.6) significantly outperforms pruning (47.9) by retaining information through averaging rather than elimination.

Compression ordering. Tab. 4 evaluates the hierarchical sequence of our compression modules and reveals that temporal-first compression yields 49.6 $\mathcal{J}\&\mathcal{F}$ while significantly surpassing the 45.7$\mathcal{J}\&\mathcal{F}$achieved by spatial-first compression. This performance gap stems from the fact that spatial-first pooling prematurely eliminates the fine-grained inter-frame differences that constitute motion signals before the temporal scanner can identify salient regions. By con-

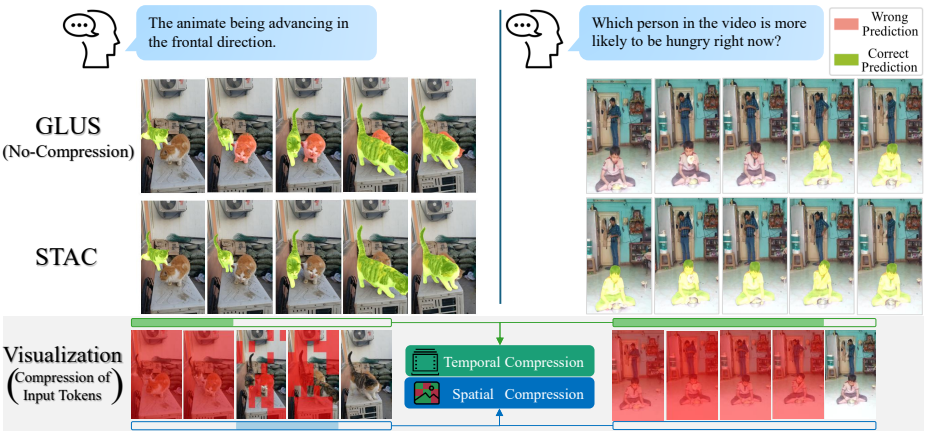

Fig. 4: Performance on Segmentation with reasoning shows that our method performs aggressive yet adaptive compression. While the RED areas indicate compressed frames or patches, the use of selective aggregation propagates all the important information within those patches without missing any critial information.

trast, our full four-stage pipeline consisting a cascade of spatial enrichment, temporal enrichment, temporal compression, and spatial compression optimizes the trade-off between context density and motion preservation. This specific sequence first captures motion patterns during the enrichment phase and prioritizes temporal reduction to ensure that critical dynamic features are preserved through the final spatial aggregation stage which architecturally enforces the principled separation of spatial context and causal temporal evolution.

Online vs. offline inference. With temporal recurrence being causal by design (Eq. (2),Eq. (3)), STAC runs identically in both modes, with only the EMA statistics (Eq. (5)) warming up over the first few frames. The cost is small on referring queries, where the online STAC reaches 72.5 $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS17 against 73.6 offline. The gap widens on reasoning (48.8 vs. 52.3 on ReasonVOS), where queries rely on future frames and get updated as the frames come in.

Qualitative analysis. A qualitative comparison between GLUS and STAC on complex reasoning queries is shown in Fig. 4. Both methods successfully segment targets, but STAC maintains comparable quality while using only 15% of tokens. For both queries, STAC correctly identifies and tracks the target throughout the sequence with greater temporal consistency than GLUS. The bottom visualization shows compression decisions, where red patches indicate discarded tokens. Temporal compression retains motion-rich frames while merging static ones, and spatial compression preserves foreground subjects while compressing the background. Through learned pooling, as discussed in Sec. 3, thresholds adapt automatically to content characteristics, achieving over 90% compression in static regions while retaining 30%-50% of tokens in dynamic sequences containing discriminative motion patterns essential for temporal reasoning.

5 Conclusion

This paper presents STAC, demonstrating that the redundancy signal intrinsic to SSM recurrence enables principled video token compression at linear cost. Exploiting this signal, STAC enriches features through selective state conditioning before compressing them, decoupling bidirectional spatial from causal temporal scanning to support streaming-compatible deployment, with task-grounded hierarchical compression to achieve 85% token reduction and 1.8× speedup while surpassing compression-free baselines on reasoning segmentation benchmarks. Zero-shot transfer results further suggest that enrichment-informed compression learns generalisable representations beyond the training distribution. Future directions include quantifying the relationship between recurrent state capacity and achievable compression ratios, and extending the framework to multimodal streams such as audio-visual reasoning and embodied perception.

