STAC：面向视频推理分割的选择性时空聚合与压缩

Syed Ariff Syed Hesham[1,2], 刘云[3,4,5⋆], 孙国磊[3,4,5], 杨静[6],
丁恒辉[7], 耿雪[2], 蒋旭东[1]

[1] 南洋理工大学电气与电子工程学院，新加坡
[2] 新加坡科技研究局（A*STAR）资讯通信研究院，新加坡
[3] 南开大学计算机学院 VCIP，中国天津
[4] 南开大学前沿交叉学科研究院，中国天津
[5] 南开国际先进研究院（深圳福田），中国深圳
[6] 贵州大学，中国贵州
[7] 复旦大学，中国上海

摘要 视频推理分割需要在复杂自然语言查询的引导下，对数百帧中的目标进行像素级精确跟踪。这会产生密集的时空标记（token），而自注意力的计算开销随 token 数量呈平方级增长，使长视频处理的成本难以承受。现有方法采用 token 压缩来应对这一问题，但通常直接处理缺乏时间上下文的编码器特征，在尚不能可靠评估内容冗余时就必须进行选择。合理的压缩决策需要上下文感知能力，但在完整分辨率下获取这种能力，也会产生随序列长度呈平方级增长的计算开销，而这恰恰是 token 压缩旨在降低的开销。状态空间模型能够解决这一矛盾：其线性递推机制以 $\mathcal{O}(T)$ 的计算代价，选择性地将时间上下文融入每个 token 的表示，从而使内容冗余得以评估。基于这一思路，选择性时空聚合与压缩（**S**elective Spatio**T**emporal **A**ggregation and **C**ompression，**STAC**）通过解耦的双向空间扫描和因果时间扫描增强特征，利用递推产生的冗余信号进行分层压缩，并以分割目标优化自适应阈值。在零样本且支持流式处理的设置下，STAC 将 token 数量减少 85%，实现 1.8× 加速，同时在推理分割基准上超越不进行压缩的基线。代码见此链接。

关键词： 视频推理分割 · 视频 token 压缩 · 时空建模 · 状态空间模型 · 多模态大语言模型

1 引言

大语言模型（Large Language Model，LLM）[1,18,34,73] 推动了人工智能领域的变革，在推理和任务完成方面展现出突出能力。在此基础上，多模态大语言模型（Multimodal Large Language Model，MLLM）将这些能力拓展至视觉理解，在图像理解 [5,13,42,48,49] 和视频分析 [14,46,56,71,74,84] 中取得了显著成果。其中，视频推理分割 [6,41,79] 是一项尤为复杂的挑战，要求模型理解复杂的自然语言查询，推理时空关系，并在长视频序列中生成像素级精确的分割掩码。

多数视频 MLLM 使用预训练图像编码器 [61,83] 独立编码每一帧，将得到的 token 拼接后输入语言模型 [5,48,49]。对于持续数秒的短片，这一策略行之有效：token 序列长度仍可控制，自注意力机制也能发现时间关系。然而，对于更长的

⋆ 通信作者：刘云 (liuyun@nankai.edu.cn)

Table 1: 面向推理分割的视频压缩方法的架构特性比较。括号中的勾号 (✓) 表示部分支持。

方法	时间整合	任务感知压缩	层次化理解	在线处理
局部压缩方法				
FastV [11]	✗	✗	✗	✗
ATP-LLaVA [81]	✗	✓	✗	✗
TempMe [66]	✓	(✓)	✓	✗
自适应压缩方法				
LongVU [67]	✓	✓	✗	✗
M-LLM [31]	✓	✓	✗	✗
TSPO [69]	✓	✓	✗	✗
双向 SSM 方法				
VideoMamba [43]	✓	✗	✗	✗
BIMBA [32]	✓	(✓)	✗	✗
STORM [35]	✓	✗	✗	✗
STAC（本文）	✓	✓	✓	✓

视频，注意力计算复杂度的平方级增长 [51] 与 token 数量的激增使该策略难以奏效。例如，一个以 2 FPS 采样、长 2 分钟的视频，若视觉编码器 [83] 每帧生成 729 个 token，将产生 61K 个 token，接近典型 32K 上下文窗口 [18,34] 的两倍，使长视频处理难以实际应用 [12]。

为应对这些挑战，多数长视频方法采用不同的压缩策略来缩短整体 token 序列。帧选择方法 [12,40,47,67] 针对时间冗余，但可能丢失跟踪所需的关键运动线索。另一类策略采用空间压缩 [11,68,69,81]，通过池化或基于注意力的聚合减少每帧的 token 数量。然而，局限于独立帧的处理方式难以利用跨帧依赖。近期工作 [31,46,67,84] 将两种策略结合起来，但压缩决策通常仍依赖预设启发式规则或帧级特征，而非根据全局时空上下文学习得到的选择机制。

这些压缩策略虽然有效，却具有一个共同的根本局限：它们直接作用于原始编码器特征，在模型理解视频的时空结构之前就进行 token 缩减。这使压缩与理解之间形成矛盾，因为以完整分辨率获取上下文感知能力，恰恰会产生压缩希望避免的开销。我们观察到，状态空间模型（State-Space Model，SSM）[23,25–27] 能够化解这一矛盾。其线性递推以 $\mathcal{O}(T)$ 的代价，选择性地根据当前及历史内容更新各个隐藏状态，使已被吸收的 token 产生近乎相同的增强表示。这提供了模型动态过程内生的冗余信号，是持续状态累积特有的能力。在 SSM 中，Mamba [22] 引入选择性状态传播机制 [24]，在过滤冗余的同时动态保留相关信息。视频领域的扩展工作 [10,43,52,58] 证实了这些机制在时间建模中的作用，但多数方法依赖统一的双向扫描 [32,35]，将整个时空体视为一个整体，阻碍了流式部署。这一局限背后存在根本的不对称性：帧内空间关系不受因果顺序约束，目标可同时与周围所有邻域交互；时间演化则遵循严格的因果顺序，未来帧不能影响过去的观测。因此，我们将空间扫描与时间扫描解耦，以兼顾效率和流式兼容性。

基于上述观察，选择性时空聚合与压缩（**S**elective Spatio**T**emporal **A**ggregation and **C**ompression，**STAC**）利用这种源于 SSM 的冗余信号，对视频 token

进行有依据的压缩。现有的基于 SSM 的视频方法仅将状态空间模型用作处理主干，而 STAC 在压缩前先通过 SSM 增强特征，从而使冗余判断基于递推机制已经吸收的信息，而非表面相似性。具体而言，状态引导的时空聚合器（State-informed Spatiotemporal Aggregator，SSA）首先采用双向空间扫描和因果时间扫描增强编码器特征，充分考虑空间维度与时间维度各自不同的结构特性；其中，因果设计不依赖未来观测，因此支持流式部署。随后，我们的分层状态自适应压缩（Hierarchical State-adaptive Compression，HSC）模块利用这一增强后的特征空间，按照“先时间、后空间”的层次化方式进行压缩。其中，自适应阈值会根据局部内容的动态变化进行调整，在静态片段中进行较强压缩，同时保留运动边界处的 token。最后，任务驱动优化将分割梯度反向传播至压缩策略，形成由下游任务直接决定保留哪些 token 的闭环。在零样本且支持流式处理的设置下，该设计减少约 85% 的 token，实现约 1.8× 加速，并在推理分割基准上超越不进行压缩的基线。

我们的主要贡献如下：

- **状态引导的时空聚合器（SSA）**：通过解耦的双向空间扫描和因果时间扫描增强编码器特征，在压缩前构建以语义信息为基础、使内容冗余清晰可辨的特征空间，并通过因果时间设计支持流式部署。
- **分层状态自适应压缩（HSC）**：在递推增强的特征空间中，利用响应局部内容动态的自适应阈值，按先时间、后空间的顺序减少 token，在对静态区域进行较强压缩的同时保留运动边界。
- **任务驱动的可微压缩**：通过直通估计，使分割梯度能够经过离散压缩决策反向传播，从而使 token 保留策略与下游分割掩码精度相一致，得到具有内容自适应能力的压缩策略，且能够零样本迁移到未见过的推理基准数据集。

