Novelty Assessment Report

Paper: ST-SimDiff: Balancing Spatiotemporal Similarity and Difference for Efficient Video Understanding with MLLMs
PDF URL: https://openreview.net/pdf?id=he8kYNcoMA
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

Multimodal Large Language Models (MLLMs) face significant computational overhead when processing long videos due to the massive number of visual tokens required. To improve efficiency, existing methods primarily reduce redundancy by pruning or merging tokens based on importance or similarity. However, these approaches largely overlook a critical dimension of video content, i.e., changes and turning points, and they lack a collaborative model for spatio-temporal relationships. To address this, we propose a new perspective: similarity is for identifying redundancy, while difference is for capturing key events. Based on this, we designed a training-free framework named ST-SimDiff. We first construct a spatio-temporal graph from the visual tokens to uniformly model their complex associations. Subsequently, we employ a parallel dual-selection strategy: 1) similarity-based selection uses community detection to retain representative tokens, compressing static information; 2) temporal difference-based selection precisely locates content-changing points to preserve tokens that capture key dynamic shifts. This allows it to preserve both static and dynamic content with a minimal number of tokens. Extensive experiments show our method significantly outperforms state-of-the-art approaches while substantially reducing computational costs. Our code is available in View paper.

Core Task Landscape

This paper addresses: **Efficient Video Token Compression for Multimodal Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Token Compression Mechanisms and Architectures**
- **Token Selection and Importance Criteria**
- **Temporal Modeling and Video-Specific Strategies**
- **Training Paradigms and Optimization Strategies**
- **Training-Free and Plug-and-Play Methods**
- **Application-Specific and Domain-Adapted Compression**

Complete Taxonomy Tree

- Efficient Video Token Compression for Multimodal Large Language Models Survey Taxonomy
- Token Compression Mechanisms and Architectures
  - Layer-wise and Progressive Compression (3 papers)
  - [15] Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model (Liu Ting, 2024) View paper
  - [18] PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models (Chen-Yu Yang, 2024) View paper
  - [19] LaCo: Efficient Layer-wise Compression of Visual Tokens for Multimodal Large Language Models (Liu Jun-tao, 2025) View paper
  - Post-Encoder Compression Modules (3 papers)
  - [7] DeCo: Decoupling Token Compression from Semantic Abstraction in Multimodal Large Language Models (Yao, 2024) View paper
  - [28] LEO-MINI: An Efficient Multimodal Large Language Model using Conditional Token Reduction and Mixture of Multi-Modal Experts (Yimu Wang, 2025) View paper
  - [39] DiViCo: Disentangled Visual Token Compression for Efficient Large Vision-Language Model (Xin Wang, 2025) View paper
  - Inner-LLM Token Reduction (3 papers)
  - [12] HoliTom: Holistic Token Merging for Fast Video Large Language Models (Kele Shao, 2025) View paper
  - [21] DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models (Keda Tao, 2025) View paper
  - [24] Variation-aware Vision Token Dropping for Faster Large Vision-Language Models (Chen Junjie, 2025) View paper
  - Discrete Tokenization and Quantization (3 papers)
  - [34] Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization (Jin Yang, 2024) View paper
  - [35] VQToken: Neural Discrete Token Representation Learning for Extreme Token Reduction in Video Large Language Models (Zhang Haichao, 2025) View paper
  - [38] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation (Yu, 2023) View paper
- Token Selection and Importance Criteria
  - Attention and Explainability-Based Selection (3 papers)
  - [11] Generic Token Compression in Multimodal Large Language Models from an Explainability Perspective (Lei Lei, 2025) View paper
  - [17] FlashVLM: Text-Guided Visual Token Selection for Large Multimodal Models (Kaitong Cai, 2025) View paper
  - [44] Learning Free Token Reduction for Multi-Modal LLM (Zhao Zihui, 2025) View paper

