Novelty Assessment Report

Paper: AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration

PDF URL: https://openreview.net/pdf?id=vjEl1PuIDE

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Audiovisual video captioning aims to generate semantically rich descriptions with temporal alignment between visual and auditory events, thereby benefiting both video understanding and generation. In this paper, we present **AVoCaDO**, a powerful audiovisual video captioner driven by the temporal orchestration between audio and visual modalities. We propose a two-stage post-training pipeline: (1) **AVoCaDO SFT**, which fine-tunes the model on a newly curated dataset of 107K high-quality, temporally-aligned audiovisual captions; and (2) **AVoCaDO GRPO**, which leverages tailored reward functions to further enhance temporal coherence and dialogue accuracy while regularizing caption length and reducing collapse. Experimental results demonstrate that AVoCaDO significantly outperforms existing open-source models across four audiovisual video captioning benchmarks, and also achieves competitive performance on the VDC benchmark under visual-only settings. The model will be made publicly available to facilitate future research in audiovisual video understanding and generation.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Audiovisual Video Captioning with Temporal Alignment**

A total of **50 papers** were analyzed and organized into a taxonomy with **35 categories**.

Taxonomy Overview

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

- **Audiovisual Representation Learning and Alignment**
- **Audiovisual Caption Generation Architectures**
- **Large Language Model-Based Audiovisual Understanding**
- **Audiovisual Reasoning and Question Answering**
- **Cross-Modal Generation and Synchronization**
- **Retrieval and Matching with Temporal Alignment**
- **Datasets and Evaluation Frameworks**

Complete Taxonomy Tree

- Audiovisual Video Captioning with Temporal Alignment Survey Taxonomy
- Audiovisual Representation Learning and Alignment
  - Continual and Localized Audio-Visual Pre-training (1 papers)
  - [2] STELLA: continual audio-video pre-training with spatio-temporal localized alignment (Lee Jae-Woo, 2023) View paper
  - Unified Multimodal Pre-training with Audio Synchronization (1 papers)
  - [8] Unified video-language pre-training with synchronized audio (Mo, 2024) View paper
  - Structural and Fine-Grained Spatio-Temporal Alignment (1 papers)
  - [7] Enhancing video-language representations with structural spatio-temporal alignment (Hao Fei, 2024) View paper
  - Compositional and Temporal Alignment Benchmarking (1 papers)
  - [4] VideoComp: Advancing Fine-Grained Compositional and Temporal Alignment in Video-Text Models (Dahun Kim, 2025) View paper
- Audiovisual Caption Generation Architectures
  - Temporally Orchestrated Audiovisual Captioning ★ (1 papers)
  - [0] AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration (Anon et al., 2026) View paper
  - Hierarchical and Cross-Modal Attention for Captioning (2 papers)
  - [26] Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning (Xin Wang, 2018) View paper
  - [39] Visually-Aware Audio Captioning With Adaptive Audio-Visual Attention (Liu Xubo, 2023) View paper
  - Dense Video Captioning with Multimodal Fusion (2 papers)
  - [47] Multi-modal Dense Video Captioning (Vladimir E. Iashin, 2022) View paper
  - [48] Exploring Audio-Visual Concepts for Dense Video Captioning (Zhuyang Xie, 2024) View paper
  - Visual-Centric Dense and Temporal Captioning (5 papers)
  - [15] NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative (Nadeem, 2024) View paper
  - [20] Learning semantic concepts and temporal alignment for narrated video procedural captioning (Botian Shi, 2020) View paper
  - [24] Explicit Temporal-Semantic Modeling for Dense Video Captioning via Context-Aware Cross-Modal Interaction (Mingda Jia, 2025) View paper
  - [45] Whats in a Video: Factorized Autoregressive Decoding for Online Dense Video Captioning (Piergiovanni, 2024) View paper
  - [46] Enhanced visual multi-modal fusion framework for dense video captioning (Ruizhe Zhong, 2023) View paper
  - Adaptive and Memory-Augmented Captioning Frameworks (3 papers)

