HAS: Highlight-guided Attention Steering for Multimodal LLM Video Summarization

Rui Chu and Yingjie Lao

Tufts University, Medford, MA, USA

Abstract. Video understanding has become more and more important with the growth of Artificial Intelligence (AI) for video generation. Recently, Multimodal Large Language Model (M-LLM) has shown its capability in video understanding. Video summarization, a specific domain of video understanding, has proven its importance for efficient navigation and retrieval. Both video understanding and video summarization require a good selection of key frames in a video. Current video summarization methods heavily focus on the selected key frames and correlated segment captions. However, existing approaches overlook the perspective of treating the importance of the frames globally. We argue that using discrete selected frames for summarization will not only reduce the understanding coherence, but also lost important information in the video, as well as wasting the original capacity of the MLLMs. In this paper, we propose **HAS**, a **H**ighlight-guided **A**ttention **S**teering method for video summarization. We consider a challenging but practical setting where the video given to MLLMs for summarize should be continuous but with highlight guidance. HAS mainly consists of two parts: The first part is to find a continuous frame-level highlight distribution for the video globally. The second part is to apply the highlight distribution as an attention steering vector for the MLLM, targeting a better understanding of the video, and thus during the model inference time, putting more attention on the highlighted frames, while avoiding lost entire information on less highlighted frames through putting less attention instead of forgetting them. We evaluated HAS on a variety of benchmarks, and it has shown convincing performance in video summarization.

Keywords: Attention Steering · Multimodal LLM · Video Summary

1 Introduction

Video understanding [5] has become more and more important with the development of video generation [49, 51], image and 3D generative visual modeling and robustness [16, 32], knowledge retrieval and in context learning [7, 8], animation rendering [11], and interactive applications, since the video data has exceeded human capacity for consumption. Video summarization, a specific sub-domain of video understanding, requires an efficient processing method which *firstly,* capture essential content (frames) and *secondly,* process the content into concise summaries [24]. Recent multimodal large language models (M-LLMs) [48]

process video as a sequence of visual tokens (time-ordered embeddings of frame-level tubelets) via cross-attention [27] and output instruction-following answers, making both video understanding and summarization feasible. Although MLLM can handle varieties of video understanding tasks, how to guide the model efficiently summarize the video during inference time while avoiding losingtoo much information is worth exploring.

Earlier video summarization through neural networks always require a training progress. Video summarization tasks, compared to traditional video understanding tasks, requires more consistency of the understanding of the entire video [6]. Many existing methods can summarize the main content, a common pipeline is still to first select discrete frames or segments (e.g., keyframes/keyshots) and then summarize based on the selected subset [24,34]. In brief, 1) finding the important frames; 2) understanding each frames, and summarize based on the combination of single frame understandings. However, such hard selection overwhelmingly rely on discrete selected highlight frame sequence window [24], which not only overlooked the MLLM capability of finding highlight through **attention** mechanism [44], but also impact the coherence of video understanding and summarization, harming downstream retrieval.

Even though early works has made a video highlight into an import-score distribution coorelated to each frame time, using the highlight distribution to continuous enhance the video summarization has been neglected. In the Nature Language Processing (NLP) domain, inference-time attention steering has been studied as a way to guide a model to focus on user-specified important parts . For example, PASTA proposes a post-hoc attention steering approach that reweights attention at inference time, so the model can read emphasized tokens more like human readers. With the development of multimodal LLMs (MLLMs), inference-time attention intervention has also been explored for vision-language settings. FarSight shows that modifying the decoding-time token interaction (via causal-mask-based intervention) can mitigate hallucinations in MLLMs, and it is effective on both image and video benchmarks [42]. This suggests that inference-time manipulation of attention-related mechanisms can be a practical lever for improving multimodal reasoning and generation [42].

More broadly, lightweight inference-time control methods are popular, because they can adjust model behaviors without expensive finetuning. At the systems level, complementary advances in efficient [13], hardware-aware, and privacy-preserving AI computation also aim to make advanced model inference more practical under real deployment constraints [45,50]. A representative direction is vector/activation steering, where a steering vector is injected into hidden activations during generation to slightly shift the output distribution [23, 40]. This is similar to human reading: if a reader has a guidance about what is important, it is easier to understand the content; in models, such steering signals can also act as a soft guidance to allocate computation and attention [23].

Similarly, Considering the way of a real human summarizing a video is to *firstly*, watch the entire video, and *secondly*, recalling the video and summarizing by memorized more on important moments and less on less exciting mo-

ments [43]. The current work of video summarization [24] is more close the scenario to skimming the video first and summarize based on the skimmed frames, which can cause inconsistency. However, it is still unclear how to use a simple and controllable steering signal to guide a Video-MLLM to summarize videos more precisely, while keeping the video input continuous and avoiding losing video details. Meanwhile, the community has built query-conditioned highlight supervision for videos (e.g., clip-wise saliency scores conditioned on natural-language queries) [25, 33], but these signals are mostly used for highlight detection or moment retrieval, not as an internal attention prior for generative video summarization. These observation motivate the research question we are trying to address: *How to 1) have a highlight distribution of a video (acting as a teacher), and thus 2) using the highlight distribution to guide the M-LLM generate a more consistent and comprehensive video summarization?*

In this work, we propose **HAS**, a *Highlight-guided Attention Steering* framework for video summarization. HAS inspired from both text summarization and image caption generation, where the performance can be improved through steering method. Meanwhile, HASis also inspired by the LLM quantization [36] and distillation [21], both aim to compress information while retaining essential knowledge: we consider highlighting video as a quantization progress, and use the highlight as a teacher signal to guide the MLLM generating progress, similar to a distillation process. HAS consists of two parts: 1) introduces a smooth way to find a continuous highlight distribution for a selected video based on the entire frame sequence; 2) introduces a novel *Attention Steering* method for MLLM to better use the highlight distribution as a steering vector to guide the attention of MLLM towards the high scoring pieces while remaining less attention towards other frames for final summarization. The high-level workflow can be illustrated in Fig. 1.