2 相关工作

2.1 语言引导的视频分割

视频推理分割需要理解复杂的自然语言查询，生成像素级精确的掩码，并在长序列中保持时间一致性 [16, 41, 78]。传统指代表达视频目标分割采用具有显式跨模态融合的专用 Transformer [9, 20, 38]。例如，ReferFormer [77] 使用可变形注意力处理多帧，MTTR [9] 则引入端到端时间建模架构。但面对需要世界知识或多跳推断的复杂查询时，这些方法缺乏灵活推理能力 [47]。基于 MLLM 的方法将视觉基础模型与大语言模型结合 [39, 41, 49]；其中，LISA [41] 开创了将嵌入映射为掩码的推理分割范式，后续视频扩展引入时间建模 [6, 63, 79]。VISA [79] 采用带有时序传播机制的分层编码方法；VideoLISA [6] 提出稀疏–密集采样及单 token 分割全部帧（one-token-seg-all）的方式；GLUS [47] 则进行综合的全局–局部推理。与此同时，VRS-HQ [21] 引入分层的帧级 token 与时间 token，结合动态聚合机制和基于 SAM2 的遮挡感知关键帧选择；Sa2VA [82] 将 SAM2 与 LLaVA 统一到共享 token 空间，在该空间内，用 LLM 生成的指令 token 引导图像和视频中的掩码预测。尽管这些方法在离线场景下效果良好，但它们对所有视频 token 采用统一的处理方式，并没有显式地进行 token 压缩 [6, 47, 79]，即对所有视频帧赋予相同的计算权重，并且需要访问完整的视频序列。这种统一处理方式限制了模型的可扩展性，使其无法在线部署，也忽略了视频中的时间冗余；利用这些冗余，有望在保留推理所必需的细粒度运动线索的同时提高计算效率。

2.2 高效视频理解

传统视频理解方法已广泛探索高效时空建模，以减少冗余计算，同时保留对运动敏感的表示 [3, 4, 19, 28, 37, 75]。这一问题在现代基于 token 的视频推理系统中更加突出：视频 MLLM 使用预训练编码器 [49, 61, 83] 为每帧生成数百个 token，很快就会超出上下文限制 [18, 34]。*Token* 压缩通过学习式选择 [57, 62] 或基于注意力的剪枝 [8, 11, 81] 缓解空间冗余。DynamicViT [62] 对 token 进行分层剪枝，ToMe [8] 通过二分图匹配进行 token 合并，FastV [11] 在早期进行剪枝，ATP-LLaVA [81] 则引入逐层自适应阈值。视频扩展方法利用时间冗余 [35, 66, 68]，例如 TempMe [66] 合并跨帧 token，STORM [35] 使用 Mamba 层。帧选择通过均匀下采样 [12, 47] 或自适应评分 [31, 40, 67, 69] 降低时间维度。LongVU [67] 使用基于相似度的过滤机制，M-LLM [31] 和 TSPO [69] 则采用空间池化或强化学习。然而，这些方法往往依赖不适用于不同复杂度场景的固定比例，或采用缺乏监督的启发式方法，并且均需要以离线方式访问完整的视频序列。

2.3 用于视觉理解的状态空间模型

建立在经典理论 [36] 之上的状态空间模型通过结构化表示实现高效序列处理 [23, 25–27]。Mamba [22] 引入选择性扫描，通过依赖输入的状态传播实现线性复杂度，在过滤冗余的同时保留显著特征 [24]。视觉应用通过方向扫描适配这些原理 [50, 59, 87]：Vision Mamba [87] 使用双向扫描，VMamba [50] 引入交叉扫描机制。视频领域的扩展方法将 SSM 应用于时空特征体 [10, 43, 52, 58, 80]，其中 VideoMamba [43] 采用双向扫描，VideoMambaPro [52] 通过带掩码的反向计算缓解历史状态衰减。近期工作将 Mamba 融入视频 MLLM [32, 35, 85]，例如 BIMBA [32] 结合交错视觉查询进行双向时空 token 选择。然而，现有架构对完整视频序列进行统一双向扫描 [32, 35, 43, 58]，以相同方式处理空间与时间维度，忽视了时间演化的严格因果性与空间关系的对称性之间的差异。这要求访问完整视频，无法执行在线处理所需的逐帧增量计算 [32, 66, 76]。此外，已有工作侧重端到端处理 [32, 35]，而未利用选择性状态调节，在面向密集预测任务进行选择之前，先增强中间表示。

针对这些不足，我们通过选择性状态调节在模型上游压缩视觉 token，在架构上解耦双向空间扫描与因果时间扫描，并通过任务驱动优化使分割梯度能够通过压缩决策进行反向传播，从而实现高效且支持流式处理的推理分割。

3 方法

3.1 架构概述

视频推理分割需要在长视频序列中区分时间冗余内容与对语义理解至关重要的运动线索。直接压缩原始编码器特征缺乏这种时间感知能力，因此 STAC 在压缩之前引入选择性状态空间调节，确保 token 缩减依据的是融合时间上下文后的冗余，而非仅依据原始外观（Fig. 1）。给定视频 $\mathbf{V} \in \mathbb{R}^{T\times 3\times H\times W}$，其帧为 $\{\mathbf{I}_1, \ldots, \mathbf{I}_T\}$，以及文本查询 q，我们生成分割掩码 $\mathbb{M} = \{m_1, \ldots, m_T\}$，其中 $m_t \in \{0,1\}^{H\times W}$。遵循标准流程 [5, 48]，使用冻结的 CLIP ViT-L/14 [61] 编码视频帧，每帧生成 $N = 256$ 个 token，特征为 $\mathbf{F}_e \in \mathbb{R}^{T\times N\times d_e}$，其中 $d_e = 1024$。整个框架由三个级联阶段构成，每个阶段均遵循空间与时间维度各自的因果结构：

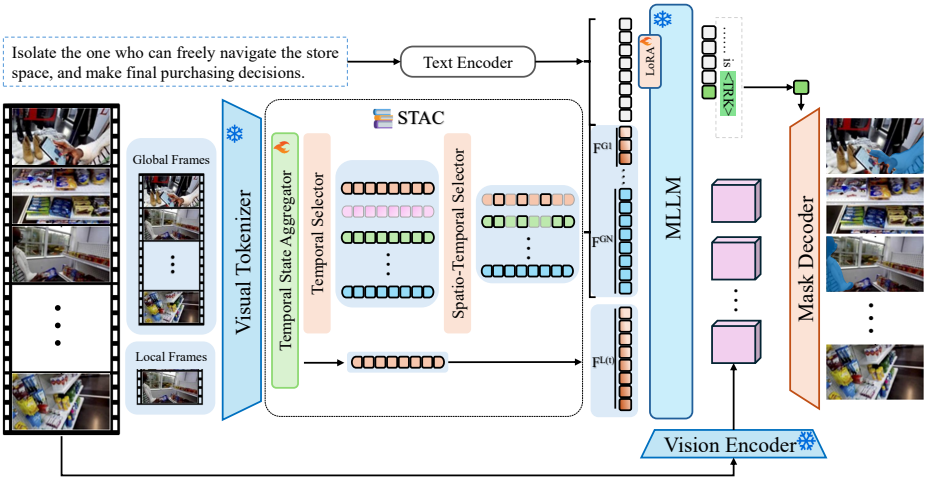


Fig. 1: 本架构引入通用时空摘要模块，通过时间选择和时空压缩降低计算开销，同时提升模型的整体性能。

1. **阶段 1：解耦时空扫描**。利用基于级联 Mamba 模块 [15, 22] 的解耦扫描架构，将独立帧 token 转换为具有全局感知能力的表示 $\mathbf{F}_{\text{ST}} \in \mathbb{R}^{T\times N\times d_e}$。不同于将时空特征体作为统一各向同性块处理的架构，我们系统地解耦两个维度：通过双向空间扫描保证语义完整性，通过严格因果的时间扫描建模运动演化。这种分离保证 $\mathcal{O}(TNd_e)$ 的线性复杂度，并防止未来帧信息泄漏，从而支持增量推理。
2. **阶段 2：非对称因果压缩**。为避免离线聚类造成的延迟瓶颈，我们采用非对称的“先预测、后压缩”策略：当前帧 $\mathbf{F}_{\text{ST}}[t]$ 在当前时间步以完整分辨率处理，而所有先前帧随后被压缩为紧凑的历史表示 $\mathbf{F}_c$（$T' \ll T$），作为时间记忆。
3. **阶段 3：任务目标优化**。将压缩特征 $\mathbf{F}_c$ 投影到 $D = 4096$ 维，生成用于掩码解码 [64] 的 ⟨TRK⟩ token，并将压缩视为可微组件。我们不依赖通用的启发式重建指标，而是进行端到端训练，通过直通估计 [7] 使分割梯度能够直接经过离散压缩决策反向传播，从而使压缩策略与分割任务目标直接对齐。

3.2 状态引导的时空聚合器

标准视觉编码器产生的特征 $\mathbf{F}_e$ 表现为缺乏全局或时间上下文的独立空间图块。这些表示的帧间相似度反映的是表面视觉相似性，而非语义冗余，从根本上限制了据此作出的压缩决策的质量。所提出的状态引导的时空聚合器（SSA）通过双阶段架构弥补了这一不足，在压缩之前利用时空上下文对每个 token 的表示进行增强。SSA 采用两个级联阶段完成时空上下文建模：先通过双向扫描整合帧内空间信息，再以因果方式跨帧传播时间信息，从而形成一个能够显式揭示内容冗余的特征空间。