- [14] ATMNet: Adaptive Two-Stage Modular Network for Accurate Video Captioning (Tianyang Xu, 2025) View paper
    - [17] MIRA-CAP: Memory-Integrated Retrieval-Augmented captioning for State-of-the-Art image and video captioning (Sabina Umirzakova, 2024) View paper
    - [34] SkimCap: A Transformer-Based Video Captioning Method with Adaptive Attention and Hierarchical Skimming Features (L. V. Cardoso, 2025) View paper
    - Preference-Aligned and Training-Free Captioning (2 papers)
    - [27] Toward Scalable Video Narration: A Training-free Approach Using Multimodal Large Language Models (Wu, 2025) View paper
    - [35] AVC-DPO: Aligned Video Captioning via Direct Preference Optimization (Tang, 2025) View paper
    - Visual-Syntactic and Temporal Composition Models (2 papers)
    - [28] Understanding temporal structure for video captioning (Shagan Sah, 2020) View paper
    - [29] Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding* (Jesus Perez-Martin, 2021) View paper
    - Multimodal Feature Fusion for Video Description (2 papers)
    - [25] Video description combining visual and audio features (Yongbo Li, 2023) View paper
    - [41] Center-enhanced video captioning model with multimodal semantic alignment. (Benhui Zhang, 2024) View paper
    - Enhanced Architectures with Temporal Coherence (2 papers)
    - [31] Video Captioning with Spatio-Temporal Graph Transformers (Muksimova Shakhnoza, 2025) View paper
    - [33] Enhanced Hybrid Framework and Comparative Analysis of Deep Learning Architectures for Video Captioning (PP Bhusare, 2025) View paper
- Large Language Model-Based Audiovisual Understanding
    - Pseudo-Untrimmed Video Learning for LLMs (1 papers)
    - [3] Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding (Yunlong Tang, 2025) View paper
    - Fine-Grained Audio-Visual LLMs with Alignment (1 papers)
    - [13] Aligned better, listen better for audio-visual large language models (GUO Yuxin, 2025) View paper
    - Pixel-Temporal and Object-Level Alignment for Video LLMs (1 papers)
    - [18] Pite: Pixel-temporal alignment for large video-language model (Liu Yang, 2024) View paper
    - Unified Causal Video-Language Modeling (1 papers)
    - [32] UNICORN: A Unified Causal Video-Oriented Language-Modeling Framework for Temporal Video-Language Tasks (Yuanhao Xiong, 2024) View paper
- Audiovisual Reasoning and Question Answering
    - Audio-Visual QA with Temporal Alignment Benchmarks (1 papers)
    - [1] Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities (Zhou Ziwei, 2025) View paper
    - Target-Aware Spatio-Temporal Reasoning for AVQA (1 papers)
    - [21] Target-Aware Spatio-Temporal Reasoning via Answering Questions in Dynamics Audio-Visual Scenarios (Jiang Yuan-Yuan, 2023) View paper
    - Dialogue and Interactive Audiovisual Understanding (2 papers)
    - [38] LiveChat: Video Comment Generation from Audio-Visual Multimodal Contexts (Julien Lalanne, 2023) View paper
    - [49] DialogMCF: Multimodal Context Flow for Audio Visual Scene-Aware Dialog (Zhe Chen, 2023) View paper
- Cross-Modal Generation and Synchronization
    - Audio-to-Video Generation with Temporal Control (2 papers)
    - [5] Diverse and aligned audio-to-video generation via text-to-video model adaptation (Yariv, 2024) View paper
    - [23] Audio-Sync Video Generation with Multi-Stream Temporal Control (Weng, 2025) View paper
    - Video-to-Audio Generation with Semantic and Rhythmic Alignment (1 papers)
    - [6] Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation (Xinyi Tong, 2025) View paper
    - Bidirectional Audio-Video Generation with Diffusion (1 papers)
    - [9] Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation (Menapace, 2025) View paper
    - Text-to-Audio-Video Generation with Alignment (2 papers)
    - [10] Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation (J Zheng, 2025) View paper
    - [16] Text-to-audio generation synchronized with videos (Mo, 2024) View paper
    - Multi-Scene Text-to-Video with Temporal Captions (1 papers)
    - [11] Talc: Time-aligned captions for multi-scene text-to-video generation (Bansal, 2024) View paper
    - Multimodal Cinematic Video Synthesis (1 papers)
    - [36] Multimodal Cinematic Video Synthesis Using Text-to-Image and Audio Generation Models (S. Sridhar, 2025) View paper
    - Music Visualization and Audio-Driven Video Generation (1 papers)
    - [30] A deep learning approach for music visualization: From audio features to descriptive video generation (Fan Huang, 2025) View paper
- Retrieval and Matching with Temporal Alignment
    - Text-Video Retrieval with Local Alignment (1 papers)
    - [19] Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision (Xiaohan Wang, 2022) View paper
    - Temporal Video-Language Pre-training for Localization (1 papers)
    - [12] Temporal perceiving video-language pre-training (Ma Fan, 2023) View paper
    - Weakly-Synchronized Sequence Matching (1 papers)
    - [40] Video2Subtitle: Matching Weakly-Synchronized Sequences via Dynamic Temporal Alignment (Ben Xue, 2022) View paper
    - Caption Alignment and Temporal Localization (1 papers)
    - [42] Caption alignment for low resource audio-visual data (Vighnesh Reddy Konda, 2020) View paper
    - Retrieval-Augmented Generation with Aligned Captions (1 papers)
    - [50] Video Enriched Retrieval Augmented Generation Using Aligned Video Captions (Kevin Dela Rosa, 2024) View paper
- Datasets and Evaluation Frameworks
    - Large-Scale Movie Description Datasets (1 papers)
    - [22] Movie description (Atousa Torabi, 2017) View paper
    - Surveys and Comparative Analyses (2 papers)
    - [43] Generative AI for Text-to-Video Generation: Recent Advances and Future Directions (K Hayawi, 2025) View paper