- Cross-Modal and Instruction-Guided Selection (3 papers)
- [2] Hybrid-level instruction injection for video token compression in multi-modal large language models (Liu Zhihang, 2025) View paper
- [4] Tokencarve: Information-preserving visual token compression in multimodal large language models (Tan Xudong, 2025) View paper
- [47] Language-Guided Adaptive Vision Token Pruning for Efficient Multimodal Large Language Models (Omer Faruk Deniz, 2025) View paper
- Similarity and Redundancy-Based Merging (3 papers)
- [26] Framefusion: Combining similarity and importance for video token reduction on large vision language models (Fu Tian-Yu, 2025) View paper
- [30] Recoverable compression: A multimodal vision token recovery mechanism guided by text information (Yi Chen, 2025) View paper
- [32] Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs (Hyun, 2025) View paper
- Spatiotemporal Redundancy Analysis ★ (3 papers)
- [0] ST-SimDiff: Balancing Spatiotemporal Similarity and Difference for Efficient Video Understanding with MLLMs (Anon et al., 2026) View paper
- [13] LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding (Shen, 2024) View paper
- [31] Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding (Wang Xiao, 2025) View paper

• Temporal Modeling and Video-Specific Strategies
- Explicit Temporal Encoding and Compression (3 papers)
- [10] Exploring the Role of Explicit Temporal Modeling in Multimodal Large Language Models for Video Understanding (Li Yun, 2025) View paper
- [20] VidCompress: Memory-Enhanced Temporal Compression for Video Understanding in Large Language Models (Lan, 2024) View paper
- [33] State Space Model based Temporal Adaptive Token Compression for Efficient Multimodal Large Language Model (Semi Kwon, 2025) View paper
- Adaptive Frame Sampling and Dynamic Compression (3 papers)
- [46] Dynamic-vlm: Simple dynamic visual token compression for videollm (Wang Han, 2025) View paper
- [48] DynTok: Dynamic Compression of Visual Tokens for Efficient and Effective Video Understanding (Zhang Hong-zhi, 2025) View paper
- [50] Timechat-online: 80% visual tokens are naturally redundant in streaming videos (Yao, 2025) View paper
- Slow-Fast and Multi-Granularity Architectures (3 papers)
- [5] Mavors: Multi-granularity video representation for multimodal large language model (Shi Yang, 2025) View paper
- [9] Slow-fast architecture for video multi-modal large language models (Shi Min, 2025) View paper
- [36] TS-LLaVA: Constructing Visual Tokens through Thumbnail-and-Sampling for Training-Free Video Large Language Models (Qu, 2024) View paper
- Trajectory and Object-Centric Tokenization (2 papers)
- [27] LLaVA-Scissor: Token Compression with Semantic Connected Components for Video LLMs (Sun BoYuan, 2025) View paper
- [43] One Trajectory, One Token: Grounded Video Tokenization via Panoptic Sub-object Trajectory (Zheng Chenhao, 2025) View paper

• Training Paradigms and Optimization Strategies
- End-to-End Trained Compression (3 papers)
- [1] Llava-prumerge: Adaptive token reduction for efficient large multimodal models (Shang, 2025) View paper
- [8] Less is more: Vision representation compression for efficient video generation with large language models (Y Zhou, 2024) View paper
- [37] Efficient large multi-modal models via visual context compression (Jie-neng Chen, 2024) View paper
- Reinforcement Learning and Memory-Augmented Training (2 papers)
- [25] MARC: Memory-Augmented RL Token Compression for Efficient Video Understanding (Wu Peiran, 2025) View paper
- [41] video-SALMONN S: Streaming Audio-Visual LLMs Beyond Length Limits via Memory (Sun Guang-zhi, 2025) View paper
- Vision-Centric and Image-to-Video Transfer (3 papers)
- [14] Videollama 3: Frontier multimodal foundation models for image and video understanding (Zhang, 2025) View paper
- [29] Vision-centric Token Compression in Large Language Model (Xing Ling, 2025) View paper
- [40] Fewer tokens and fewer videos: Extending video understanding abilities in large vision-language models (Chen Shi-min, 2024) View paper