空间聚合。　单帧内的空间关系不受因果顺序约束，各像素可同时与周围所有邻域交互。我们对全部 T 帧分别进行双向选择性状态扫描 [87]，以捕获全局上

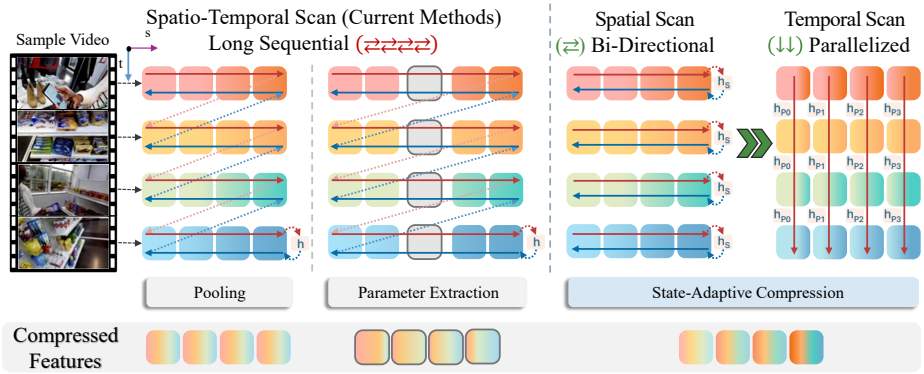


Fig. 2: 扫描策略比较。现有方法（左）将视频展平为长度为 $T \times H \times W$ 的序列，需要对完整视频进行双向处理，并采用固定池化 [35,43] 或可学习参数提取 [32] 进行压缩。我们的方法（右）将双向空间扫描与在空间位置上并行的因果时间扫描解耦，通过 HSC 模块实现支持流式处理的内容自适应压缩，并获得更高的 token 缩减幅度。

下文，避免强加人为的顺序偏置。具体地，对前向（$\mathcal{M}^b_{\rightarrow}$）和反向（$\mathcal{M}^b_{\leftarrow}$）状态扫描结果取平均：

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

这一操作有效消除了方向偏置，使所得特征能够表征精确的目标边界和丰富的帧内语义，为像素级分割提供必要的信息。

时间聚合。　与之相对，时间演化遵循严格的物理因果关系。为支持在线流式处理，我们遵循这一约束，采用因果选择性状态扫描 $\mathcal{M}_{\text{causal}}$ [15]，并在 N 个相互独立的空间位置上高效并行执行。该机制通过递推更新累积时间上下文：

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

其中，位置 n 的因果选择性状态空间机制通过以下过程累积时间上下文：

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

依赖输入的矩阵 $\bar{\mathbf{A}}_t, \bar{\mathbf{B}}_t \in \mathbb{R}^{d_e \times d_e}$ 在过滤冗余的同时，选择性传播显著运动模式。这种因果结构支持增量推理：每个新到达的帧更新隐藏状态 $\mathbf{h}_{n,t}$，无需重新计算过去的表示。相比之下，统一双向方法 [32,43,58] 需要访问完整视频，无法流式部署。最终获得的增强特征 $\mathbf{F}_{\text{ST}}$ 为后续压缩决策提供基础。

3.3 分层状态自适应压缩

所提出的分层状态自适应压缩（HSC）模块在增强特征空间 $\mathbf{F}_{\text{ST}}$ 中分阶段缩减 token；在该空间中，SSA 的选择性状态调节（Sec. 3.2）已将时间上下文融入每个 token。对于内容已被递推过程表示的连续帧，其输出近乎相同，从而揭示缺少时间整合的原始编码器特征无法显现的冗余。HSC 通过先时间、后空间的层

次结构（Table 3）利用这一冗余，先识别并合并冗余帧，再进行帧内空间缩减，因为过早的空间池化会破坏这一冗余信号所依赖的帧间差异。

自适应时间压缩。 视频中的相邻帧具有显著时间冗余 [40]，但运动信息恰恰存在于均匀下采样会丢弃的帧间差异中。我们通过高效的 $\mathcal{O}(T)$ 顺序相似度计算识别冗余。对于每一帧 t，先对空间位置取平均，得到帧级表示，再计算其与前一帧的相似度：

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

为克服固定超参数缺乏灵活性的问题并兼容流式处理，我们不使用全局统计量，而是通过指数移动平均（Exponential Moving Average，EMA）在线估计分布。令 μ_t 和 σ_t^2 分别表示第 t 步的运行均值和方差，动量 $\alpha = 0.1$：

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

自适应阈值定义为 $\tau_t = \mu_t + k\sigma_t$，其中参数 k 通过轻量级多层感知机（Multilayer Perceptron，MLP）学习。满足 $s_t > \tau_t$ 的连续帧通过取平均进行合并，得到压缩后的历史表示 $\mathbf{F}_{\mathrm{temp}}$。仅依据帧间内容动态，这一步通常即可减少 60%–70% 的 token。

自适应空间压缩。 时间压缩减少了帧数，但背景区域仍存在空间冗余。为此，我们在时间压缩后的序列上，为每个空间位置 n 计算跨帧一致性 c_n：

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

较高的一致性表示该图块在时间上保持静态，其跨帧表示适合合并；较低的一致性则突出需要保留的动态前景区域。我们将 Eq. (5) 中的统计量更新公式应用于各个空间位置，利用局部统计量 $(\mu_c^{(n)}, \sigma_c^{(n)})$ 计算自适应阈值。对于超过阈值的位置，将其时间维上的 token 取平均并合并；动态位置则保留完整的时间分辨率。这一步主要针对静态区域，进一步减少 40%–50% 的 token，得到压缩特征 $\mathbf{F}_c$，其 token 总数满足 $|\mathbf{F}_c| \ll T' \times N$，两步合计减少约 85%。

3.4 任务驱动的可微压缩

以重建为目标优化的标准压缩指标，未必能保证保留对分割至关重要的边界。本框架将压缩作为可微组件融入分割流程，使任务梯度通过联合目标直接影响 token 保留：

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

其中，$\mathcal{L}_{\mathrm{seg}}$ 包含标准交叉熵损失和 Dice 损失 [64]，$\lambda = 0.001$ 用于平衡两个目标。压缩正则项 $\mathcal{L}_{\mathrm{comp}}$ 随视频复杂度自适应变化：

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

这里，B 为批大小，c_i 为保留的 token 数，T_i 为原始 token 数；$\hat{C}_i = \frac{1}{2}[(1-\mu_i)+\sigma_i]$ 基于特征方差估计视频复杂度，使动态程度较高的视频受到更小的压缩惩罚；平方根归一化因子 $1/\sqrt{T_i}$ 保证长序列下的次线性缩放。这种缩放降低每个 token 的惩罚，在鼓励提高计算效率的同时，避免过度压缩长序列而丢失关键运动信息。

压缩与分割之间的对齐通过Sec. 3.3中的学习式阈值实现，分割梯度借助直通估计器 [7] 调整这些阈值。尽管选择过程是离散的，阈值演化（Eq. (5)）仍能稳定这些更新，使训练收敛到内容自适应策略：对静态片段进行较强压缩，同时保留运动边界处的 token。该策略形成完整设计闭环：解耦扫描（Sec. 3.2）利用时空上下文对特征进行增强，HSC 将由此显现的冗余转化为压缩决策，而任务目标根据分割精度校准这些决策。如Fig. 4所示，静态视频的压缩幅度可超过90%，而动作序列保留 30%–50% 的 token，以保证边界精度。

4 实验与分析

4.1 实验设置

数据集。 我们在五个基准上评估模型，并按查询复杂度将其分为两类。*指代表达基准*要求根据语言描述定位目标。Refer-YouTube-VOS [78] 包含 3,978 个视频和 15,458 条表达，侧重外观属性与空间关系；MeViS [16] 包含 2,006 个视频和 28,570 条表达，需要通过运动模式而非静态特征区分目标。Ref-DAVIS17 [38] 在 DAVIS-2017 [60] 基础上扩展了语言标注，包含 90 个视频，用于评估遮挡和外观变化下的性能。*推理基准*要求结合世界知识进行多步推断。ReVOS [79] 包含 1,042 个视频和 35,074 个指令–掩码对，要求进行超越直接视觉匹配的上下文推断；ReasonVOS [6] 包含 91 个视频和 458 个样本，专门评估长上下文中的时间理解与因果推理。我们在官方验证划分上报告区域相似度 $\mathcal{J}$、轮廓精度 $\mathcal{F}$ 及其平均值 $\mathcal{J}\&\mathcal{F}$。