- ◦ [44] RECENT ADVANCES IN AUDIO-VISUAL-LANGUAGE MODELING (Kairui Zhang, 2025) View paper
- ◦ Natural Video Description Systems (1 papers)
- ◦ [37] Generating Natural Video Descriptions via Multimodal Processing. (Qin Jin, 2016) View paper

Narrative

Core task: audiovisual video captioning with temporal alignment. The field addresses how to generate natural language descriptions of video content by jointly modeling visual frames and audio signals while respecting their temporal structure. The taxonomy reveals several complementary research directions. Audiovisual Representation Learning and Alignment focuses on learning shared or coordinated embeddings that capture cross-modal correspondences, often through contrastive or alignment objectives. Audiovisual Caption Generation Architectures explores encoder-decoder frameworks and attention mechanisms tailored to fuse multimodal streams and produce coherent captions. Large Language Model-Based Audiovisual Understanding investigates how pretrained language models can be adapted or prompted to incorporate video and audio inputs. Audiovisual Reasoning and Question Answering extends beyond captioning to interactive tasks requiring deeper semantic inference. Cross-Modal Generation and Synchronization examines bidirectional synthesis problems, such as generating audio from video or vice versa, which inform alignment strategies. Retrieval and Matching with Temporal Alignment studies how to index and retrieve video segments based on multimodal queries, while Datasets and Evaluation Frameworks provides the benchmarks and metrics that ground empirical progress across all branches.

Several active lines of work highlight key trade-offs and open questions. One cluster emphasizes explicit temporal modeling: methods like STELLA[2] and Temporal Perceiving[12] design architectures that track event boundaries and align audio-visual cues at fine-grained time scales, whereas others adopt coarser segment-level fusion. Another line leverages large-scale pretraining and prompting strategies, as seen in Daily-Omni[1] and Video Enriched RAG[50], which integrate audiovisual encoders with language models to handle diverse reasoning tasks. The original paper, AVoCaDO[0], sits within the Temporally Orchestrated Audiovisual Captioning branch, emphasizing coordinated temporal orchestration of audio and visual streams during caption generation. Compared to earlier works like Watch Listen Describe[26] or Align and Tell[19], AVoCaDO[0] appears to prioritize tighter synchronization mechanisms and richer temporal context, aligning closely with recent efforts such as NarrativeBridge[15] and MIRA-CAP[17] that also stress fine-grained temporal alignment. The central challenge remains balancing computational efficiency with the need to capture long-range dependencies and subtle audio-visual interactions across extended video sequences.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Sibling Subtopics