• Training-Free and Plug-and-Play Methods
- Inference-Time Token Pruning and Merging (3 papers)
- [6] Voco-llama: Towards vision compression with large language models (Xubing Ye, 2025) View paper
- [16] Llava-mini: Efficient image and video large multimodal models with one vision token (Zhang, 2025) View paper
- [23] Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models (Xuyang Liu, 2025) View paper
- Cross-Modal Audio-Visual Compression (1 papers)
- [22] OmniZip: Audio-Guided Dynamic Token Compression for Fast Omnimodal Large Language Models (Keda Tao, 2025) View paper

• Application-Specific and Domain-Adapted Compression
- Long-Context and Streaming Video Processing (1 papers)
- [49] Improving LLM Video Understanding with 16 Frames Per Second (Li, 2025) View paper
- High-Resolution and Multi-View Compression (1 papers)
- [42] Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution (Liu, 2024) View paper
- Video Coding and Generation Applications (2 papers)
- [3] When video coding meets multimodal large language models: A unified paradigm for video coding (Zhang Ping-ping, 2024) View paper
- [45] A survey on generative ai and llm for video generation, understanding, and streaming (Pengyuan Zhou, 2024) View paper

Narrative

Core task: efficient video token compression for multimodal large language models. As video understanding becomes central to multimodal LLMs, the computational burden of processing dense visual tokens has driven a rich landscape of compression strategies. The field organizes around several complementary branches: Token Compression Mechanisms and Architectures develop novel structural designs (e.g., merging layers, quantization schemes) to reduce token counts; Token Selection and Importance Criteria focus on identifying which tokens matter most through attention scores, similarity metrics, or spatiotemporal redundancy analysis; Temporal Modeling and Video-Specific Strategies exploit the unique structure of video data, such as frame-level dependencies and motion patterns; Training Paradigms and Optimization Strategies address how to learn compression policies end-to-end or via auxiliary objectives; Training-Free and Plug-and-Play Methods offer lightweight alternatives that adapt existing models without retraining; and Application-Specific and Domain-Adapted Compression tailors techniques to particular downstream tasks or video domains. Representative works like LongVU[13] and Videollama 3[14] illustrate how temporal redundancy can be leveraged, while methods such as Llava-prumerge[1] and Tokencarve[4] demonstrate diverse architectural choices for token reduction.

A particularly active line of inquiry centers on spatiotemporal redundancy analysis, where methods quantify and exploit the overlap between frames and spatial regions. ST-SimDiff[0] sits squarely in this cluster, emphasizing similarity-based filtering to discard redundant tokens across both space and time. Nearby works like LongVU[13] also target long-form video efficiency through temporal pooling, while Adaretake[31] adapts token selection dynamically based on content variation. These approaches contrast with training-free methods such as DeCo[7] and Less is more[8], which apply heuristic pruning without model updates, and with architecture-driven solutions like Mavors[5] or Voco-llama[6], which integrate compression directly into the model backbone. The central trade-off across these branches involves balancing compression ratio against semantic fidelity: aggressive pruning risks losing fine-grained details, while conservative strategies may not sufficiently reduce computational cost. ST-SimDiff[0] navigates this by focusing on spatiotemporal similarity metrics, positioning itself as a middle ground that preserves task-relevant information while achieving substantial token reduction, complementing both the temporal-centric designs of LongVU[13] and the adaptive selection logic of Adaretake[31].