Acknowledgements

This work is supported by the Natural Science Foundation of China (No. 62576176) and A*STAR Graduate Scholarship. The computational resources are supported by the Supercomputing Center of Nankai University (NKSC).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Adam, K.D.B.J., et al.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 **1412**(6) (2014)
3. An, Z., Li, Z., Ye, M., Qiao, F., Li, J., Wu, Z., Thengane, V., Li, C., Li, L., Gool, L.V., Sun, G., Belongie, S.: Video understanding: From geometry and semantics to unified models. Machine Intelligence Research (2026)
4. Ariff, S.H.S., Liu, Y., Sun, G., Yang, J., Ding, H., Geng, X., Jiang, X.: Evaluating sam2 for video semantic segmentation. Machine Intelligence Research (2026)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
6. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. In: NeurIPS. pp. 6833–6859 (2024)
7. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
8. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
9. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: CVPR. pp. 4985–4995 (2022)

10. Chen, G., Huang, Y., Xu, J., Pei, B., Chen, Z., Li, Z., Wang, J., Li, K., Lu, T., Wang, L.: Video mamba suite: State space model as a versatile alternative for video understanding. arXiv preprint arXiv:2403.09626 (2024)
11. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: ECCV. pp. 19–35. Springer (2024)
12. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al.: Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188 (2024)
13. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023)
14. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
15. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)
16. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: ICCV. pp. 2694–2703 (2023)
17. Ding, Z., Hui, T., Huang, J., Wei, X., Han, J., Liu, S.: Language-bridged spatial-temporal interaction for referring video object segmentation. In: CVPR. pp. 4964–4973 (2022)
18. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
19. Feng, Y., Yan, Z., Jia, Y., Chen, E.Q., Qin, J.: Training-free dense video captioning with large-scale pre-trained models. Machine Intelligence Research (2026)
20. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: CVPR. pp. 5958–5966 (2018)
21. Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., Lu, H.: The devil is in temporal token: High quality video reasoning segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29183–29192 (2025)
22. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024)
23. Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: Hippo: Recurrent memory with optimal polynomial projections. In: NeurIPS. vol. 33 (2020)
24. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
25. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: ICLR (2022)
26. Gu, A., Gupta, A., Goel, K., Ré, C.: On the parameterization and initialization of diagonal state space models. In: NeurIPS. vol. 35 (2022)
27. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state-space layers. In: NeurIPS. vol. 34 (2021)
28. Han, C., Fan, J., Wu, N., Dai, J., Bao, H., Lu, X.: Object-centric video prediction with mask-guided spatiotemporal diffusion. Machine Intelligence Research (2026)
29. He, S., Ding, H.: Decoupling static and hierarchical motion perception for referring video segmentation. In: CVPR. pp. 13332–13341 (2024)

30. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
31. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., et al.: M-llm based video frame selection for efficient video understanding. In: CVPR. pp. 13702–13712 (2025)
32. Islam, M.M., Nagarajan, T., Wang, H., Bertasius, G., Torresani, L.: Bimba: Selective-scan compression for long-range video question answering. In: CVPR. pp. 29096–29107 (2025)
33. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: Int. Conf. Machine Learning. pp. 4651–4664. PMLR (2021)
34. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv:2401.04088 (2024)
35. Jiang, J., Li, X., Liu, Z., Li, M., Chen, G., Li, Z., Huang, D.A., Liu, G., Yu, Z., Keutzer, K., et al.: Token-efficient long video understanding for multimodal llms. arXiv preprint arXiv:2503.04130 (2025)
36. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82**(1), 35–45 (1960)
37. Karacan, L., Sarıgül, M.: Full-frame video stabilization via spatiotemporal transformers. Computational Visual Media **11**(3), 655–667 (2025)
38. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Asian conference on computer vision. pp. 123–141. Springer (2018)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
40. Korbar, B., Tran, D., Torresani, L.: Scsampler: Sampling salient clips from video for efficient action recognition. In: ICCV. pp. 6232–6242 (2019)
41. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
42. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
43. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: ECCV. pp. 237–255. Springer (2024)
44. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
45. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
46. Lin, L., Yu, X., Pang, Z., Wang, Y.X.: Glus: Global-local reasoning unified into a single large language model for video segmentation. In: CVPR. pp. 8658–8667 (2025)
47. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, (Accessed: 2026-06-25)
48. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. vol. 36, pp. 34892–34916 (2023)

49. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. NeurIPS **37**, 103031–103063 (2024)
50. Liu, Y., Wu, Y.H., Sun, G., Zhang, L., Chhatkuli, A., Van Gool, L.: Vision transformers with hierarchical attention. Machine Intelligence Research **21**(4), 670–683 (2024)
51. Lu, H., Salah, A.A., Poppe, R.: Videomambapro: A leap forward for mamba in video understanding. arXiv e-prints pp. arXiv–2406 (2024)
52. Luo, Z., Xiao, Y., Liu, Y., Li, S., Wang, Y., Tang, Y., Li, X., Yang, Y.: Soc: Semantic-assisted object cluster for referring video object segmentation. In: NeurIPS. vol. 36, pp. 26425–26437 (2023)
53. Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Spectrum-guided multi-granularity referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 920–930 (2023)
54. Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan, F.S., Khan, S.: Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In: CVPR. pp. 19036–19046 (2025)
55. Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., Yuan, L.: Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. Computational Visual Media **12**(1), 71–84 (2026)
56. Pan, B., Panda, R., Jiang, Y., Wang, Z., Feris, R., Oliva, A.: Ia-red 2: Interpretability-aware redundancy reduction for vision transformers. In: NeurIPS. vol. 34, pp. 24898–24911 (2021)
57. Park, J., Kim, H.S., Ko, K., Kim, M., Kim, C.: Videomamba: Spatio-temporal selective state space model. In: ECCV. pp. 1–18. Springer (2024)
58. Pei, X., Huang, T., Xu, C.: Efficientvmamba: Atrous selective scan for light weight visual mamba. In: AAAI. pp. 6443–6451 (2025)
59. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Machine Learning. pp. 8748–8763. PmLR (2021)
61. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: NeurIPS. vol. 34, pp. 13937–13949 (2021)
62. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
63. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
64. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223. Springer (2020)
65. Shen, L., Hao, T., He, T., Zhao, S., Zhang, Y., Liu, P., Bao, Y., Ding, G.: Tempme: Video temporal token merging for efficient text-video retrieval. arXiv preprint arXiv:2409.01156 (2024)
66. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)

67. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR. pp. 18221–18232 (2024)
68. Tang, C., Han, Z., Sun, H., Zhou, S., Zhang, X., Wei, X., Yuan, Y., Xu, J., Sun, H.: Tspo: Temporal sampling policy optimization for long-form video language understanding. arXiv preprint arXiv:2508.04369 (2025)
69. Tang, J., Zheng, G., Yang, S.: Temporal collection and distribution for referring video object segmentation. In: ICCV. pp. 15466–15476 (2023)
70. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
71. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
72. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 (2023)
73. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European Conference on Computer Vision. pp. 396–416. Springer (2024)
74. Wang, Z., Shao, D., Zhang, L., Zhang, Z., Wang, B.: SAMDistill: SAM-based spatial-temporal distillation for robust 3d object detection. Machine Intelligence Research (2026)
75. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: ICCV. pp. 2761–2770 (2023)
76. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022)
77. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
78. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: ECCV. pp. 98–115. Springer (2024)
79. Yang, Y., Xing, Z., Yu, L., Huang, C., Fu, H., Zhu, L.: Vivim: A video vision mamba for medical video segmentation. arXiv preprint arXiv:2401.14168 (2024)
80. Ye, X., Gan, Y., Ge, Y., Zhang, X.P., Tang, Y.: Atp-llava: Adaptive token pruning for large vision language models. In: CVPR. pp. 24972–24982 (2025)
81. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
82. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
83. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
84. Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S., Wang, D.: Cobra: Extending mamba to multi-modal large language model for efficient inference. In: AAAI. pp. 10421–10429 (2025)
85. Zhu, J., Cheng, Z.Q., He, J.Y., Li, C., Luo, B., Lu, H., Geng, Y., Xie, X.: Tracking with human-intent reasoning. arXiv preprint arXiv:2312.17448 (2023)

86. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: Int. Conf. Machine Learning. pp. 62429–62442 (2024)