Our contributions are summarized as followed,

- To the best of our knowledge, HAS is the first work to use attention steering to guide the MLLM, and for a specific goal, video summarization.
- HAS treat the video summarization from a continuous perspective, enhancing the consistency of LLM-based video summarization works.
- HAS is a inference-time light-weight, backbone model training free framework, which is easy to be adapted towards pre-trained models.
- Through extensive experiments, HAS achieved an outstanding performance on both coherence, summarization performance and scalability.

2 Related Works

2.1 Video summarization

A standard formulation in video summarization is to first predict a frame- or clip-level importance curve and then convert it into hard keyframe or keyshot selection under a budget [24]. Early work in this line is mostly visual-only and

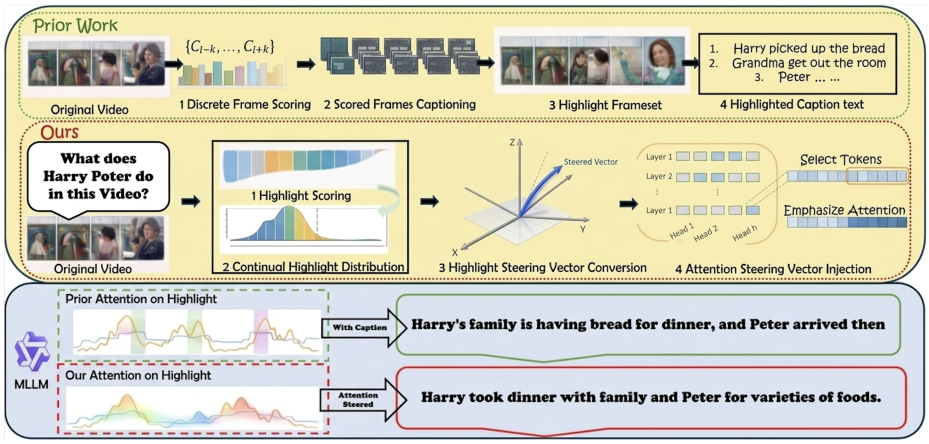


Fig. 1: Top: Unlike prior works assign *discrete* importance scores and *hard-select* a few highlights before summarizing, HASpreserves the full video context and treats highlighting as *continual* attention guidance through steering vector. **Bottom:** While smoothly bias a frozen video MLLM toward highlight moments, HAS does not neglect the peace time steps, better exploiting model capacity while reducing missed evidence for more coherent and faithful summaries.

relies on supervised temporal modeling; TVSum is a representative benchmark and formulation that summarizes videos from shot-level importance annotations [39]. CLIP-It extends this paradigm to multimodal summarization, and unifies generic and query-focused settings by learning frame importance from a language-guided transformer [34]. Scaling Up Video Summarization Pretraining with Large Language Models further shows that LLM-generated supervision can improve summarization pretraining at scale [2]. Recent efforts also expand the summarization setting beyond unimodal V2V: V2Xum-LLM unifies video-to-video, video-to-text, and joint cross-modal summarization via temporal prompt instruction tuning [20]. In parallel, VISTA builds a large-scale video-to-text benchmark for scientific presentations and highlights the challenges of long-form, domain-specific abstractive summarization. More recent MLLM-based summarization still follows the same score-then-select pipeline: LLMVS converts frames into captions, estimates local importance with an LLM, and refines it with a global aggregator before summary construction [24]. A related efficiency line also performs hard visual selection before downstream reasoning, including M-LLM Based Video Frame Selection, Flexible Frame Selection, Adaptive Keyframe Sampling, and the agentic AKeyS framework [5, 10, 19, 42]. A zero-shot alternative further shows that contrastive features can already approximate frame importance without task-specific retraining [35]. Different from all these methods, we do not use the highlight signal as a final selector; we use a coarse continuous highlight distribution only as a steering prior for MLLM summarization.

2.2 Query-based highlight

Query-based moment retrieval and highlight detection are closely related because they also learn a query-conditioned temporal relevance curve, but their output is still a moment boundary or a highlight score rather than a generated summary. Classic highlight detection methods (without language queries) also learn a temporal highlight score curve via ranking supervision, e.g., the pairwise deep ranking framework for first-person video summarization [46]. QVHighlights establishes this setting with natural-language queries, relevant moment annotations, and clip-level saliency labels [25]. UMT unifies moment retrieval and highlight detection in a shared multimodal transformer. QD-DETR further learns query-dependent video representations through early cross-attention and negative video-query training, leading to better localization and saliency estimation [33]. This line is the closest to our highlight prior, but the highlight curve is their final prediction target, while in our method it is only an intermediate signal that guides summarization.

2.3 Attention Steering

Another relevant line shows that attention can be modified at inference time as a practical control interface. Attention Biasing and Context Augmentation demonstrates that encoder-decoder transformers can be controlled in zero-shot fashion by directly biasing cross-attention during generation [17]. PASTA makes this idea explicit for LLMs by reweighting selected attention heads so that the model attends more to user-emphasized text spans, without parameter updates. In multimodal and video generation, FarSight modifies the causal mask during decoding to improve visual token propagation and reduce hallucination [42]. Closely related video-MLLM work also changes how visual tokens compete for attention: Vista-LLaMA adjusts relative token distance to prevent visual evidence from being ignored in long generations [30], MASH-VLM disentangles spatial and temporal attention to reduce action-scene hallucination [3], and SEAL learns semantic attention over compact long-video units instead of dense raw frames. These works support the general paradigm that attention-level control is effective, but none of them injects a frame-level highlight distribution for video summarization.

A final related question is how to construct and inject a reasonable highlighting prior. Highlight-Transformer provides direct evidence from text summarization: it introduces a highlighting matrix that explicitly increases attention weights on key phrases. Post-hoc attribution methods give another source of such priors. Excitation Backprop for RNNs localizes spatiotemporal evidence in video models without retraining [4]. Grad-CAM gives a simple gradient-based localization signal for visual backbones and is widely used as a saliency baseline [37]. Extremal Perturbations estimates smooth masks by measuring which input regions most affect the prediction [12]. These methods are mainly explanation tools rather than control mechanisms, but they support our assumption that a coarse continuous saliency signal can be extracted without dense labels and then reused to guide generation.