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding

Authors: Shen, Xiaoqian, Xiong Yunyang, Xiaoqian Shen, Zhao Changsheng, et al. (42 authors total) | **Year/Venue**: 2024 • International Conference on Machine Learning | **URL**: View paper

Abstract

Multimodal Large Language Models (MLLMs) have shown promising progress in understanding and analyzing video content. However, processing long videos remains a significant challenge constrained by LLM's context size. To address this limitation, we propose LongVU, a spatiotemporal adaptive compression mechanism thats reduces the number of video tokens while preserving visual details of long videos. Our idea is based on leveraging cross-modal query and inter-frame dependencies to adaptively reduce ...

Relationship Analysis

Both papers belong to the Spatiotemporal Redundancy Analysis category, focusing on exploiting spatial and temporal redundancy patterns in video tokens for efficient compression. They overlap in their core approach of analyzing both spatial (within-frame) and temporal (across-frame) redundancy to reduce token counts while preserving video understanding capabilities. The key difference is that ST-SimDiff constructs a spatio-temporal graph and uses a dual-selection strategy (similarity-based community detection for redundancy and difference-based detection for key events), while LongVU employs a sequential adaptive compression pipeline (DINOv2-based frame removal, text-guided cross-modal queries, and temporal dependency-based spatial reduction).

---

2. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding

Authors: Wang Xiao, Si, Qingyi, Xiao Wang, Wu, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Multimodal Large Language Models (MLLMs) have revolutionized video understanding, yet are still limited by context length when processing long videos. Recent methods compress videos by leveraging visual redundancy uniformly, yielding promising results. Nevertheless, our quantitative analysis shows that redundancy varies significantly across time and model layers, necessitating a more flexible compression strategy. We propose AdaReTaKe, a training-free method that flexibly reduces visual redundan...

Relationship Analysis

Both papers belong to the Spatiotemporal Redundancy Analysis category, focusing on exploiting spatial and temporal redundancy patterns in video tokens for efficient MLLM processing. They overlap in their core approach of analyzing token similarity across frames and spatial positions to identify redundant information for compression. However, ST-SimDiff uniquely combines similarity-based community detection with temporal difference-based event detection using a spatio-temporal graph framework, while AdaRETAKE focuses on adaptive compression ratio allocation across timestamps and LLM layers based on heavy-hitter analysis and attention scores, without explicit event detection mechanisms.

Contributions Analysis

Overall novelty summary. The paper proposes ST-SimDiff, a training-free framework for video token compression in multimodal LLMs that combines similarity-based redundancy removal with temporal difference-based event detection. It resides in the 'Spatiotemporal Redundancy Analysis' leaf of the taxonomy, which contains only three papers total. This leaf sits within the broader 'Token Selection and Importance Criteria' branch, indicating the work focuses on defining selection criteria rather than architectural redesign. The small leaf size suggests this specific combination of spatial and temporal redundancy modeling remains relatively underexplored compared to other compression strategies.

The taxonomy reveals several neighboring research directions. The sibling leaf 'Similarity and Redundancy-Based Merging' addresses token clustering without explicit temporal modeling, while 'Attention and Explainability-Based Selection' uses LLM-derived importance scores rather than content-based redundancy. Adjacent branches include 'Temporal Modeling and Video-Specific Strategies', which encompasses explicit temporal encoders and adaptive frame sampling, and 'Training-Free and Plug-and-Play Methods', which shares the inference-time approach but may not emphasize spatiotemporal structure. ST-SimDiff bridges these areas by applying training-free selection criteria specifically to spatiotemporal redundancy patterns.

Among twenty candidates examined across three contributions, the dual perspective on similarity and difference shows one refutable candidate from ten examined, suggesting some prior work addresses similar conceptual framing. The ST-SimDiff framework itself found no refutations among ten candidates, indicating the specific combination of spatio-temporal graph modeling with dual-selection may be less directly anticipated. The parallel dual-selection strategy was not separately evaluated. These statistics reflect a limited semantic search scope rather than exhaustive coverage, meaning additional related work may exist beyond the top-twenty matches examined.