- **Adaptive and Memory-Augmented Captioning Frameworks** (leaves: 1, papers: 3)
- Scope: Frameworks incorporating adaptive attention, memory banks, retrieval augmentation, or hierarchical feature selection to enhance caption quality and temporal coherence.
- Exclude: Excludes static attention or non-memory-based methods; those belong under standard captioning architectures.
- **Dense Video Captioning with Multimodal Fusion** (leaves: 1, papers: 2)
- Scope: Methods for dense video captioning that localize and describe multiple events using multimodal inputs including audio, visual, and optionally speech features.
- Exclude: Excludes single-event or visual-only dense captioning; those belong under visual-centric dense captioning.
- **Enhanced Architectures with Temporal Coherence** (leaves: 1, papers: 2)
- Scope: Advanced architectures employing transformers, spatio-temporal graphs, or hybrid frameworks to enhance temporal coherence and computational scalability in video captioning.
- Exclude: Excludes basic encoder-decoder models; those belong under standard captioning architectures.
- **Hierarchical and Cross-Modal Attention for Captioning** (leaves: 1, papers: 2)
- Scope: Architectures employing hierarchical attention, cross-modal fusion, or globally-locally aligned attention mechanisms to integrate audio and visual features for caption generation.
- Exclude: Excludes single-modality or non-hierarchical approaches; those belong under visual-only or simple fusion categories.
- **Multimodal Feature Fusion for Video Description** (leaves: 1, papers: 2)
- Scope: Approaches combining visual and audio features through concatenation, fusion networks, or hybrid pipelines for generating video descriptions.
- Exclude: Excludes methods with sophisticated attention or alignment mechanisms; those belong under cross-modal attention categories.
- **Preference-Aligned and Training-Free Captioning** (leaves: 1, papers: 2)
- Scope: Methods using preference optimization, direct preference learning, or training-free pipelines to align generated captions with human preferences or leverage off-the-shelf models.
- Exclude: Excludes standard supervised training approaches; those belong under supervised captioning architectures.
- **Visual-Centric Dense and Temporal Captioning** (leaves: 1, papers: 5)
- Scope: Dense video captioning approaches primarily leveraging visual features with temporal modeling, semantic concept extraction, or causal-temporal narrative structures.
- Exclude: Excludes multimodal fusion methods; those belong under multimodal dense captioning.
- **Visual-Syntactic and Temporal Composition Models** (leaves: 1, papers: 2)
- Scope: Models explicitly incorporating syntactic structure, visual-syntactic embeddings, or temporal composition to improve grammatical correctness and temporal understanding in captions.
- Exclude: Excludes purely semantic or attention-based methods; those belong under semantic captioning categories.

Contributions Analysis

Overall novelty summary. The paper introduces AVoCaDO, an audiovisual video captioner emphasizing temporal orchestration between audio and visual modalities through a two-stage post-training pipeline. According to the taxonomy, this work resides in the 'Temporally Orchestrated Audiovisual Captioning' leaf under 'Audiovisual Caption Generation Architectures'. Notably, this leaf contains only the original paper itself with no sibling papers, suggesting this specific formulation of temporal orchestration represents a relatively sparse research direction within the broader field of 50 papers across 35 leaf nodes.

The taxonomy reveals that AVoCaDO's neighboring research directions include 'Hierarchical and Cross-Modal Attention for Captioning' (2 papers), 'Dense Video Captioning with Multimodal Fusion' (2 papers), and 'Visual-Centric Dense and Temporal Captioning' (5 papers). These adjacent leaves focus on attention mechanisms, event localization, and temporal modeling but without the explicit 'orchestration' framing. The broader 'Audiovisual Caption Generation Architectures' branch contains 8 distinct approaches, indicating moderate diversity in architectural strategies. AVoCaDO's emphasis on coordinated temporal alignment distinguishes it from methods prioritizing hierarchical fusion or memory-augmented frameworks in neighboring categories.