3 Methodology

In this section, we present HAS , the Highlight-guided Attention Steering method for video summarization.

We begin by introducing the problem statement for video summarization in Section 3.1 for better formulation, and then show the overview of our method in Section 3.2, illustrating notations and explaining Figure. 1. Afterwards, we illustrate HAS step by step in the following parts.

3.1 Problem Statement

The goal of video summarization is to find the best video summarization output text $\mathbf{O}$ through MLLM $\mathbf{M}$ given a selected video, considering as a sequence of frames $\mathbf{F}$; and a user query/prompt q. Let $\mathbf{F} = [\mathbf{F}_1, \ldots, \mathbf{F}_T]$ be a video sequence frame, T denotes the temporal length of the video. We are aiming at using continuous video information instead of discrete frame pieces for outputing best $\mathbf{O}$, as prior LLM-based video summarization works did [24], and maintaining best original capacity of MLLMs.

To this end, the optimization goal is to find the proper *attention steering* vector $\mathbf{V}$ for $\mathbf{M}$ so that during inference time, best guide MLLM $\mathbf{M}_\mathbf{V}$ through steering to generate the best optimized summarization output text $\mathbf{O}^*$, where

$$\mathbf{O}^* = \mathbf{M}_\mathbf{V}(\mathbf{F}, q) \tag{1}$$

3.2 Method Overview

As shown in Fig. 1, HAS consists of two stages: (i) constructing a prompt-conditioned temporal continuous highlight distribution prior from the input video, and (ii) converting the distribution into vectors and injecting into a MLLM at inference time via a attention steering policy.

Highlight construction Given a video $\mathbf{F}$ and q, we first obtain a *raw* highlight score sequence $\hat{\mathbf{h}} = [\hat{h}_1, \ldots, \hat{h}_T]$ using an off-the-shelf highlight/moment module. We then *calibrate* it to a smooth and bounded highlight distribution $\mathbf{h} = [h_1, \ldots, h_T] \in [0, 1]^T$ by applying temporal smoothing.

Attention Steering Injection We map the distribution to visual tokens and form a token-level attention (denoted as $\mathbf{V}$) for the Video-MLLM. The injection is governed by a small set of steering parameters following, which we tune/choose on a validation set for stable and effective summarization.

3.3 Highlight Distribution

Highlight generator. We reuse an off-the-shelf highlight generator $\mathcal{H}$ and only calibrate its output for steering. Given the input video $\mathbf{F}$ and prompt q, the generator produces a raw temporal score sequence $\hat{\mathbf{h}} = \mathcal{H}(\mathbf{F}, q)$. When the generator outputs clip-level scores, we linearly interpolate them to length T.

Algorithm 1 Prompt-conditioned highlight steering vector generation

Require: video frames $\mathbf{F} = [F_1, \ldots, F_T]$, Highlight Identifying Module $\mathcal{H}$, query q.
1: $\hat{\mathbf{h}} \leftarrow \mathcal{H}(\mathbf{F}, q)$
2: $\hat{\mathbf{h}} \leftarrow \text{Interp}(\hat{\mathbf{h}}, T)$
3: $\mathbf{h} \leftarrow \text{MinMaxNorm}(\hat{\mathbf{h}})$
4: $\mathbf{V} \leftarrow \text{Vectorize}(\mathbf{h})$
5: **return V**

Temporal calibration. Raw highlight curves are noisy and not continuous (smooth) because it is based on the frame selection and length of frame time window. In HAS needs to be continuous enough to guide attention. Thus, we firstly align the time frame the required T; and then, we normalize the score into the range of $[0, 1]$ (line 1 to 3 in Algorithm. 1). We solve a lightweight one-dimensional calibration problem on top of the off-the-shelf scores.

The final result is a bounded and continuous highlight distribution $\mathbf{h} = [h_1, \ldots, h_T]$, which will be converted into steering vectors in Section 3.4.

3.4 From Highlight Distribution to Steering Vector

As it is shown in line 4 of Algorithm. 1, vectorization is needed. The vectorization is inspired. Since the target MLLM attends over visual tokens rather than scalar frame scores, HAS lifts this frame-level distribution to the token level before attention intervention. Following prior work that converts explicit highlighting or external emphasis into attention bias or attention reweighting [17], we repeat each frame score over the P visual tokens extracted from that frame and form the steering vector

$$\mathbf{V} = \log(\text{Repeat}(\mathbf{h}, P) + \epsilon) \in \mathbb{R}^{TP}, \tag{2}$$

where $\text{Repeat}(\mathbf{h}, P)$ copies h_t to all P visual tokens of frame F_t. We add a small constant ϵ for numerical stability, which avoids collapsing low-highlight frames to $-\infty$ in log-space and keeps HAS as soft steering rather than hard frame selection. The logarithm makes $\mathbf{V}$ directly usable as an additive bias on attention logits in the next subsection. (Similar inference-time attention intervention has also been shown effective in MLLMs [42]).

3.5 Inference-time Attention Steering

Given the steering vector $\mathbf{V}$, HASintervenes on a small subset of visual cross-attention heads during decoding. This follows the same inference-time steering spirit as attention biasing [17] and PASTA, and is also consistent with the plug-and-play attention intervention view of FarSight [42]. For a selected head $(\ell, m) \in \mathcal{S}$, let $\mathbf{A}^{(\ell,m)}$ denote the pre-softmax attention logits from the current text queries to all visual tokens. To enable *differentiable* policy learning over

Algorithm 2 HAS inference-time attention steering

Require: video frames $\mathbf{F}$, prompt q, frozen MLLM M, Steering Vector $\mathbf{V}$, Attention steering head set $\mathcal{S}$, gate parameters $\{a_{\ell,m}\}$, strengths $\{\beta_{\ell,m}\}$