Given the constrained search scope and the sparse population of the taxonomy leaf, the work appears to occupy a relatively distinct position within spatiotemporal redundancy analysis. The explicit focus on difference-based event detection alongside similarity-based

compression differentiates it from purely redundancy-focused methods. However, the limited candidate pool and single refutation suggest careful positioning relative to existing temporal modeling and training-free compression literature would strengthen claims of novelty.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Dual perspective on similarity and difference for video token compression

Description: The authors introduce a conceptual framework that treats similarity as a mechanism for compressing redundant static content in videos, while treating difference as essential for capturing key dynamic events and turning points that drive video narratives.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Don't Look Twice: Faster Video Transformers with Run-Length Tokenization

URL: View paper

Brief Assessment

Run-Length Tokenization[54] focuses on identifying and removing temporally redundant patches through run-length encoding, not on a dual framework that explicitly balances similarity-based compression with difference-based event detection as conceptualized in the original paper.

---

2. Timechat-online: 80% visual tokens are naturally redundant in streaming videos

URL: View paper

Brief Assessment

Timechat-online[50] focuses on temporal redundancy in streaming videos by preserving significant temporal changes between consecutive frames, rather than proposing a conceptual framework that treats similarity and difference as dual mechanisms for video token compression.

---

3. Framefusion: Combining similarity and importance for video token reduction on large vision language models

URL: View paper

Prior Art Analysis

Framefusion[26] demonstrates that the concept of using similarity for video token compression was established prior to the original paper's submission. The candidate paper explicitly focuses on similarity-based token merging as a mechanism for identifying and compressing redundant visual content in videos, particularly between adjacent frames. While the original paper claims to be the first to propose treating similarity and difference with equal importance, Framefusion[26] already implements similarity-based compression strategies for redundant static content, which directly overlaps with the original paper's claimed novelty of using similarity to identify redundancy.

Evidence

Evidence 1 - **Rationale**: Both papers establish similarity as a mechanism for identifying and compressing redundancy in video content. Framefusion[26] explicitly uses similarity to reduce redundancy, which is the core claim of the original paper's similarity component. - **Original**: we propose a new perspective: similarity is for identifying redundancy, while difference is for capturing key events - **Candidate**: we argue that even important tokens can introduce redundancy due to visual similarities, particularly among adjacent frames. By merging these highly similar tokens, redundancy can be significantly reduced without compromising essential visual information

Evidence 2 - **Rationale**: Both papers use similarity to compress static/redundant information. Framefusion[26] implements similarity-based token merging to handle redundant content, demonstrating prior work on this concept. - **Original**: similarity-based selection uses community detection to retain representative tokens, compressing static information - **Candidate**: framefusion computes token similarities exclusively between corresponding visual tokens from adjacent frames, applies token merging at initial successive layers followed by pruning in deeper layers

Evidence 3 - **Rationale**: While Framefusion[26] combines similarity with importance rather than difference, it establishes that similarity-based approaches for token reduction existed before the original paper, challenging the claim of being first to emphasize similarity in VLM efficiency. - **Original**: we are the first to propose that the similarity and difference of tokens should be given equal importance in vlm efficiency research - **Candidate**: we propose framefusion, a novel token reduction approach integrating similarity-based merging with importance-based pruning

Evidence 4 - **Rationale**: Framefusion[26] explicitly analyzes and utilizes spatiotemporal similarity for compression, including both spatial and temporal dimensions, which overlaps with the original paper's claimed contribution. - **Original**: spatiotemporal similarity (left) can be utilized to compress redundant information both spatially within a frame and temporally across adjacent frames - **Candidate**: token similarity predominantly occurs between spatially corresponding tokens from adjacent frames, (2) similarities exhibit high values particularly at shallow layers, and (3) the token similarity rankings are highly consistent across layers

---

4. When Tokens Talk Too Much: A Survey of Multimodal Long-Context Token Compression across Images, Videos, and Audios

URL: View paper

Brief Assessment

Tokens Talk Survey[52] is a survey paper that categorizes existing compression methods but does not propose the dual similarity-difference framework for video token compression that the original paper introduces.