实现细节。 我们使用预训练且冻结的 CLIP ViT-L/14 [61] 作为视觉编码器，输入分辨率为 224×224，每帧产生 256 个 1024 维 token。特征首先经过保持 1024 维隐藏状态的双阶段 SSA 模块进行时空增强，再由 HSC 模块按照Sec. 3中的学习式自适应阈值进行分层压缩。压缩 token 被投影为 4096 维后输入 LLaVA-7B [49, 72]。低秩适配（Low-Rank Adaptation，LoRA）微调 [30] 仅应用于注意力投影矩阵，秩为 8、α=16，语言模型的基础参数保持冻结。随后，SAM2 [64] 将与语言对齐的嵌入解码为分割掩码。训练仅使用指代表达分割数据集 MeViS [16] 和 Refer-YouTube-VOS [78]，不使用任何推理分割数据，以在 Ref-DAVIS17 [38]、ReVOS [79] 和 ReasonVOS [6] 上评估零样本迁移。训练采用 AdamW [2] 优化器，β_1=0.9、β_2=0.999，权重衰减为 0.05，学习率为 3×10^{-4}，并进行 100 步线性预热。每次迭代使用批大小 2，每个视频动态采样 2–32 个上下文帧。训练在 2× NVIDIA A40 GPU 上运行 3,000 次迭代，耗时约 30 小时。

基线。 我们与两类基线比较，以区分压缩策略的影响。*无压缩基线*包括 ReferFormer [77]、OnlineRefer [76] 和 TempCD [70] 等专用指代表达分割架构，以及基于 MLLM 的 LISA [41]、VISA [79]、VideoLISA [6] 和 GLUS [47]；这些方法处理完整 token 序列，不进行显式缩减。*压缩基线*包括均匀缩减时空维度的 3D 池化、以 Transformer 层替换 Mamba 模块并根据注意力分数剪枝的注意力选择方法、通过可学习跨注意力查询进行聚合的 Perceiver [33]，以及采用双向状态空间压缩的 BIMBA [32]。所有基线均采用官方实现，并将超参数与我们的配置对齐，在相同缩减比例下进行公平比较。

Table 2: 与指代表达及推理视频目标分割基准上的先进方法比较。报告 $\mathcal{J}\&\mathcal{F}$、$\mathcal{J}$ 和 $\mathcal{F}$ 分数。其他方法处理完整 token 序列，而 STAC 通过自适应压缩仅使用约 **15% 的视觉 token**，仍在各基准上取得有竞争力或更优的性能。

模型	Ref-DAVIS17			MeViS			Ref-YouTube			ReasonVOS			ReVOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
指代表达视频目标分割方法															
URVOS [65]	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 [70]	-	-	-	-	-	-	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 [76]	64.8	61.6	67.7	-	-	-	63.5	61.6	65.5	38.7	34.6	42.9	-	-	-
SOC [53]	-	-	-	-	-	-	67.3	65.3	69.3	35.9	33.3	38.5	-	-	-
SgMg [54]	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 [77]	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
基于多模态大语言模型的推理方法															
VideoGLaMM [55]	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 [86]	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 [86]	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 [79]	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 [79]	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 [47]	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（本文）	**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

4.2 主要结果

Table 2 比较了 STAC 与指代表达及推理视频目标分割领域的先进方法。尽管视觉 token 减少 85%，STAC 在强调空间定位和时间推理的基准上仍表现出色。在 Ref-DAVIS17 [38] 和 Ref-YouTube [78] 上，我们的方法超越所有无压缩基线，说明即使存在遮挡，上下文感知的选择也能增强基于外观的指代表达分割。尤其值得关注的是，STAC 虽然仅使用指代表达分割数据训练，却能有效迁移至推理基准，在零样本设置下于 ReasonVOS [6] 上超越最强的无压缩基线。这提示，有依据的压缩不仅能够保留推理能力，还可能通过促使模型仅保留运动关键信息、过滤无压缩方法所处理的静态冗余，进一步改善推理能力。

在 MeViS [16] 和 ReVOS [79] 等以运动为核心的基准上，能够利用完整序列的双向时间信息的方法表现略优。然而，STAC 在输入 MLLM 之前通过上游选择性状态调节压缩 85% 的视觉 token，仍达到接近的性能，其因果时间设计还支持仅利用过去观测进行逐帧在线处理。此外，这种先增强、后压缩的上游策略在 token 进入 MLLM 之前就缓解了平方级注意力瓶颈，说明仅通过因果处理，选择性状态调节便能捕获足够的时间上下文。由此，STAC 在现有方法中兼具较强的分割精度、显著的 token 压缩和实时部署能力。

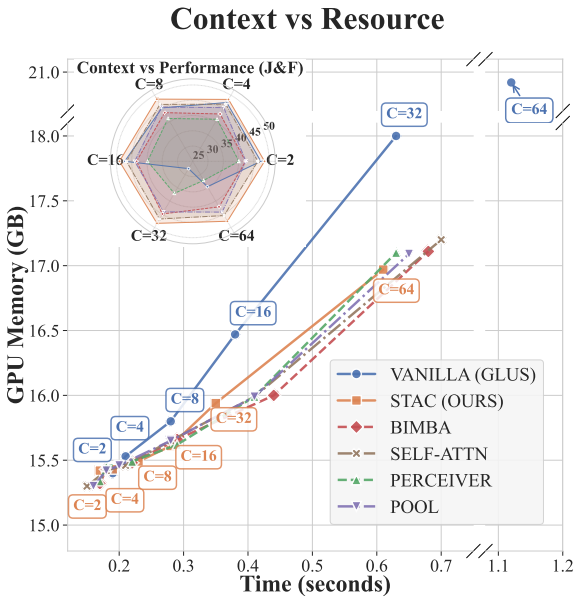


Fig. 3: 不同压缩方法的性能评估。STAC 保持了较高效率（主图），并在不同上下文长度下维持稳定表现（插图）。

4.3 消融实验

所有消融实验遵循 Sec. 4.1 中的训练方案，在指代表达数据上训练 3,000 次迭代，上下文固定为 32 帧，并在未参与训练的 ReasonVOS 上验证，以分析零样本性能。

压缩策略比较。 Fig. 3 在标准化批处理设置下，评估上下文长度从 C=2 到 64 帧时各类压缩方法的表现。主图展示显存占用与推理延迟的权衡：VANILLA [47] 方法呈二次增长，而压缩方法保持可控的资源占用。插入的雷达图显示，STAC 始终取得最高的 $\mathcal{J}\&\mathcal{F}$ 分数，且随着上下文增加而单调改善，验证了因果时间设计可通过选择性状态传播保留长程依赖。Perceiver [33] 由于查询集合是固定的，其性能会较早趋于饱和；Pooling 则因均匀平均丢失运动信息而进入平台期。BIMBA [32] 在较短上下文中仍具竞争力，但需要访问完整视频，限制了流式部署。关键在于，在线场景下 STAC 的单帧推理延迟保持在约 200 ms，不随序列长度变化；推理开销随序列长度呈线性或平方级增长的双向基线无法实现这一点。

时空增强架构。 Table 3 验证了我们的非对称扫描策略：双向空间扫描捕获对称的帧内关系，单向时间扫描则支持流式处理。与不进行增强和压缩的配置（$\mathcal{J}\&\mathcal{F}$ 为 48.1）相比，完整架构取得更高性能（49.6），说明具有全局感知能力的特征能改善 token 选择。在空间和时间两个维度上都采用双向扫描时性能较低（42.7），同时失去流式能力并引入未来信息泄漏；空间与时间均采用单向扫描时（44.1），虽然保持因果性，但空间精度受损。仅采用单向时间扫描的方法（45.2）优于双向变体，但不及我们的双阶段设计。在比较不同的聚合策略时，合

Table 3: 空间与时间扫描方向的不同组合，以及剪枝和合并策略的评估。

空间扫描	时间扫描 单向	时间扫描 双向	剪枝	合并	ReasonVOS $\mathcal{J}\&\mathcal{F}$	ReasonVOS $\mathcal{J}$	ReasonVOS $\mathcal{F}$
✗	✗	✗	✗	✗	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: 扫描与压缩顺序比较：空间优先与时间优先。

扫描 空间优先	扫描 时间优先	压缩 空间优先	压缩 时间优先	ReasonVOS $\mathcal{J}\&\mathcal{F}$	ReasonVOS $\mathcal{J}$	ReasonVOS $\mathcal{F}$
	✓	✓		31.3	27.5	35.2
	✓		✓	29.6	25.4	33.9
✓		✓		45.7	43.4	48.0
✓			✓	**49.6**	**47.2**	**52.0**