Among 20 candidates examined across three contributions, the core AVoCaDO system shows one refutable candidate from 10 examined, while the two-stage SFT-GRPO pipeline shows zero refutable candidates from 10 examined. The tailored reward functions contribution was not evaluated against candidates. This limited search scope suggests that while the overall audiovisual captioning concept has prior work, the specific combination of temporal orchestration with GRPO-based reinforcement learning appears less explored. The single refutable candidate for the core system indicates some overlap with existing audiovisual captioning approaches, though the extent of novelty depends on implementation details not captured in this top-20 semantic search.

Based on the limited literature search of 20 candidates, AVoCaDO appears to occupy a relatively underexplored niche within audiovisual captioning, particularly regarding its orchestration-focused architecture and reinforcement learning pipeline. The taxonomy structure shows this as an isolated leaf, though this may reflect the specific categorization criteria rather than absolute novelty. A more exhaustive search beyond top-20 semantic matches would be needed to definitively assess whether the temporal orchestration and GRPO integration represent substantial advances over the broader landscape of 50 papers in this taxonomy.

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

Contribution 1: AVoCaDO: Audiovisual video captioner with temporal orchestration

Description: The authors introduce AVoCaDO, a model specifically designed for audiovisual video captioning that emphasizes temporal alignment between visual and auditory events. This model addresses the limitation of existing vision-centric approaches by jointly processing both modalities to generate temporally coherent captions.

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. Masked Generative Video-to-Audio Transformers with Enhanced Synchronicity

URL: View paper

Brief Assessment

Masked Generative Transformers[61] focuses on video-to-audio generation (synthesizing audio from visual features), not audiovisual video captioning with temporal alignment. The tasks are fundamentally different: one generates audio signals, the other generates textual descriptions.

2. Pite: Pixel-temporal alignment for large video-language model

URL: View paper

Brief Assessment

PITE[18] focuses on pixel-temporal alignment for video-language models using object trajectories across spatial and temporal dimensions, not specifically on audiovisual captioning with temporal alignment between audio and visual modalities. The candidate addresses video understanding through trajectory-guided training rather than joint audio-visual processing.

3. Unified video-language pre-training with synchronized audio

URL: View paper

Brief Assessment

Unified Synchronized Audio[8] focuses on video-language pre-training for retrieval tasks using contrastive learning and masked modeling, not on generating temporally-aligned audiovisual captions as AVoCaDO does.

4. Text-Driven Synchronized Diffusion Video and Audio Talking Head Generation

URL: View paper

Brief Assessment

Text-Driven Talking Head[65] focuses on generating synchronized talking head videos from text inputs for video conferencing, not on audiovisual video captioning with temporal alignment. The candidate addresses video generation, while the original addresses video understanding and caption generation.

5. Learning audio-video modalities from image captions

URL: View paper

Brief Assessment

Learning from Captions[63] focuses on mining video-text pairs from image captions for retrieval and captioning tasks, not on temporal orchestration between audio and visual modalities for caption generation.

6. InVideo Search: Scene Description Clustering and Integrating Image and Audio Captioning for Enhanced Video Search

URL: View paper

Brief Assessment

InVideo Search[64] focuses on scene description clustering and integrating separate image and audio captions for video search tasks, not on developing a unified audiovisual captioning model with temporal orchestration like AVoCaDO.

7. Multi-modal Dense Video Captioning

URL: View paper

Prior Art Analysis

Multi-Modal Dense[47] demonstrates that prior work exists on audiovisual video captioning with temporal alignment. The candidate paper presents a multi-modal dense video captioning (MDVC) framework that processes audio, visual, and speech modalities jointly to generate temporally coherent captions. The paper explicitly addresses temporal alignment between modalities and demonstrates that joint processing of audio and visual information improves caption quality compared to vision-only approaches. This work predates the original paper and establishes similar technical contributions in audiovisual temporal alignment for video captioning.

Evidence

Evidence 1 - **Rationale**: Both papers address the same core problem: generating video captions that integrate audio and visual modalities. Multi-Modal Dense[47] explicitly presents a framework for audiovisual video captioning, demonstrating that this approach was already established in prior work. - **Original**: audiovisual video captioning aims to generate semantically rich descriptions with temporal alignment between visual and auditory events, thereby benefiting both video understanding and generation. in this paper, we present a vocado, a powerful audiovisual video captioner driven by the temporal orche... - **Candidate**: dense video captioning is a task of localizing interesting events from an untrimmed video and producing textual description (captions) for each localized event. most of the previous works in dense video captioning are solely based on visual information and completely ignore the audio track. however,...