---

5. Less is more: Vision representation compression for efficient video generation with large language models

URL: View paper

Brief Assessment

Less is more[8] focuses on compressing video token representations for LLM-based video generation by eliminating redundancy between adjacent tokens through learnable compressor/decompressor modules. The original paper addresses video understanding with MLLMs using graph-based similarity and difference detection for token selection, which is a fundamentally different task and technical approach.

---

6. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models

URL: View paper

Brief Assessment

DyCoke[21] focuses on temporal merging and dynamic pruning based on attention scores during decoding, not on a conceptual framework treating similarity and difference as dual mechanisms for compression. The candidate's approach is implementation-focused rather than conceptual.

---

7. Motion Guided Token Compression for Efficient Masked Video Modeling

URL: View paper

Brief Assessment

Motion Guided Compression[55] focuses on temporal motion differences between consecutive frames for masked video modeling in action recognition tasks, not on a dual framework balancing similarity-based redundancy compression with difference-based event detection for multimodal LLMs processing long videos.

---

8. The Best and Most Efficient Video Compression Methods

URL: View paper

Brief Assessment

Best Efficient Compression[53] focuses on traditional video codec compression methods (H.264/AVC, H.265/HEVC, H.266/VVC, DCT, Huffman coding) for storage and transmission, not on visual token compression for multimodal language models or the conceptual framework of similarity/difference for redundancy identification in video understanding tasks.

---

9. Fast: Efficient action tokenization for vision-language-action models

URL: View paper

Brief Assessment

Fast action tokenization[51] focuses on compressing robot action sequences using DCT and BPE for vision-language-action models, not video token compression. The candidate addresses action tokenization for robotic control, while the original paper addresses video understanding with MLLMs through spatiotemporal similarity and temporal difference analysis.

---

10. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding

URL: View paper

Brief Assessment

LongVU[13] focuses on removing redundant frames through similarity-based compression but does not explicitly introduce a dual conceptual framework that treats difference as essential for capturing key dynamic events and turning points in video narratives.

Contribution 2: ST-SimDiff framework with spatio-temporal graph modeling

Description: The authors develop a training-free framework that constructs a spatio-temporal graph to uniformly model complex spatial and temporal relationships between video tokens, enabling joint analysis of spatio-temporal correlations that existing methods fail to capture.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Dynamic spatio-temporal graph reasoning for videoqa with self-supervised event recognition

URL: View paper

Brief Assessment

Dynamic graph reasoning[56] focuses on VideoQA with question-guided dynamic graphs for object interactions and event recognition, while the original paper addresses video token compression for MLLMs using static spatio-temporal graphs with similarity/difference-based selection strategies. These are fundamentally different applications and technical approaches.

---

2. Efficient video transformers via spatial-temporal token merging for action recognition

URL: View paper

Brief Assessment

Spatial-temporal token merging[62] focuses on action recognition tasks using token merging for efficiency in video transformers, while the original paper addresses video understanding in multimodal LLMs using graph-based community detection and dual selection strategies for both similarity and difference.

---

3. Multi-stage spatio-temporal aggregation transformer for video person re-identification

URL: View paper

Brief Assessment

Multi-stage aggregation transformer[65] focuses on video person re-identification using multi-stage spatial-temporal aggregation for attribute and identity features, not on video token compression or similarity-difference balancing for efficient video understanding with MLLMs.

---

4. Spatial-temporal graphs for cross-modal text2video retrieval

URL: View paper

Brief Assessment

Spatial-temporal graphs retrieval[63] focuses on cross-modal text-to-video retrieval using spatial-temporal graphs to model object interactions, while the original paper addresses video token compression for efficient MLLM processing. These are fundamentally different tasks with different objectives.