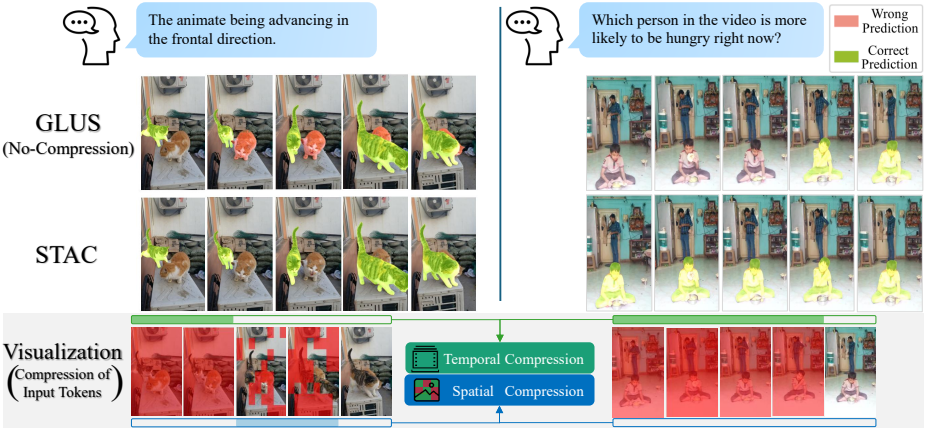


Fig. 4: 推理分割结果表明，我们的方法能够进行较强且自适应的压缩。红色区域表示被压缩的帧或图块，选择性聚合传播这些图块中的全部重要信息，避免丢失任何关键信息。

并策略（49.6）的性能显著优于剪枝策略（47.9），因为合并是通过平均操作来保留信息，而不是直接删除信息。

压缩顺序。 Table 4 评估压缩模块的层次化顺序。时间优先压缩取得 49.6 的 $\mathcal{J}\&\mathcal{F}$，显著高于空间优先压缩的 45.7。性能差距源于空间优先池化在时间扫描器识别出显著区域之前，就过早消除了构成运动信号的细粒度帧间差异。相比之下，我们由空间增强、时间增强、时间压缩和空间压缩级联组成的完整四阶段流程，优化了上下文密度与运动保留之间的权衡。这一顺序首先在增强阶段捕获运动模式，再优先进行时间缩减，确保关键动态特征能够在最终空间聚合后得到保留，从架构上落实空间上下文与因果时间演化的合理分离。

在线与离线推理。 由于时间递推在设计上具有因果性（Eq. (2)、Eq. (3)），STAC 在两种模式下采用相同的运行机制，区别仅在于 EMA 统计量（Eq. (5)）需要在最初几帧逐渐稳定。在指代表达查询上，性能下降较小：在线 STAC 在 Ref-DAVIS17 上取得 72.5 的 $\mathcal{J}\&\mathcal{F}$，离线为 73.6。在推理任务上，差距扩大，ReasonVOS 上分别为 48.8 和 52.3；这类查询依赖未来帧，并会随着新帧到达而更新。

定性分析。 Fig. 4展示了 GLUS 与 STAC 在复杂推理查询上的定性比较。两种方法均成功分割目标，但 STAC 仅使用 15% 的 token 就保持了相近的分割质量。对于两个查询，STAC 均能在整个序列中正确识别并跟踪目标，时间一致性优于 GLUS。底部可视化展示压缩决策，红色图块表示被移除的 token。时间压缩保留运动丰富的帧并合并静态帧，空间压缩则保留前景主体并压缩背景。通过Sec. 3中介绍的学习式池化机制，阈值自动适应内容特征：静态区域的 token 压缩幅度超过 90%，而包含时间推理所需判别性运动模式的动态序列保留 30%–50% 的 token。

5 结论

本文提出 STAC，表明 SSM 递推内生的冗余信号能够以线性代价支持有依据的视频 token 压缩。STAC 利用该信号，通过选择性状态调节先增强特征再进行压缩，并将双向空间扫描与因果时间扫描解耦，以支持流式部署。结合任务驱动的层次化压缩，STAC 减少 85% 的 token，实现 1.8× 加速，同时在推理分割基准上超越无压缩基线。零样本迁移结果进一步表明，基于增强特征的压缩能够学习到可泛化至训练分布之外的特征表示。未来研究方向包括量化递推状态容量与可达压缩比例之间的关系，以及将该框架拓展至视听推理和具身感知等多模态流。

致谢

本工作得到国家自然科学基金(编号 62576176)以及新加坡科技研究局(A*STAR) MTC 计划性基金（资助号 M23L7b0021）的资助。计算资源由南开大学超级计算中心（NKSC）提供支持。

参考文献

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. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. arXiv preprint arXiv:2202.07800 (2022)
46. 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)

47. 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)
48. 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/
49. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. vol. 36, pp. 34892–34916 (2023)
50. 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)
51. 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)
52. Lu, H., Salah, A.A., Poppe, R.: Videomambapro: A leap forward for mamba in video understanding. arXiv e-prints pp. arXiv–2406 (2024)
53. 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)
54. 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)
55. 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)
56. 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)
57. 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)
58. 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)
59. Pei, X., Huang, T., Xu, C.: Efficientvmamba: Atrous selective scan for light weight visual mamba. In: AAAI. pp. 6443–6451 (2025)
60. 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)
61. 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)
62. 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)
63. 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)
64. 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)

65. 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)
66. 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)
67. 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)
68. 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)
69. 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)
70. Tang, J., Zheng, G., Yang, S.: Temporal collection and distribution for referring video object segmentation. In: ICCV. pp. 15466–15476 (2023)
71. 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)
72. 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)
73. 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)
74. 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)
75. 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)
76. 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)
77. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022)
78. 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)
79. 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)
80. 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)
81. 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)
82. 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)
83. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)

84. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
85. 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)
86. 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)
87. 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)

附录

A.1 补充实现细节

A.1.1 架构与训练配置

视觉编码器与时空聚合。 视频帧分别经过预训练的 CLIP ViT-L/14 [61]，得到帧级编码特征 $\mathbf{F}_e \in \mathbb{R}^{T\times N\times d_e}$。随后，编码特征输入我们提出的状态引导的时空聚合器（SSA），以时空上下文增强编码器特征。SSA 模块通过两个顺序执行的阶段处理这些特征：首先，空间聚合阶段采用双向状态扫描 [87] 捕获对称的帧内关系，对全部 T 帧独立并行处理；随后，通过在空间位置上并行执行的因果扫描累积时间上下文。该过程可表示为：

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

因果选择性状态空间机制通过以下操作累积时间上下文 [15]：

$$\begin{aligned} \mathbf{h}_{n,t} &= \bar{\mathbf{A}}_t \mathbf{h}_{n,t-1} + \bar{\mathbf{B}}_t \mathbf{F}_{\text{spatial}}[t,n], \\ \mathbf{F}_{\text{ST}}[t,n] &= \bar{\mathbf{C}}_t \mathbf{h}_{n,t}, \quad n = 1,\ldots,N. \end{aligned} \tag{A.2}$$

其中，$\mathcal{M}^{b}_{\rightarrow}$ 和 $\mathcal{M}^{b}_{\leftarrow}$ 分别为双向选择性扫描中的前向与反向扫描，$\mathcal{M}_{\text{causal}}$ 表示因果时间扫描。对前向和反向扫描取平均可在编码精确空间上下文的同时消除方向偏置；Eq. (A.2) 中完整的状态到输出表达式通过额外的输出矩阵 $\bar{\mathbf{C}}_t$ 得到增强特征 $\mathbf{F}_{\text{ST}}[t,n] = \bar{\mathbf{C}}_t \mathbf{h}_{n,t}$，其中依赖输入的矩阵 $\bar{\mathbf{A}}_t, \bar{\mathbf{B}}_t, \bar{\mathbf{C}}_t \in \mathbb{R}^{d_e\times d_e}$ 共同实现选择性传播并过滤冗余。SSA 模块由一对空间聚合和时间聚合组成，二者均保持 $d_e = 1024$ 维隐藏状态，内部状态维度为 $d_{\text{state}} = 16$，扩展因子为 2。该扫描策略随机初始化并从头训练，构成泛化至推理任务的核心机制。

分层压缩。 经 SSA 进行时空增强后，分层状态自适应压缩（HSC）模块通过学习式自适应阈值依次缩减 token。时间压缩首先计算连续帧之间的帧级相似度分数 s_t，以识别运动关键帧。自适应阈值 τ_{temp} 根据各帧与前一帧的相似度确定合并候选（$s_t > \tau_{\text{temp}}$），生成二值保留掩码 $\mathbf{m}_{\text{temp}} \in \{0,1\}^T$，其中 $\mathbf{m}_{\text{temp}}[t] = 1$ 表示应保留帧 t。随后，空间压缩独立处理时间压缩后保留的每一帧，沿时间维度计算各位置的一致性分数 c_n 和自适应阈值 $\tau^{(n)}_{\text{spatial}}$，为每帧生成空间保留掩码 $\mathbf{m}^{(t)}_{\text{spatial}} \in \{0,1\}^N$。

Table A.1: STAC 的训练配置。

配置项	**取值**
上下文帧数	2-32
查询帧数	4
输入分辨率	224
特征下采样率	4
优化器	AdamW
优化器动量	$\beta_1, \beta_2 = 0.9, 0.999$
优化器权重衰减	0.05
学习率	3×10^{-4}
LoRA 秩	8
LoRA α	16
梯度累积	10
批大小（每设备、每步）	2
总批大小	40
预热步数	100
训练总步数	3,000
损失：λ_{ce}	1.0
损失：λ_{dice}	0.5
损失：λ_{bce}	2.0
损失：λ_{comp}	0.001