1: Encode $\mathbf{F}$ into visual tokens and feed them to M
2: **for** each decoding step **do**
3: **for** each selected cross-attention head $(\ell, m) \in \mathcal{S}$ **do**
4: Compute pre-softmax attention logits $\mathbf{A}^{(\ell,m)}$
5: $g_{\ell,m} \leftarrow \sigma(a_{\ell,m})$
6: **for** each query row i **do**
7: $\mathbf{A}^{(\ell,m)}_{i,:} \leftarrow \mathbf{A}^{(\ell,m)}_{i,:} + g_{\ell,m}\beta_{\ell,m}\mathbf{V}$
8: **end for**
9: Compute steered attention with softmax($\mathbf{A}^{(\ell,m)}$)
10: **end for**
11: Decode the next token
12: **end for**
13: **return** summary $\mathbf{O}^*$

heads, we introduce a continuous gate $g_{\ell,m} = \sigma(a_{\ell,m}) \in [0, 1]$ for each candidate head. HASapplies a row-wise additive bias

$$\tilde{\mathbf{A}}^{(\ell,m)}_{i,:} = \mathbf{A}^{(\ell,m)}_{i,:} + g_{\ell,m}\,\beta_{\ell,m}\mathbf{V}, \qquad (\ell, m) \in \mathcal{S}, \tag{3}$$

and leaves all other heads unchanged, where ℓ and m index the transformer layer and attention head, $\mathcal{S}$ denotes the candidate head set, $\sigma(\cdot)$ is the sigmoid function, $\beta_{\ell,m}$ is the head-wise steering strength, and i indexes the query row (with : spanning all visual tokens). Since $\mathbf{V}$ is constructed in log-space, Eq. (3) is equivalent to multiplicative reweighting of the unnormalized attention scores, but is easier to stabilize and easier to combine with existing causal or padding masks. The small constant ϵ in Eq. (2) avoids collapsing low-highlight frames to $-\infty$ in log-space, so HASremains a soft steering method rather than hard frame selection. Alg. 2 implements this intervention inside the standard autoregressive decoding loop: after encoding $\mathbf{F}$ into visual tokens (line 1), each generation step computes $\mathbf{A}^{(\ell,m)}$ (line 4), applies the gated bias in Eq. (3) for heads in $\mathcal{S}$ (line 5–8), and then proceeds with the usual softmax and token decoding using the modified logits (line 9–13).

The optimization target of this stage is only the steering policy $\Theta = \{a, \beta\}$, not the MLLM nor the highlight generator. We choose Θ on a held-out calibration/validation set by minimizing the standard teacher-forcing negative log-likelihood of the reference summaries [41, 44], while keeping the whole MLLM frozen, and we add a sparsity term to encourage using only a small subset of heads, following the common practice of sparse head selection for interpretability and regularization:

$$\min_{a,\beta}\ \mathcal{L}_{\mathrm{NLL}}(a, \beta)\ +\ \lambda_s \sum_{(\ell,m)\in\mathcal{S}} g_{\ell,m}. \tag{4}$$

Here λ_s is the sparsity weight that trades off summary likelihood and the number of activated heads. In practice, we optimize $\{a_{\ell,m}, \beta_{\ell,m}\}$ by minimizing the teacher-forcing NLL in Eq. (4) with a standard first-order optimizer (Adam [22]), while keeping the whole MLLM frozen. We optimize $\{a, \beta\}$ once on a held-out set; at test time, $\{a, \beta\}$ are fixed and thus Alg. 2 will performs forward-only steering.

4 Experiments

In this section, we evaluate HAS across a range of video summarization tasks. All experiments are conducted on Nvidia L40S GPUs.

4.1 Experimental Settings

Datasets and benchmarks. We evaluate HAS on three benchmark families that align with our cross-comparison tables. **(i) V2V summarization** on SumMe [15] and TVSum [39] measures the quality of frame/shot importance ranking against human annotations. **(ii) Cross-modal summarization** on VideoXum [28] evaluates *video-to-text (V2T)*, *video-to-video (V2V)*, and *joint video + text (V2VT)* summaries in a unified protocol. **(iii) Scientific video-to-text summarization** on VISTA [29] benchmarks long-form academic talk summarization, emphasizing both semantic quality and factual grounding.

Models. To avoid conclusions tied to a single backbone, we evaluate HAS as a *plug-and-play* inference-time module on a diverse set of *open-source* video MLLMs. Specifically, our backbone $\mathcal{M}$ is instantiated with: (i) **mPLUG-Owl3** [47], (ii) **LLaVA-NeXT-Interleave** [26], (iii) **Video-LLaVA**, (iv) **LLaMA-VID**, (v) **Video-ChatGPT** [31], and (vi) **Video-LLaMA**. All backbones use publicly available implementations and checkpoints released by the authors (subject to the corresponding base-model licenses). Unless stated otherwise, we keep $\mathbf{M}$ *frozen* and apply HAS purely at inference time.

Baselines. For **V2V summarization**, we compare against: **Visual-only** summarizers including VASNet [9], DSNet (anchor-based/anchor-free) [52], DMASum, PGL-SUM [1], MSVA [14], iPTNet, and CSTA [38]; **Visual+Text** methods CLIP-It [34], A2Summ [18], SSPVS, and the large-scale pretraining baseline by Argaw et al. [2]; and **LLM-centric** methods including the zero-shot LLM scoring baseline and LLMVS [24], as well as V2Xum-LLaMA from V2Xum-LLM [20]. For **cross-modal summarization** on VideoXum, we follow the established baselines and protocols from VideoXum [28] and V2Xum-LLM [20] (e.g., BLIP/Vid2Seq-based variants and V2Xum-LLaMA). For **scientific summarization**, we report representative model families and settings (zero-shot and fine-tuning) as benchmarked by VISTA [29].

Table 1: Human-aligned temporal importance estimation. Rank correlation (τ/ρ; higher is better) between predicted importance trajectories and human annotations on V2V summarization.