Evidence 2 - **Rationale**: Both papers describe architectures that process multiple modalities (audio, visual, speech) and align them temporally for caption generation. Multi-Modal Dense[47] demonstrates a complete framework for this task, establishing prior work in audiovisual temporal alignment. - **Original**: built upon qwen2.5-omni (xu et al., 2025), which aligns visual and audio signals via interleaved token sequences, a v ocado is further enhanced through a two-stage post-training pipeline: (1) a v ocado sft, where we collect and construct a dataset of 107k high-quality audiovisual video-caption pairs... - **Candidate**: to produce inputs from audio, visual, and speech modalities, we use inflated 3d convolutions (i3d) [6] for visual and vggish network [15] for audio modalities. for speech representation as a text, we employ an external asr system [1]. to represent the text into a numerical form, we use a similar text...

Evidence 3 - **Rationale**: Both papers identify the same limitation in prior work (vision-centric approaches) and propose the same solution (joint audiovisual processing). Multi-Modal Dense[47] explicitly implements temporal alignment of audio and visual modalities, demonstrating prior work on this contribution. - **Original**: despite notable progress in recent video captioning models (xu et al., 2024; chai et al., 2024; yuan et al., 2025; shi et al., 2025; ren et al., 2024), most existing approaches remain predominantly vision-centric, often overlooking the rich semantic cues embedded in audio signals. in practice, audit... - **Candidate**: in contrast, we build our model to utilize video frames, raw audio signal, and the speech content in the caption generation process. to this end, we deploy automatic speech recognition (asr) system [1] to extract time-aligned captions of what is being said (similar to subtitles) and employ it alongs...

8. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding

URL: View paper

Brief Assessment

Temporal Working Memory[62] focuses on a query-guided segment selection mechanism for multimodal understanding across various tasks (QA, captioning, retrieval), not specifically on audiovisual video captioning with temporal alignment as a primary contribution. The candidate addresses temporal modeling through selective attention mechanisms rather than joint audiovisual caption generation with temporal orchestration.

9. Diverse and aligned audio-to-video generation via text-to-video model adaptation

URL: View paper

Brief Assessment

Diverse Audio-Video Generation[5] focuses on audio-to-video generation (creating videos from audio inputs), not audiovisual video captioning (generating text descriptions from audio-visual content). The tasks are fundamentally different: generation vs. captioning.

10. VideoComp: Advancing Fine-Grained Compositional and Temporal Alignment in Video-Text Models

URL: View paper

Brief Assessment

VideoComp[4] focuses on evaluating and improving compositional understanding in video-text models through benchmark construction and preference learning, not on audiovisual video captioning with temporal alignment between audio and visual modalities.

Contribution 2: Two-stage post-training pipeline with SFT and GRPO

Description: The authors propose a two-stage training approach: first, supervised fine-tuning on 107K curated audiovisual video-caption pairs emphasizing temporal alignment; second, Group Relative Policy Optimization using tailored reward functions to improve temporal coherence, dialogue accuracy, and caption quality while reducing repetition collapse.

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. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning

URL: View paper

Brief Assessment

VideoChat-R1[51] focuses on reinforcement fine-tuning for spatio-temporal perception tasks in video understanding, not audiovisual video captioning with temporal alignment between audio and visual modalities as in the original paper.

2. VidChain: Chain-of-Tasks with Metric-based Direct Preference Optimization for Dense Video Captioning

URL: View paper

Brief Assessment

VidChain[55] focuses on dense video captioning through chain-of-tasks decomposition and metric-based DPO, not on audiovisual temporal alignment with SFT and GRPO for general video captioning.

3. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms

URL: View paper

Brief Assessment

TempSamp-R1[57] focuses on temporal video grounding tasks using a modified GRPO approach with off-policy supervision, not audiovisual video captioning with temporal alignment between audio and visual modalities as in the original paper.

4. VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning

URL: View paper

Brief Assessment

VideoRFT[56] focuses on video reasoning with chain-of-thought (CoT) generation and semantic-consistency rewards, while the original paper targets audiovisual video captioning with temporal alignment between audio and visual modalities. The tasks and optimization objectives differ fundamentally.