合并操作将信息整合到保留位置（$\mathbf{m}[i] = 1$）。对于每个保留位置 i，通过取平均，将其前方连续的所有未保留位置汇聚到一起：

$$\mathbf{F}_{\text{summary}}[i] = \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} \mathbf{F}[j], \quad \text{where } \mathcal{G}_i = \{j : i_{\text{prev}} < j \leq i, \mathbf{m}[j] = 0 \text{ for } j \neq i\} \cup \{i\}. \tag{A.3}$$

其中，$\mathcal{G}_i$ 表示合并到位置 i 的连续索引组。结合压缩与基于掩码 m 的提取操作，得到压缩特征 $\mathbf{F}_c \in \mathbb{R}^{T' \times N' \times d_e}$，在整合静态内容的同时保留运动关键帧和动态空间区域。压缩表达式及阈值学习的更多细节见主论文。

多模态大语言模型。 压缩特征经过视觉到语言投影层 $\phi_{V \to L}$，从视觉维度 $d_e = 1024$ 映射到语言模型维度 $D = 4096$：

$$\mathbf{F}_{\text{LLM}} = \phi_{V \to L}(\mathbf{F}_c). \tag{A.4}$$

语言模型随后以自回归方式处理这些特征，生成分割 token。对于每个需要分割的帧 t，MLLM 同时接收提供全局上下文的压缩摘要和该帧全分辨率的时空增强特征：

$$\langle \text{SEG} \rangle_t = \text{LLM}([R, \phi_{V \to L}(\mathbf{F}_c), \phi_{V \to L}(\mathbf{F}_{\text{ST}}[t])]), \tag{A.5}$$

其中，R 表示指代表达文本。该混合策略通过压缩历史上下文提高效率，并通过全分辨率的当前帧特征保证精度，从而在二者之间取得平衡。

掩码解码器。 我们的掩码解码器沿用 SAM2 [64] 的设计。得到 $\langle \text{SEG} \rangle_t$ 后，token 嵌入 $\hat{\mathbf{u}}_t$ 经过语言到视觉投影层 $\phi_{L \to V}$，生成掩码解码器的提示。掩码解码器 Dec 处理该提示及视觉编码器 Enc 输出的帧编码特征，同时在记忆库中维护特征记忆，以生成分割掩码。

$$M_t = \text{Dec}(\text{Enc}(\mathbf{I}_t), \phi_{L \to V}(\mathbf{u}_t), \text{FeatBank}). \tag{A.6}$$

其中，记忆库存储先前生成的掩码的编码特征，并对最多 7 个选定的历史掩码施加注意力，以提供时间一致性。所有解码器与记忆组件均与 STAC 框架进行端到端联合微调。

训练配置。 我们使用预训练 SAM2（sam2_hiera_large）[64] 权重初始化图像编码器 Enc、掩码解码器 Dec、记忆编码器和记忆注意力模块。LLM 骨干网络（LLaVA-7B [49, 72]）、投影层（$\phi_{V \to L}$、$\phi_{L \to V}$）和视觉 token 化器（CLIP ViT-L/14 [61]）均由已有的预训练权重 [41] 初始化。LLM 通过 LoRA [30] 微调，秩为 $r = 8$、缩放系数为 $\alpha = 16$，仅应用于注意力投影矩阵，LoRA 参数随机初始化。训练目标为：

$$\mathcal{L}_{\text{total}} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{comp}\mathcal{L}_{comp}, \tag{A.7}$$

该训练目标将主论文中的 $\mathcal{L}_{\text{seg}}$ 展开为交叉熵损失（$\mathcal{L}_{ce}$）、Dice 损失（$\mathcal{L}_{dice}$）和逐像素 BCE 损失（$\mathcal{L}_{bce}$），并加入压缩正则项（$\mathcal{L}_{comp}$），其中 $\lambda_{comp} = 0.001$。训练超参数汇总于 Table A.1。

A.1.2 推理流程

STAC 支持针对不同部署场景设计的两种推理模式。

Algorithm 1 STAC 离线推理

输入： 视频 $\mathbf{V} = \{\mathbf{I}_1, \ldots, \mathbf{I}_T\}$, 文本查询 q
输出： 分割掩码 $\mathbb{M} = \{m_1, \ldots, m_T\}$
$\mathbf{F}_e \leftarrow \text{Encoder}(\mathbf{V})$
// 阶段 1：状态引导的时空聚合器（SSA）
// 空间聚合
for $t = 1$ to T **并行执行 do**
  $\mathbf{F}_{\text{spatial}}[t] \leftarrow \frac{1}{2}\left(\mathcal{M}^b_{\rightarrow}(\mathbf{F}_e[t]) + \mathcal{M}^b_{\leftarrow}(\mathbf{F}_e[t])\right)$
end for
// 时间聚合
for $n = 1$ to N **并行执行 do**
  $\mathbf{h}_{n,t} \leftarrow \bar{\mathbf{A}}_t \mathbf{h}_{n,t-1} + \bar{\mathbf{B}}_t \mathbf{F}_{\text{spatial}}[t, n], \quad t = 1, \ldots, T$
  $\mathbf{F}_{\text{ST}}[:, n] \leftarrow \mathcal{M}_{\text{causal}}(\mathbf{F}_{\text{spatial}}[:, n])$
  存储隐藏状态：$\mathbf{h}^*_n \leftarrow \mathbf{h}_{n,T}$
end for
// 阶段 2：自适应分层压缩
$\tau_{\text{temp}} \leftarrow \text{AdaptiveThreshold}(\mathbf{F}_{\text{ST}}, \text{temporal})$
$\mathbf{F}_{\text{temp}} \leftarrow \text{Compress}_{\text{temporal}}(\mathbf{F}_{\text{ST}}, \tau_{\text{temp}})$ // $T \rightarrow T'$
$\{\tau^{(n)}_{\text{spatial}}\}_{n=1}^{N} \leftarrow \text{AdaptiveThreshold}(\mathbf{F}_{\text{temp}}, \text{spatial})$
$\mathbf{F}_c \leftarrow \text{Compress}_{\text{spatial}}(\mathbf{F}_{\text{temp}}, \{\tau^{(n)}_{\text{spatial}}\})$ // $N \rightarrow N'$
// 始终保留最后一帧的完整特征。
// 阶段 3：逐帧分割
$\mathbf{F}_{\text{summary}} \leftarrow \text{Proj}(\mathbf{F}_c)$
for $t = 1$ to T **do**
  $\mathbf{F}^{(t)}_{\text{full}} \leftarrow \text{SSA}(\mathbf{I}_t, \{\mathbf{h}^*_n\}_{n=1}^{N})$ // 增强帧特征
  $\text{trk}_t \leftarrow \text{MLLM}(q, \mathbf{F}_{\text{summary}}, \text{Proj}(\mathbf{F}^{(t)}_{\text{full}}))$
  $m_t \leftarrow \text{Mask-Predictor}(\text{trk}_t, \mathbf{I}_t)$
end for
返回 $\{m_t\}_{t=1}^{T}$

离线批处理。 当能够访问完整视频时，STAC 通过 SSA 模块处理全部 T 帧，以全面的时空上下文增强每个 token。随后，分层压缩模块依次执行先时间、后时空的压缩，生成由 $T'N'$ 个 token 组成的紧凑摘要。

为生成掩码，STAC 采用兼顾全局效率与局部精度的混合策略。表示整个视频上下文的压缩摘要 $\mathbf{F}_c$ 为 MLLM 提供全局理解。对于每个需要分割的帧 I_t，该架构提供全分辨率的时空增强特征 $\mathbf{F}_{\text{ST}}[t]$，其中包含该帧全部 N 个空间 token。MLLM 同时处理 $[\phi_{V\rightarrow L}(\mathbf{F}_c), \phi_{V\rightarrow L}(\mathbf{F}_{\text{ST}}[t])]$，利用压缩历史上下文进行时间推理，同时保留完整空间细节，以精确定位边界。这避免了仅依赖压缩表示进行分割时可能出现的细粒度空间信息丢失。完整离线流程见 Algorithm 1。