---

5. Constructing holistic spatio-temporal scene graph for video semantic role labeling

URL: View paper

Brief Assessment

Holistic scene graph[58] focuses on video semantic role labeling with scene-event mapping for event structure detection, not on visual token compression for efficient video understanding in MLLMs.

---

6. Video relation detection with spatio-temporal graph

URL: View paper

Brief Assessment

Video relation detection[64] focuses on detecting object relationships in videos using spatial-temporal graphs for relation triplet prediction, not on video token compression for MLLMs. The technical goals and applications are fundamentally different.

---

7. Enhancing video-language representations with structural spatio-temporal alignment

URL: View paper

Brief Assessment

Structural alignment[57] focuses on video-language representation learning with scene graph structures for cross-modal alignment, not on token compression for efficiency in MLLMs. The candidate addresses a fundamentally different problem domain (representation learning vs. computational efficiency).

---

8. Exploring spatioâ□□temporal graph convolution for video-based humanâ□□object interaction recognition

URL: View paper

Brief Assessment

Spatio-temporal graph convolution[61] focuses on human-object interaction recognition in videos using graph convolutions to model object relationships, not on video token compression for multimodal large language models.

---

9. Cross-attentional spatio-temporal semantic graph networks for video question answering

URL: View paper

Brief Assessment

Cross-attentional graph[60] focuses on video question answering with heterogeneous graphs for multimodal reasoning, not on efficient video token compression for MLLMs. The technical contexts and objectives are fundamentally different.

---

10. ODTrack: Online Dense Temporal Token Learning for Visual Tracking

URL: View paper

Brief Assessment

ODTrack[59] focuses on visual object tracking with online token propagation for frame-to-frame association in tracking tasks, not on video token compression for multimodal LLMs through spatio-temporal graph-based community detection and dual selection strategies.

Contribution 3: Parallel dual-selection strategy for token compression

Description: The authors propose a novel dual token selection strategy that operates in parallel: similarity-based selection applies community detection to compress redundant static content, while difference-based selection identifies temporal turning points to preserve tokens capturing key dynamic shifts.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 21 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Framefusion: Combining similarity and importance for video token reduction on large vision language models

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] ST-SimDiff: Balancing Spatiotemporal Similarity and Difference for Efficient Video Understanding with MLLMs View paper
- [1] Llava-prumerge: Adaptive token reduction for efficient large multimodal models View paper
- [2] Hybrid-level instruction injection for video token compression in multi-modal large language models View paper
- [3] When video coding meets multimodal large language models: A unified paradigm for video coding View paper
- [4] Tokencarve: Information-preserving visual token compression in multimodal large language models View paper
- [5] Mavors: Multi-granularity video representation for multimodal large language model View paper
- [6] Voco-llama: Towards vision compression with large language models View paper
- [7] DeCo: Decoupling Token Compression from Semantic Abstraction in Multimodal Large Language Models View paper
- [8] Less is more: Vision representation compression for efficient video generation with large language models View paper
- [9] Slow-fast architecture for video multi-modal large language models View paper
- [10] Exploring the Role of Explicit Temporal Modeling in Multimodal Large Language Models for Video Understanding View paper
- [11] Generic Token Compression in Multimodal Large Language Models from an Explainability Perspective View paper
- [12] HoliTom: Holistic Token Merging for Fast Video Large Language Models View paper
- [13] LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding View paper
- [14] Videollama 3: Frontier multimodal foundation models for image and video understanding View paper
- [15] Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model View paper
- [16] Llava-mini: Efficient image and video large multimodal models with one vision token View paper
- [17] FlashVLM: Text-Guided Visual Token Selection for Large Multimodal Models View paper
- [18] PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models View paper
- [19] LaCo: Efficient Layer-wise Compression of Visual Tokens for Multimodal Large Language Models View paper
- [20] VidCompress: Memory-Enhanced Temporal Compression for Video Understanding in Large Language Models View paper
- [21] DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models View paper
- [22] OmniZip: Audio-Guided Dynamic Token Compression for Fast Omnimodal Large Language Models View paper
- [23] Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models View paper
- [24] Variation-aware Vision Token Dropping for Faster Large Vision-Language Models View paper
- [25] MARC: Memory-Augmented RL Token Compression for Efficient Video Understanding View paper
- [26] Framefusion: Combining similarity and importance for video token reduction on large vision language models View paper
- [27] LLaVA-Scissor: Token Compression with Semantic Connected Components for Video LLMs View paper
- [28] LEO-MINI: An Efficient Multimodal Large Language Model using Conditional Token Reduction and Mixture of Multi-Modal Experts View paper
- [29] Vision-centric Token Compression in Large Language Model View paper
- [30] Recoverable compression: A multimodal vision token recovery mechanism guided by text information View paper