	SumMe		TVSum	
Method	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
Visual				
VASNet	0.160	0.170	0.160	0.170
DSNet-AB	0.051	0.059	0.108	0.129
DSNet-AF	0.037	0.046	0.113	0.138
DMASum	0.063	0.089	0.203	0.267
PGL-SUM	–	–	0.206	0.157
MSVA	0.200	0.230	0.190	0.210
iPTNet	0.101	0.119	0.134	0.163
CSTA	0.246	0.274	0.194	0.255
Visual + Text				
CLIP-It	–	–	0.108	0.147
A2Summ	0.108	0.129	0.137	0.165
SSPVS	0.192	0.257	0.181	0.238
Argaw et al.	0.130	0.152	0.155	0.186
LLM-centric				
LLM (zero-shot)	0.170	0.189	0.051	0.056
LLMVS	0.253	0.282	0.211	0.275
V2Xum-LLaMA	0.296	**0.378**	0.222	0.293
HAS (ours)	**0.298** ± 0.02	0.335 ± 0.03	**0.224** ± 0.02	**0.299** ± 0.03

Metrics. We report official metrics for each benchmark family.

SumMe/TVSum, we use Kendall's τ and Spearman's ρ rank correlations (higher is better), following prior practice and LLMVS [24]. For **VideoXum**, we evaluate V2T by BLEU-4 / METEOR / ROUGE-L / CIDEr, V2V by F1 / Spearman / Kendall, and V2VT by semantic alignment metrics (FCLIP / Cross-FCLIP) [20, 28]. For **VISTA**, we report ROUGE-Lsum (RLsum), BERTScore, and two video-grounded quality metrics VideoScore and FactVC [29]. Unless noted otherwise, larger values indicate better performance.

4.2 Main Performance

Human-Aligned Temporal Importance Estimation

Human-aligned temporal importance estimation. Table 1 shows that HASbetter aligns temporal importance with human annotations under both Kendall's τ and Spearman's ρ. The gains over score-then-select pipelines suggest that inference-time steering with a calibrated temporal prior improves *global* salience order-

Table 2: Joint cross-modal summarization and video–text consistency. We evaluate V2T text quality (BLEU-4/METEOR/ROUGE-L/CIDEr), V2V temporal salience (F1/Spearman/Kendall), and V2VT semantic alignment (FCLIP/Cross-FCLIP); higher is better.

	V2T				V2V			V2VT	
Method	B-4	M	R-L	C	F1	Spr.	Kend.	FCLIP	Cross-FCLIP
Frozen-BLIP	0.0	0.4	1.4	0.0	16.1	0.011	0.008	–	–
Vid2Seq-HCY	2.3	8.2	19.0	7.6	24.2	–	–	0.888	0.214
Vid2Seq-HC	2.7	8.5	19.8	8.4	24.5	–	–	0.892	0.217
Vid2Seq-HCV	2.7	8.4	19.8	8.3	25.1	–	–	0.899	0.200
VSUM-BLIP	–	–	–	–	21.7	0.207	0.131	–	–
TSUM-BLIP	5.6	11.8	24.9	20.9	–	–	–	–	–
VTSUM-BLIP	5.8	12.2	25.1	23.1	23.5	0.258	0.196	0.894	0.247
V2Xum-LLaMA-7B	5.8	12.3	26.3	26.9	29.0	0.298	0.204	0.931	0.253
V2Xum-LLaMA-13B	5.7	12.3	26.2	25.3	31.6	0.276	0.200	0.957	0.251
HAS (ours)	**6.0** ±0.03	12.7 ±0.3	**26.7** ±0.3	**28.0** ±0.8	32.0 ±0.5	0.304 ±0.02	0.230 ±0.02	**0.963** ±0.01	**0.258** ±0.007

ing rather than applying a trivial re-scaling. Overall, HASnarrows the gap between classical visual summarizers and LLM-centric approaches while remaining training-free.

Cross-Modal Consistency without Sacrificing Text Quality

Cross-modal summarization. Table 2 indicates that HASimproves cross-modal summarization in a balanced way: text quality (V2T) is maintained while temporal salience (V2V) and video–text consistency (V2VT) are strengthened.

The gain on semantic alignment metrics (FCLIP/Cross-FCLIP) suggests that HASbetter ties generated summaries to evidence-bearing segments even when exact temporal boundaries are ambiguous.This supports soft steering as a lightweight alternative to discrete extraction for joint V2T/V2V/V2VT evaluation.

Faithful Summarization under Long-Context Evidence

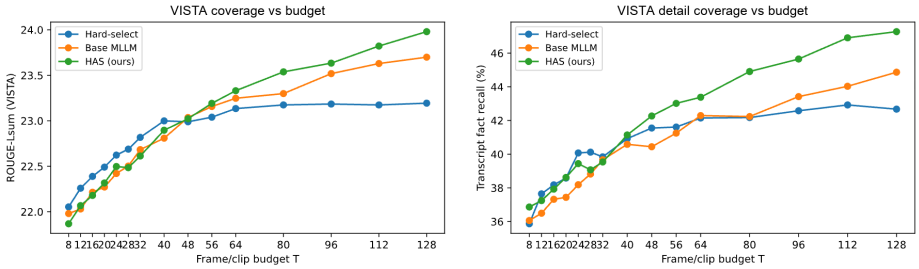
Grounded long-form summarization. Table 3 shows that HASimproves groundedness on long scientific talks: video–text alignment and factual consistency trend upward without sacrificing overall summary quality. This suggests that calibrated temporal steering encourages evidence usage during generation, reducing unsupported details that commonly arise in long-context settings. Taken together, HASoffers a practical inference-time mechanism for faithful summarization in high-stakes domains.

4.3 Coverage Performance

Coverage and information retention on VISTA. Fig. 2 exhibits a shared saturation trend: marginal gains diminish as T increases. However, the discrete extraction baseline plateaus substantially earlier, revealing an irreversible

Table 3: Grounded long-form summarization. We report summary quality (RLsum/BERTScore) together with video-grounded alignment (VideoScore) and factual consistency (FactVC); higher is better.