5. Video-lmm post-training: A deep dive into video reasoning with large multimodal models

URL: View paper

Brief Assessment

Video-LMM Post-Training[53] surveys general post-training methodologies for video-LMMs across multiple approaches, while the original paper proposes a specific two-stage pipeline tailored for audiovisual video captioning with temporal alignment emphasis and custom reward functions for dialogue accuracy and repetition reduction.

6. Thinking With Bounding Boxes: Enhancing Spatio-Temporal Video Grounding via Reinforcement Fine-Tuning

URL: View paper

Brief Assessment

Thinking Bounding Boxes[54] focuses on spatio-temporal video grounding (localizing objects in videos with bounding boxes) using reinforcement fine-tuning, not audiovisual video captioning. The technical objectives and reward designs are fundamentally different from AVOCADO's caption-oriented SFT+GRPO pipeline.

7. Tempo-R0: A Video-MLLM for Temporal Video Grounding through Efficient Temporal Sensing Reinforcement Learning

URL: View paper

Brief Assessment

Tempo-R0[60] applies a two-stage training approach (SFT followed by GRPO) specifically for temporal video grounding tasks, not audiovisual video captioning. The technical focus, reward design, and task objectives differ fundamentally from the original paper's audiovisual caption generation pipeline.

8. TEMPLE: Temporal Preference Learning of Video LLMs via Difficulty Scheduling and Pre-SFT Alignment

URL: View paper

Brief Assessment

TEMPLE[58] uses DPO (Direct Preference Optimization) with a progressive pre-SFT alignment strategy, not GRPO. The training order and methodology differ fundamentally from the original paper's two-stage SFT-then-GRPO approach.

9. OwlCap: Harmonizing Motion-Detail for Video Captioning via HMD-270K and Caption Set Equivalence Reward

URL: View paper

Brief Assessment

OwlCap[59] uses a two-stage pipeline (SFT on HMD-270k + GRPO with CSER reward), but focuses on balancing motion-detail in video captioning rather than audiovisual temporal alignment. The reward design (CSER with unit-to-set matching) differs fundamentally from AVOCADO's temporal coherence and dialogue accuracy rewards.

10. Reinforcement Learning Tuning for VideoLLMs: Reward Design and Data Efficiency

URL: View paper

Brief Assessment

Reinforcement Learning Tuning[52] focuses on GRPO-based reinforcement learning tuning for video understanding tasks with dual-reward formulation (discrete semantic + continuous temporal IoU rewards), not on audiovisual video captioning with temporal alignment between audio and visual modalities as in the original paper.

Contribution 3: Tailored reward functions for audiovisual captioning optimization

Description: The authors design three complementary reward functions for GRPO training: a checklist-based reward for comprehensive audiovisual keypoint coverage, a dialogue-based reward for ASR fidelity and speaker identification, and a length-regularized reward to mitigate repetition collapse and control caption length.

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

No high-similarity text segments were detected across any compared papers.

References

- [0] AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration View paper
- [1] Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities View paper
- [2] STELLA: continual audio-video pre-training with spatio-temporal localized alignment View paper
- [3] Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding View paper
- [4] VideoComp: Advancing Fine-Grained Compositional and Temporal Alignment in Video-Text Models View paper
- [5] Diverse and aligned audio-to-video generation via text-to-video model adaptation View paper
- [6] Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation View paper
- [7] Enhancing video-language representations with structural spatio-temporal alignment View paper
- [8] Unified video-language pre-training with synchronized audio View paper
- [9] Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation View paper
- [10] Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation View paper
- [11] Talc: Time-aligned captions for multi-scene text-to-video generation View paper
- [12] Temporal perceiving video-language pre-training View paper
- [13] Aligned better, listen better for audio-visual large language models View paper
- [14] ATMNet: Adaptive Two-Stage Modular Network for Accurate Video Captioning View paper
- [15] NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative View paper
- [16] Text-to-audio generation synchronized with videos View paper
- [17] MIRA-CAP: Memory-Integrated Retrieval-Augmented captioning for State-of-the-Art image and video captioning View paper
- [18] Pite: Pixel-temporal alignment for large video-language model View paper
- [19] Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision View paper
- [20] Learning semantic concepts and temporal alignment for narrated video procedural captioning View paper
- [21] Target-Aware Spatio-Temporal Reasoning via Answering Questions in Dynamics Audio-Visual Scenarios View paper
- [22] Movie description View paper
- [23] Audio-Sync Video Generation with Multi-Stream Temporal Control View paper
- [24] Explicit Temporal-Semantic Modeling for Dense Video Captioning via Context-Aware Cross-Modal Interaction View paper
- [25] Video description combining visual and audio features View paper
- [26] Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning View paper
- [27] Toward Scalable Video Narration: A Training-free Approach Using Multimodal Large Language Models View paper
- [28] Understanding temporal structure for video captioning View paper