- [31] Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding View paper
- [32] Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs View paper
- [33] State Space Model based Temporal Adaptive Token Compression for Efficient Multimodal Large Language Model View paper
- [34] Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization View paper
- [35] VQToken: Neural Discrete Token Representation Learning for Extreme Token Reduction in Video Large Language Models View paper
- [36] TS-LLaVA: Constructing Visual Tokens through Thumbnail-and-Sampling for Training-Free Video Large Language Models View paper
- [37] Efficient large multi-modal models via visual context compression View paper
- [38] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation View paper
- [39] DiViCo: Disentangled Visual Token Compression for Efficient Large Vision-Language Model View paper
- [40] Fewer tokens and fewer videos: Extending video understanding abilities in large vision-language models View paper
- [41] video-SALMONN S: Streaming Audio-Visual LLMs Beyond Length Limits via Memory View paper
- [42] Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution View paper
- [43] One Trajectory, One Token: Grounded Video Tokenization via Panoptic Sub-object Trajectory View paper
- [44] Learning Free Token Reduction for Multi-Modal LLM View paper
- [45] A survey on generative ai and llm for video generation, understanding, and streaming View paper
- [46] Dynamic-vlm: Simple dynamic visual token compression for videollm View paper
- [47] Language-Guided Adaptive Vision Token Pruning for Efficient Multimodal Large Language Models View paper
- [48] DynTok: Dynamic Compression of Visual Tokens for Efficient and Effective Video Understanding View paper
- [49] Improving LLM Video Understanding with 16 Frames Per Second View paper
- [50] Timechat-online: 80% visual tokens are naturally redundant in streaming videos View paper
- [51] Fast: Efficient action tokenization for vision-language-action models View paper
- [52] When Tokens Talk Too Much: A Survey of Multimodal Long-Context Token Compression across Images, Videos, and Audios View paper
- [53] The Best and Most Efficient Video Compression Methods View paper
- [54] Don't Look Twice: Faster Video Transformers with Run-Length Tokenization View paper
- [55] Motion Guided Token Compression for Efficient Masked Video Modeling View paper
- [56] Dynamic spatio-temporal graph reasoning for videoqa with self-supervised event recognition View paper
- [57] Enhancing video-language representations with structural spatio-temporal alignment View paper
- [58] Constructing holistic spatio-temporal scene graph for video semantic role labeling View paper
- [59] ODTrack: Online Dense Temporal Token Learning for Visual Tracking View paper
- [60] Cross-attentional spatio-temporal semantic graph networks for video question answering View paper
- [61] Exploring spatioâ□□temporal graph convolution for video-based humanâ□□object interaction recognition View paper
- [62] Efficient video transformers via spatial-temporal token merging for action recognition View paper
- [63] Spatial-temporal graphs for cross-modal text2video retrieval View paper
- [64] Video relation detection with spatio-temporal graph View paper
- [65] Multi-stage spatio-temporal aggregation transformer for video person re-identification View paper