Algorithm 2 STAC 在线推理

输入： 视频流 $\{\mathbf{I}_1, \mathbf{I}_2, \ldots\}$, 文本查询 q
输出： 掩码 $\{m_1, m_2, \ldots\}$ （增量输出）
// 初始化
$\mathbf{F}_{\text{summary}} \leftarrow \emptyset$, $\{\mathbf{h}_{n,0}\}_{n=1}^{N} \leftarrow 0$, $t \leftarrow 0$
$\{\mu, \sigma\}_{\text{temp/spatial}} \leftarrow 0$ // 运行统计量
while 新帧 $\mathbf{I}_t$ 到达 **do**
  $t \leftarrow t + 1$
  $\mathbf{F}_e[t] \leftarrow \text{Encoder}(\mathbf{I}_t)$
  // 阶段 1：增量 SSA
  // 空间聚合（逐帧独立）
  $\mathbf{F}_{\text{spatial}}[t] \leftarrow \frac{1}{2}\left(\mathcal{M}^{b}_{\rightarrow}(\mathbf{F}_e[t]) + \mathcal{M}^{b}_{\leftarrow}(\mathbf{F}_e[t])\right)$
  // 时间聚合（更新隐藏状态）
  for $n = 1$ to N **并行执行 do**
    $\mathbf{h}_{n,t} \leftarrow \bar{\mathbf{A}}_t \mathbf{h}_{n,t-1} + \bar{\mathbf{B}}_t \mathbf{F}_{\text{spatial}}[t, n]$
    $\mathbf{F}_{\text{ST}}[t, n] \leftarrow \mathcal{M}_{\text{causal}}(\mathbf{F}_{\text{spatial}}[t, n], \mathbf{h}_{n,t})$
  end for
  // 阶段 2：自适应压缩
  更新 $\{\mu, \sigma\}_{\text{temp}}$ // 使用帧 t
  $\tau^{(t)}_{\text{temp}} \leftarrow \text{AdaptiveThreshold}(\{\mu, \sigma\}_{\text{temp}})$
  $\mathbf{F}_{\text{temp}}[t] \leftarrow \text{Compress}_{\text{temporal}}(\mathbf{F}_{\text{ST}}[t], \tau^{(t)}_{\text{temp}})$
  更新 $\{\mu, \sigma\}_{\text{spatial}}$ // 使用帧 t
  $\{\tau^{(n)}_{\text{spatial}}\}^{(t)} \leftarrow \text{AdaptiveThreshold}(\{\mu^{(n)}_c, \sigma^{(n)}_c\})$
  $\mathbf{F}_c[t] \leftarrow \text{Compress}_{\text{spatial}}(\mathbf{F}_{\text{temp}}[t], \{\tau^{(n)}_{\text{spatial}}\}^{(t)})$
  $\mathbf{F}_{\text{summary}} \leftarrow \mathbf{F}_{\text{summary}} \oplus \mathbf{F}_c[t]$
  // 阶段 3：生成分割掩码
  $\text{trk}_t \leftarrow \text{MLLM}(q, \text{Proj}(\mathbf{F}_{\text{summary}}), \text{Proj}(\mathbf{F}_{\text{ST}}[t]))$
  $m_t \leftarrow \text{Mask-Predictor}(\text{trk}_t, \mathbf{I}_t)$
  输出 m_t
end while

在线流式处理。 帧依次到达的实时场景要求在无法访问未来帧的情况下进行增量处理。不同于同时处理全部帧的离线模式，在线模式在 SSA 中为 N 个空间位置分别维护隐藏状态 $\{\mathbf{h}_{n,t}\}_{n=1}^{N}$，随着帧的到达累积时间上下文。当帧 I_t 到达时，首先通过逐帧独立的双向扫描进行空间聚合。空间聚合后的特征随后进入因果时间扫描，利用已有隐藏状态生成具有充分上下文的表示，同时通过 Eq. (A.2) 中的递推表达式，用当前信息更新这些状态；上述过程在所有空间位置 n 上并行执行。这避免了重新计算过去的表示，从而实现单帧计算开销恒定的真正增量处理。

HSC 模块中的自适应阈值在在线推理期间动态演化。网络不再像离线模式那样对完整视频计算统计量，而是在已观测帧 $[1, t]$ 上维护均值 μ_t 和标准差 σ_t 的运行估计。这些运行统计量输入阈值预测 MLP，在每个时间步生成自适应阈值。压缩后，压缩特征与当前帧的全分辨率特征拼接，并输入 MLLM 生成掩码预测。完整在线流程见 Algorithm 2。

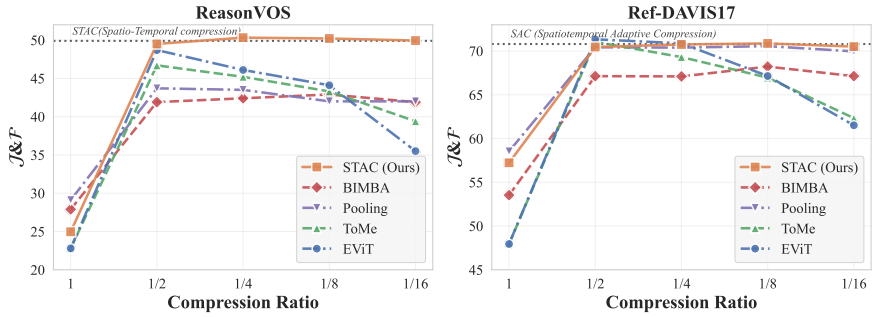


Fig. A.1: 空间压缩策略比较表明，STAC 通过保留 SSA 识别的运动关键 token，在不同压缩比例下保持稳健性能。

Table A.2: 局部查询帧数量对 ReasonVOS 上性能的影响。

局部查询帧数	GLUS			STAC（本文）		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	48.1	45.4	51.0	**49.6**	**47.2**	**52.0**
2	45.6	42.7	48.5	48.9	46.5	51.3
3	44.3	41.6	47.0	49.1	46.8	51.4
4	47.2	44.4	50.0	48.5	46.3	50.7

A.2 补充消融实验

A.2.1 空间压缩策略比较

我们将基于一致性的压缩与其他 token 选择策略进行比较，评估其能否有效保留 SSA 扫描所学习的时空表示。STAC、BIMBA [32] 和 Pooling 使用相同骨干网络，分别结合各自的压缩模块进行端到端训练；ToMe [8] 和 EViT [45] 则应用于训练完成的 GLUS [47] 模型。随后，在推理时改变压缩比例 $r \in \{1, 1/2, 1/4, 1/8, 1/16\}$，并在 ReasonVOS [6] 和 Ref-DAVIS17 [38] 上评估。如图 A.1 所示，STAC 在所有比例下均保持十分稳定的性能，即使在极端的 16× 缩减下，性能下降也很小。然而，其他方法在强压缩下出现明显的性能崩溃，其中基于注意力的方法（ToMe、EViT）下降最为严重。这一差异源于基于注意力的方法优先考虑帧内显著性，却丢弃了其从未显式建模的时间上下文。相比之下，我们的 SSA 扫描机制将运动关键信息编码到时空增强表示中，使其对大幅 token 缩减具有内在鲁棒性。基于一致性的选择利用跨帧一致性显式识别动态区域，保留扫描策略学到的关键细节，进一步补充了这种鲁棒性。

A.2.2 局部查询帧的影响

我们研究：提供多个连续的全分辨率局部查询帧，能否在压缩全局上下文之外提供额外局部细节，从而改善分割。为公平比较，我们在相同设置下训练 STAC 和基线 GLUS [47]，并固定上下文长度。推理时，将查询帧数从 $Q = 1$ 改为 $Q \in \{2, 3, 4\}$，在 ReasonVOS [6] 数据集上评估。如表 A.2 所示，STAC 在所有

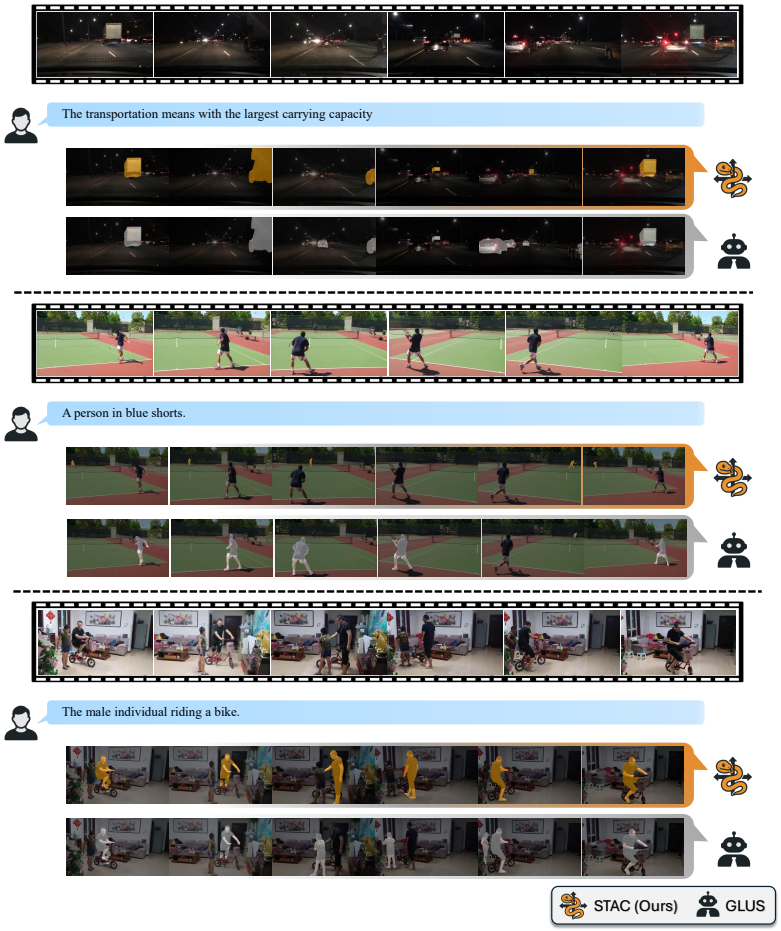


Fig. A.2: 时间一致性与目标身份保持。尽管存在局部遮挡和视角变化，STAC 仍能保持目标的跨帧稳定分割。因果时间扫描通过选择性状态更新传播目标身份，产生平滑且无闪烁的掩码过渡。