Method (setting)	RLsum	BERTScore	VideoScore	FactVC
GPT-o1 (zero-shot)	24.37	82.63	2.17	51.36
Gemini 2.0 (zero-shot)	24.29	82.64	2.02	52.02
LLaVA-NeXT-Interleave (zero-shot)	22.68	81.40	1.73	40.12
mPLUG-Owl3 (zero-shot)	22.84	81.39	1.77	42.07
Plan-mPLUG-Owl3* (zero-shot)	22.97	81.45	1.86	47.37
mPLUG-Owl3 (full FineTune)	32.91	84.22	3.28	71.94
Plan-mPLUG-Owl3 (full FineTune)	**33.25**	**84.37**	**3.33**	**75.41**
HAS (ours, averaged on selected backbones)	23.94	82.05	2.03	50.11



(a) Coverage (ROUGE-Lsum) vs. budget T. **(b)** Coverage (fact-unit recall) vs. budget T.

Fig. 2: Coverage vs. budget on VISTA. (a) ROUGE-Lsum and (b) transcript fact-unit recall versus T. Hard selection saturates early, while HASkeeps improving at mid-to-high budgets, indicating reduced information loss under fixed-budget inference.

bottleneck—once low-scored segments are removed, additional budget cannot restore missing evidence. HASshows delayed saturation and a clear divergence from hard selection at mid-to-high budgets, suggesting improved retention of dispersed details common in long scientific talks. The consistent separation on both ROUGE-Lsum and transcript fact recall further indicates that the gains reflect improved coverage rather than metric-specific tuning. Together, these trends support our hypothesis that soft attention steering reduces information loss under fixed-budget inference.

4.4 Scalability

Zero-shot generalization Table 4 reports a zero-shot transfer from SumMe to MR.HiSum without adaptation. HASremains competitive under this distribution shift and improves both Kendall's τ and Spearman's ρ, suggesting more stable global importance ordering rather than overfitting to training-set idiosyn-

Table 4: Zero-shot evaluation on MR.HiSum. Following LLMVS [24], models are trained on SumMe and directly evaluated on a chosen subset of 50 MR.HiSum videos. Higher is better.

Method	τ	ρ
VASNet [9]	0.364	0.364
PGL-SUM [1]	0.375	0.375
DSNet-AB [52]	0.362	0.362
DSNet-AF [52]	0.342	0.342
LLMVS [24]	0.440	0.440
HAS (ours)	**0.45**	**0.45**

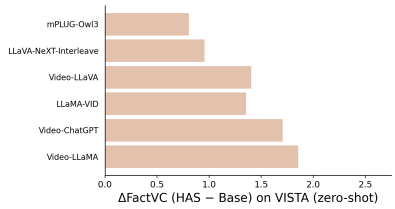


Fig. 3: Gains across open-source MLLMs. The improvement (Averaged) in groundedness on VISTA after adding HAS to all selected open-source video MLLM backbone. Positive values indicate that HAS increases factual consistency.

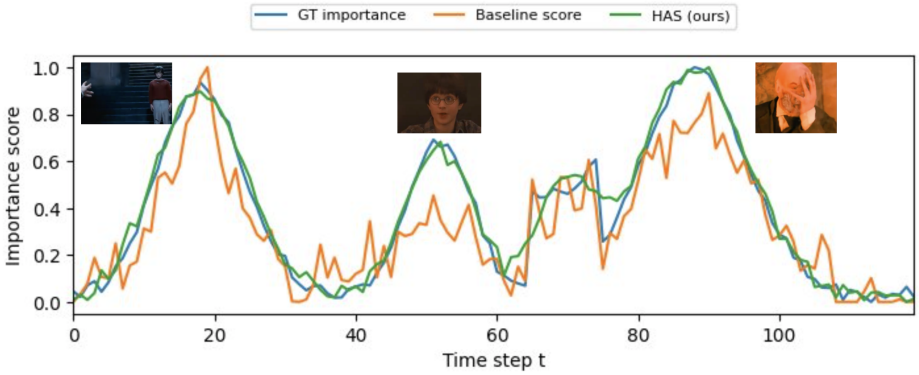
crasies. These results indicate that calibrated temporal attention steering is robust and practical when retraining on the target domain is infeasible.

Performance Across Different MLLMs. Fig. 3 shows that HAS yields consistent positive ΔFactVC across a diverse set of open-source video MLLMs under the same zero-shot protocol. This directly addresses the single-backbone concern: the gains are not tied to a particular architecture or training recipe, but stem from the inference-time steering mechanism. Moreover, the variation in Δ across backbones suggests that HAS acts as a complementary control layer—often providing larger benefits when the base model is more prone to missing or misusing evidence—while remaining beneficial even for stronger interleaving-style backbones. Overall, these results support HAS as a plug-and-play module that improves grounded long-form summarization in a backbone-agnostic manner.

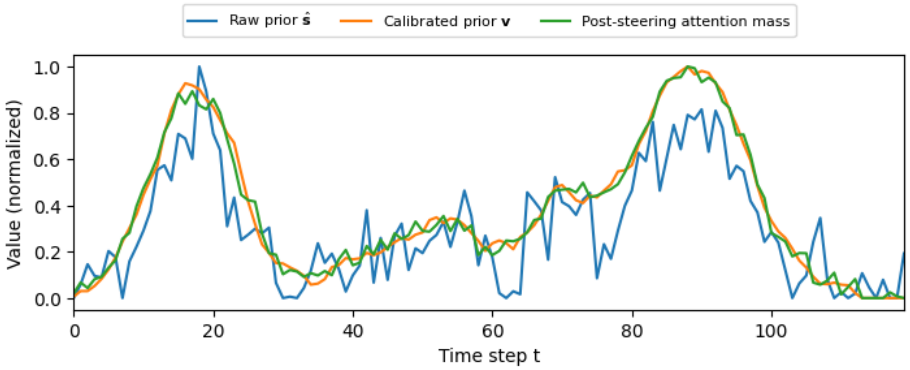
4.5 Controllable Highlight Distribution Visualization given Queries

Qualitative importance curves. Fig. 4(A) compares the predicted importance trajectory with ground-truth (GT) scores. Compared to the baseline, HASaligns salient peaks more accurately in both timing and duration while suppressing spurious oscillations on less important segments, consistent with the higher Kendall's τ and Spearman's ρ in Table 1. This suggests our steering reshapes temporal salience allocation rather than applying a global re-scaling.