- [29] Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding* View paper
- [30] A deep learning approach for music visualization: From audio features to descriptive video generation View paper
- [31] Video Captioning with Spatio-Temporal Graph Transformers View paper
- [32] UNICORN: A Unified Causal Video-Oriented Language-Modeling Framework for Temporal Video-Language Tasks View paper
- [33] Enhanced Hybrid Framework and Comparative Analysis of Deep Learning Architectures for Video Captioning View paper
- [34] SkimCap: A Transformer-Based Video Captioning Method with Adaptive Attention and Hierarchical Skimming Features View paper
- [35] AVC-DPO: Aligned Video Captioning via Direct Preference Optimization View paper
- [36] Multimodal Cinematic Video Synthesis Using Text-to-Image and Audio Generation Models View paper
- [37] Generating Natural Video Descriptions via Multimodal Processing. View paper
- [38] LiveChat: Video Comment Generation from Audio-Visual Multimodal Contexts View paper
- [39] Visually-Aware Audio Captioning With Adaptive Audio-Visual Attention View paper
- [40] Video2Subtitle: Matching Weakly-Synchronized Sequences via Dynamic Temporal Alignment View paper
- [41] Center-enhanced video captioning model with multimodal semantic alignment. View paper
- [42] Caption alignment for low resource audio-visual data View paper
- [43] Generative AI for Text-to-Video Generation: Recent Advances and Future Directions View paper
- [44] RECENT ADVANCES IN AUDIO-VISUAL-LANGUAGE MODELING View paper
- [45] Whats in a Video: Factorized Autoregressive Decoding for Online Dense Video Captioning View paper
- [46] Enhanced visual multi-modal fusion framework for dense video captioning View paper
- [47] Multi-modal Dense Video Captioning View paper
- [48] Exploring Audio-Visual Concepts for Dense Video Captioning View paper
- [49] DialogMCF: Multimodal Context Flow for Audio Visual Scene-Aware Dialog View paper
- [50] Video Enriched Retrieval Augmented Generation Using Aligned Video Captions View paper
- [51] Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning View paper
- [52] Reinforcement Learning Tuning for VideoLLMs: Reward Design and Data Efficiency View paper
- [53] Video-lmm post-training: A deep dive into video reasoning with large multimodal models View paper
- [54] Thinking With Bounding Boxes: Enhancing Spatio-Temporal Video Grounding via Reinforcement Fine-Tuning View paper
- [55] VidChain: Chain-of-Tasks with Metric-based Direct Preference Optimization for Dense Video Captioning View paper
- [56] VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning View paper
- [57] Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms View paper
- [58] TEMPLE: Temporal Preference Learning of Video LLMs via Difficulty Scheduling and Pre-SFT Alignment View paper
- [59] OwlCap: Harmonizing Motion-Detail for Video Captioning via HMD-270K and Caption Set Equivalence Reward View paper
- [60] Tempo-R0: A Video-MLLM for Temporal Video Grounding through Efficient Temporal Sensing Reinforcement Learning View paper
- [61] Masked Generative Video-to-Audio Transformers with Enhanced Synchronicity View paper
- [62] Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding View paper
- [63] Learning audio-video modalities from image captions View paper
- [64] InVideo Search: Scene Description Clustering and Integrating Image and Audio Captioning for Enhanced Video Search View paper
- [65] Text-Driven Synchronized Diffusion Video and Audio Talking Head Generation View paper