配置下均保持十分稳定的性能，变化很小（$\mathcal{J}\&\mathcal{F}$ 分数变化在 1.1 分以内），且始终优于基线。这种稳定性表明，SSA 扫描策略成功地将运动关键信息编码到时空增强表示中，使额外的局部帧成为冗余；单个查询帧即足以实现准确分割，有利于提高部署效率。

A.3 扩展定性分析

A.3.1 成功案例

以下示例展示了所提出的 STAC 架构在具有挑战性的视频推理场景中，通过双阶段扫描策略保持时间一致性与空间精度的能力。

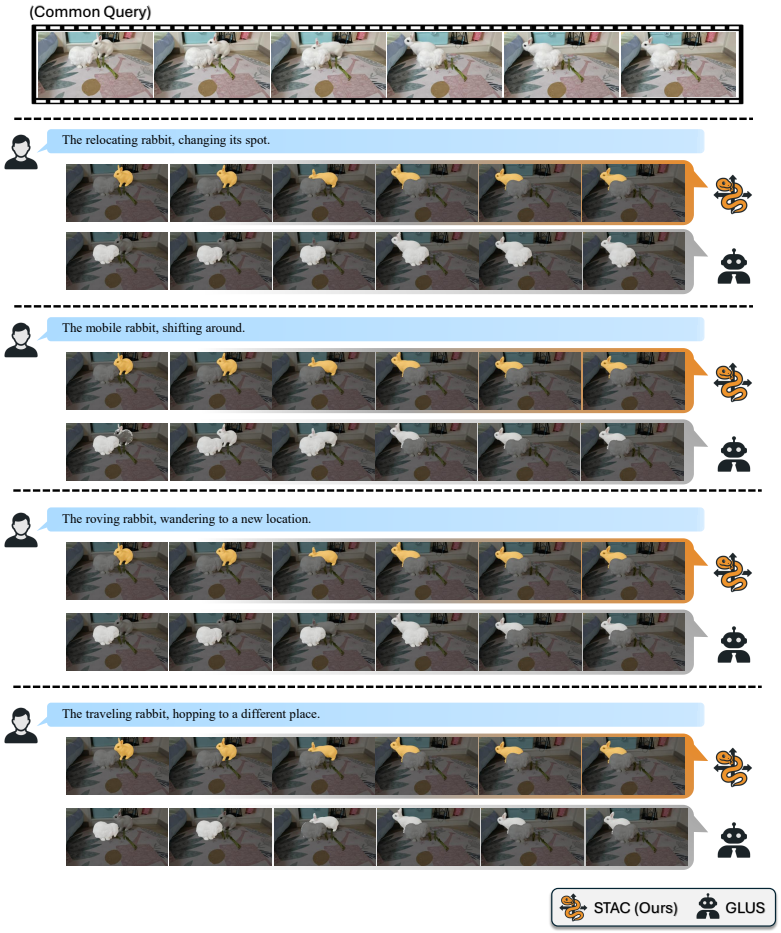


Fig. A.3: 语言不变性与跨查询一致性。给定多个语义等价但用词不同、指向同一目标的查询，STAC 生成相同的分割掩码。这表明双阶段扫描学到了以目标为中心、对语言变化保持不变的表示。

时间一致性与目标身份保持。 给定无歧义的指代表达，STAC 能够保持跨帧稳定分割。如 Fig. A.2 所示，即使存在局部遮挡和视角变化，模型仍能一致地跟踪目标，帧间过渡平滑，掩码闪烁很小。这种稳定性源于因果时间扫描：它通过选择性隐藏状态更新传播目标身份，同时过滤时间噪声。分层压缩即使在大幅缩减下，也能保留这些时间上连贯的表示。

语言不变性与跨查询一致性。 面对指向同一目标的多个不同但无歧义的查询 (Fig. A.3)，STAC 在所有语言变体下均一致地分割同一目标。这表明，双阶段扫描学习到的时空增强表示能够捕获以目标为中心、对措辞差异保持不变的特征。双向空间扫描编码具有判别性的外观特征，因果时间扫描则保持目标身份，从而实现可靠的跨查询识别。

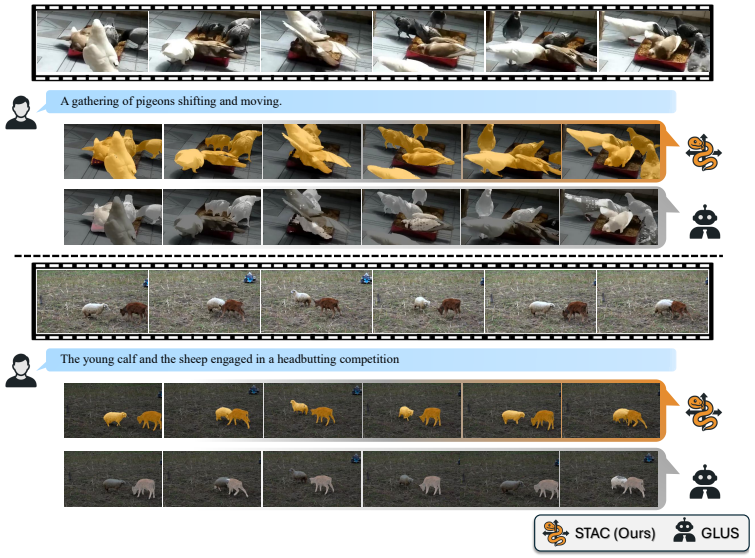


Fig. A.4: 复杂多目标推理。 STAC 在一次前向传播中正确识别并分割多个被指代目标，同时在整个序列中为每个目标保持清晰区分的实例边界和时间一致性。

复杂多目标推理。 STAC 能够成功处理同时指向多个目标的查询，通过一次前向传播正确识别并分割场景中所有被指代目标，同时保持清晰区分的边界。如 Fig. A.4 所示，模型为每个目标分别保持时间一致性，展示了对复杂多目标时空关系进行推理的能力。这种能力源于扫描策略能够学习多个动态实体之间的丰富交互，而分层压缩则保留目标发生交互的运动关键区域。

A.3.2 困难场景

尽管 STAC 在多个基准上表现出色，但包含多个视觉上相似、同时进行复杂运动的目标的场景仍具有挑战性。如 Fig. A.5 所示，这类场景要求在长序列中联合推理细粒度外观差异、空间关系和各目标的运动模式。我们的压缩框架通过合并空间上连贯的区域优先提高效率，因此有时可能模糊相邻相似目标之间的细微边界。值得注意的是，不进行压缩的基线方法也面临相近的困难，这提示上述现象反映了当前视频推理架构的根本局限，而非压缩特有的伪影。这些案例为未来研究细粒度判别机制提供了方向。

A.3.3 定性结果视频

本投稿附有补充视频，可视化展示序列中的分割结果与压缩动态。为说明压缩行为，我们使用红色渐变叠加层：颜色强度增大表示图块正在依次聚合，颜色刷新则标记累积信息的摘要更新。



Fig. A.5: 困难场景。多个视觉上相似的目标同时进行复杂运动的场景，对视频推理系统构成挑战。在压缩过程中合并空间上连贯的区域时，细微的边界差异可能变得模糊。

A.4 未来方向

本研究主要关注为视频推理分割设计高效的时空扫描策略，提出了一种新的双阶段机制，即使仅在指代表达数据上训练，也能成功学习可泛化的表示。在此基础上，未来可探索针对多个视觉相似目标场景的细粒度外观判别专用机制，例如在保持计算效率的同时，利用基于注意力的特征细化，处理扫描策略学到的丰富时空表示。此外，在多个时间尺度上运行的分层记忆结构，可以进一步增强模型在长序列中保持细粒度区分的能力。

双阶段扫描原则的成功提示，它有望扩展至更广泛的视频理解领域，包括视听推理和具身人工智能；在这些领域中，从架构上分离空间与时间处理，为高效时空推理提供了一般原则。我们希望将 STAC 的扫描架构与更长的上下文窗口及视频原生预训练相结合，进一步释放其在这些不同应用领域中的潜力。