Mechanism from prior calibration to attention steering. Fig. 4(B) visualizes the internal signals of HAS. Calibration converts the noisy raw highlight prior into a temporally coherent distribution $\mathbf{v}$, and the post-steering attention mass closely follows $\mathbf{v}$, indicating effective injection of the prompt-conditioned control signal into inference-time attention. Unlike discrete extraction (score→select→summarize), which discards low-scored segments and creates an irreversible information bottleneck, HASretains full context and reduces



(A) Qualitative importance curves



(B) Mechanistic visualization of HAS

Fig. 4: Qualitative and mechanistic visualizations of HAS

conversion loss from prompt intent to evidence usage, making prompt-relevant details more accessible during generation.

5 Conclusion

In this work, we presented **HAS**, a highlight-guided attention steering framework for multimodal LLM video summarization. Unlike prior pipelines that rely on *discrete* highlight selection, HAS turns query-conditioned highlight scores into a *continuous* temporal prior and keeps the video input intact. At inference time, HAS aligns this prior to visual tokens and injects it into a small set of visual cross-attention heads to smoothly bias computation toward salient moments while avoiding the loss of low-scored but useful evidence. Experiments show consistent gains over strong baselines, improving temporal salience agreement, cross-modal consistency, and grounded long-form summarization across datasets.

6 Societal Impact

On the positive side, HAS can improve the accessibility of long videos by helping users navigate, retrieve, and summarize video content more efficiently. This may be useful for educational videos, scientific presentations, and other information-dense long-form media, where grounded summarization can reduce the burden of reviewing the entire video manually. More generally, improving evidence-aware video summarization may support better human access to large-scale video archives.

At the same time, summaries are inherently lossy. If the highlight prior, user query, or underlying MLLM is biased or incomplete, the generated summary may over-emphasize certain moments while omitting important context. As a result, the system could produce partial or misleading narratives, especially in high-stakes settings where factual completeness matters. Because HAS is query-conditioned, malicious or leading prompts could also be used to generate selectively framed summaries that reinforce a desired interpretation of the video.

7 Future Work

First, future work can explore stronger prompt-conditioned highlight priors. Although our calibration step smooths raw highlight scores and improves their stability, the quality of the steering signal still depends on the reliability of the initial highlight estimate. More adaptive prior-generation strategies, uncertainty-aware calibration, or jointly learned saliency estimators may further improve robustness under out-of-domain queries, weak clip descriptions, and videos with temporally dispersed evidence.

Second, an interesting direction is to combine our coarse continuous steering prior with finer temporal localization mechanisms. The current design intentionally avoids hard keyframe selection and instead provides stable soft guidance over the full video. Future extensions could use a coarse-to-fine policy, where continuous attention steering preserves global context while a secondary module handles abrupt event transitions, short salient moments, or tasks requiring precise temporal boundaries.

Third, future work can further study how inference-time steering interacts with different video MLLM backbones. Our results show that HAS can improve groundedness without updating the backbone, but the final summary quality is still influenced by the model's visual understanding ability, context budget, and instruction-following behavior. A broader study over stronger open-source and fully fine-tuned video MLLMs may clarify when steering provides the largest benefit and how it can complement model-scale improvements.

Acknowledgments

This research was supported in part by the National Science Foundation (NSF) SaTC-2426299 and SaTC-2413046.

References

1. Apostolidis, E., Balaouras, G., Mezaris, V., Patras, I.: Combining global and local attention with positional encoding for video summarization. In: Proceedings of the IEEE International Symposium on Multimedia (ISM) (2021)
2. Argaw, D.M., Yoon, S., Caba Heilbron, F., Deilamsalehy, H., Bui, T., Wang, Z., Dernoncourt, F., Chung, J.S.: Scaling up video summarization pretraining with large language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024),
3. Bae, K., Kim, J., Lee, S., Lee, S., Lee, G., Choi, J.: Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13744–13753 (2025)
4. Bargal, S.A., Zunino, A., Kim, D., Zhang, J., Murino, V., Sclaroff, S.: Excitation backprop for rnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1440–1449 (2018)
5. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the "video" in video-language understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2917–2927 (2022)
6. Chen, G., Huang, Y., Xu, J., Pei, B., Wang, J., Chen, Z., Li, Z., Lu, T., Wang, L.: Video mamba suite: State space model as a versatile alternative for video understanding. Int. J. Comput. Vis. **134**(1), 20 (2026). ,
7. Chu, R., Zhao, B., Jiang, H., Aeron, S., Lao, Y.: Bam-icl: Causal hijacking in-context learning with budgeted adversarial manipulation. Advances in Neural Information Processing Systems **38**, 14152–14183 (2025)
8. Chu, R., Zhao, B., Le, T.Q.H., Hoang, D.C., Lin, H., Li, P., Zhao, W., Doan, K.D., Lao, Y.: Debiasrag: A tuning-free path to fair generation in large language models through retrieval-augmented generation (2026),
9. Fajtl, J., Sadeghi Sokeh, H., Argyriou, V., Monekosso, D., Remagnino, P.: Summarizing videos with attention. In: Computer Vision – ACCV 2018 Workshops. Lecture Notes in Computer Science, vol. 11367, pp. 39–54. Springer (2019)
10. Fan, S., Guo, M.H., Yang, S.: Agentic keyframe search for video question answering. arXiv preprint arXiv:2503.16032 (2025).
11. Feng, X., Zou, K., Cen, C., Huang, T., Guo, H., Huang, Z., Zhao, Y., Zhang, M., Zheng, Z., Wang, D., Zou, Y., Li, D.: Linkto-anime: A 2d animation optical flow dataset from 3d model rendering. CoRR **abs/2506.02733** (2025). ,
12. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2950–2958 (2019)
13. Fu, Z., Avaliani, A., Donato, M.: Heterogeneous memory integration and optimization for energy-efficient multi-task nlp edge inference. In: Proceedings of the 29th ACM/IEEE International Symposium on Low Power Electronics and Design. p. 1–6. ISLPED '24, Association for Computing Machinery, New York, NY, USA (2024). ,
14. Ghauri, J.A., Hakimov, S., Ewerth, R.: Supervised video summarization via multiple feature sets with parallel attention. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) (2021)
15. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol. 8695, pp. 505–520. Springer (2014)

16. Han, Y., Zhao, B., Chu, R., Luo, F., Sikdar, B., Lao, Y.: Uibdiffusion: Universal imperceptible backdoor attack for diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 19186–19196. Computer Vision Foundation / IEEE (2025). ,
17. Hazarika, D., Namazifar, M., Hakkani-Tür, D.: Attention biasing and context augmentation for zero-shot control of encoder-decoder transformers for natural language generation. Proceedings of the AAAI Conference on Artificial Intelligence **36**(10), 10738–10748 (2022).
18. He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., Wang, Z.: Align and attend: Multimodal summarization with dual contrastive losses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
19. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., Chilimbi, T.: M-llm based video frame selection for efficient video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13702–13712 (2025)
20. Hua, H., Tang, Y., Xu, C., Luo, J.: V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 3599–3607. AAAI Press (2025). ,
21. Huang, Z., Long, R., Chen, H., Wu, M., Li, Q., Na, J.: LLMQR: efficient knowledge distillation for questionnaire recommendation via large language models. Expert Syst. Appl. **303**, 130632 (2026). ,
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015),
23. Konen, K., Jentzsch, S., Diallo, D., Schütt, P., Bensch, O., El Baff, R., Opitz, D., Hecking, T.: Style vectors for steering generative large language models. In: Findings of the Association for Computational Linguistics: EACL 2024. pp. 782–802. Association for Computational Linguistics, St. Julian's, Malta (March 2024). ,
24. Lee, M.J., Gong, D., Cho, M.: Video summarization with large language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 18981–18991. Computer Vision Foundation / IEEE (2025). ,
25. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: Advances in Neural Information Processing Systems. vol. 34 (2021),
26. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models (2024),
27. Liao, P., Lee, H., Wang, H.: Cross-attention reprogramming for ASR: bridging discrete speech units and pretrained language models. IEEE Access **14**, 662–678 (2026). ,
28. Lin, J., Hua, H., Chen, M., Li, Y., Hsiao, J., Ho, C., Luo, J.: Videoxum: Cross-modal visual and textural summarization of videos. IEEE Transactions on Multimedia (2024).
29. Liu, D., Whitehouse, C., Yu, X., Mahon, L., Saxena, R., Zhao, Z., Qiu, Y., Lapata, M., Demberg, V.: What is that talk about? A video-to-text summarization dataset

for scientific presentations. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025. pp. 6187–6210. Association for Computational Linguistics (2025),

30. Ma, F., Jin, X., Wang, H., Xian, Y., Feng, J., Yang, Y.: Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13151–13160 (2024)
31. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-ChatGPT: Towards detailed video understanding via large vision and language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). ,
32. Meng, N., Manicke, C., Sahu, R., Ding, C., Lao, Y.: Advancing adversarial robustness in gnerfs: The il2-nerf attack. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 16388–16397. Computer Vision Foundation / IEEE (2025). ,
33. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23023–23033 (June 2023),
34. Narasimhan, M., Rohrbach, A., Darrell, T.: Clip-it! language-guided video summarization. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 13988–14000 (2021),
35. Pang, Z., Nakashima, Y., Otani, M., Nagahara, H.: Unleashing the power of contrastive learning for zero-shot video summarization. Journal of Imaging **10**(9), 229 (2024).
36. Park, S., Chung, K.: Spinout: Enhanced rotation-based quantization for LLM by outlier injection. IEEE Access **14**, 24082–24095 (2026). ,
37. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
38. Son, J., Park, J., Kim, K.: Csta: Cnn-based spatiotemporal attention for video summarization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
39. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5179–5187 (2015)
40. Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., Nushi, B.: Improving instruction-following in language models through activation steering. In: International Conference on Learning Representations (ICLR) (2025),
41. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 3104–3112 (2014),

42. Tang, F., Liu, C., Xu, Z., Hu, M., Huang, Z., Xue, H., Chen, Z., Peng, Z., Yang, Z., Zhou, S., Li, W., Li, Y., Song, W., Su, S., Feng, W., Su, J., Lin, M., Peng, Y., Cheng, X., Razzak, I., Ge, Z.: Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26147–26159 (June 2025),
43. de la Torre, P.G., Pérez-Verdugo, M., Barandiaran, X.E.: Attention is all they need: cognitive science and the (techno)political economy of attention in humans and machines. AI Soc. **41**(1), 5–21 (2026). ,
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017),
45. Wang, A., Zhang, K., Parhi, K.K., Lao, Y.: Hermes: Homomorphic encryption over residual number system for multi-level evaluations. In: Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design. ICCAD '24, Association for Computing Machinery, New York, NY, USA (2025). ,
46. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 982–990. IEEE Computer Society (2016). ,
47. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models (2024),
48. Yu, Q., Wang, Z., Wei, G., Yu, H.: Deep learning for video summarization: Systematic review, challenges and opportunities. IEEE CAA J. Autom. Sinica **13**(1), 21–42 (2026). ,
49. Zhang, H., Shi, P., Yang, G.: Tdtoon: Two-stage diffusion for controllable cartoon video generation via sketch enhancement. Displays **91**, 103269 (2026). ,
50. Zhang, K., Wang, A., Parhi, K.K., Lao, Y.: Hardware acceleration for fully homomorphic encryption scheme switching from ckks to fhew. In: 2024 58th Asilomar Conference on Signals, Systems, and Computers. pp. 1792–1796 (2024).
51. Zhao, C., Ding, G., Wang, W., Yang, Z., Liu, Z., Chen, H., Shen, C.: Freercustom: Training-free multi-concept customization for image and video generation. Int. J. Comput. Vis. **134**(1), 17 (2026). ,
52. Zhu, W., Lu, J., Li, J., Zhou, J.: Dsnet: A flexible detect-to-summarize network for video summarization. IEEE Transactions on Image Processing **30**, 948–962 (2021)