



SEASON: Mitigating Temporal Hallucination in Video Large Language Models via Self-Diagnostic Contrastive Decoding

Chang-Hsun Wu^{1,†}, Kai-Po Chang¹, Yu-Yang Sheng¹,
Hung-Kai Chung¹, Kuei-Chun Wang¹, and Yu-Chiang Frank Wang^{1,2,‡}

¹ Graduate Institute of Communication Engineering, National Taiwan University ² NVIDIA

[†]r14942083@ntu.edu.tw, [‡]frankwang@nvidia.com

Abstract

Video Large Language Models (VideoLLMs) have shown remarkable progress in video understanding. However, these models still struggle to effectively perceive and exploit rich temporal information in videos when responding to user queries. Therefore, they often generate descriptions of events that are temporally inconsistent or causally implausible, causing severe hallucination issues. While most prior studies have focused on spatial hallucinations (e.g. object mismatches), temporal reasoning in video understanding remains relatively underexplored. To address this issue, we propose **Self-Diagnostic Contrastive Decoding (SEASON)**, a training-free method that adaptively enhances temporal and spatial faithfulness for each output token. It achieves this by dynamically diagnosing each token’s hallucination tendency and applying adaptive contrastive decoding against its corresponding temporal and spatial negatives. Extensive experiments demonstrate that SEASON outperforms all existing training-free hallucination mitigation approaches on three hallucination examination benchmarks, while further improves VideoLLMs across four general video understanding benchmarks. The code will be released upon acceptance.

1. Introduction

Multimodal large language models (MLLMs) [1, 20, 24, 25] have recently achieved remarkable success in bridging vision and language, enabling unified reasoning across tasks such as visual captioning [28, 30], visual question answering [2, 32], and video understanding [5, 15]. However, these models remain prone to produce the textual content that is often inconsistent with the given visual evidence, which causes severe hallucination issues [10, 18]. This presents serious risks for the model deployment in critical applications demanding high-standard reliability and trustworthiness, such as healthcare robots and self-driving vehicles.

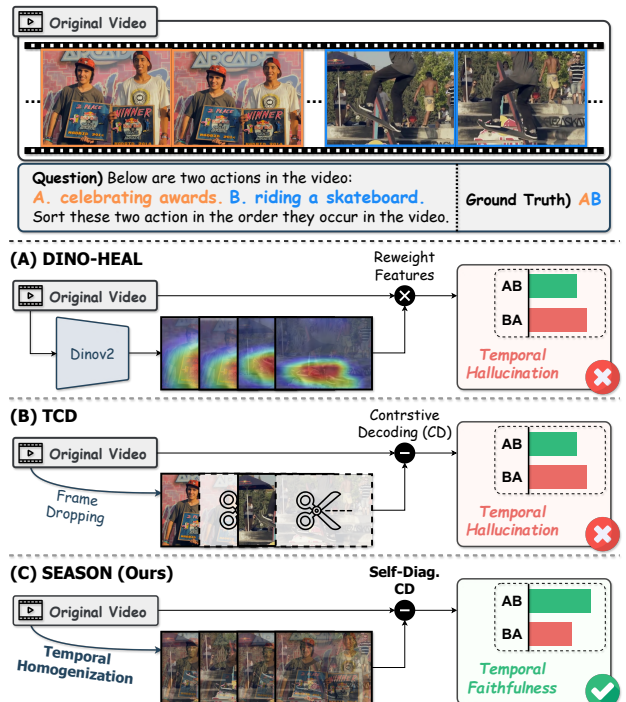


Figure 1. Suppressing hallucination in video LLMs. (A) DINO-HEAL [14] exploits spatial saliency but misses temporal order, (B) TCD [33] contrasts frame-dropped videos but ignores causal relations, and (C) our SEASON achieves temporal faithfulness for each output token.

As a result, alleviating hallucinations in multimodal models has become an important research focus across academia and industry.

To deal with this challenge, early works [11, 12, 27, 35] have primarily focused on mitigating hallucinations in image understanding tasks, where hallucinations often appear as spatial inconsistencies, such as describing nonexistent objects or incorrect attributes [12]. We refer to this issue as *spatial hallucination*. To handle this issue, they aim to enforce the output description exempted from spurious correlation caused by language prior, via contrasting output

distributions derived from original and distorted visual inputs [11, 27]. Although these approaches are effective when the visual input is a *static* image. However, directly extending them to video large language models is not sufficient, because videos introduce rich temporal structure. Despite reduced spatial hallucinations, the model would still misunderstand event causality and thus produce descriptions that are temporally inconsistent with visual content [14, 26]. This issue is known as *temporal hallucination*, which remains a key obstacle for reliable video understanding.

To mitigate temporal alongside spatial hallucinations in video, multiple benchmarks [14, 26, 33] have been established. Building upon these, recent approaches have extended the ideas from image hallucination mitigation to video domain, generally following two research lines. Training-based methods [17, 23, 29] exhibits improved temporal faithfulness via reinforcement learning [29] or preference optimization [17, 23], but require both expensive re-training and high-quality preference data. This motivates the pursuit of training-free approaches that bypass these cost and could be easily applied to different models during inference. However, existing training-free approaches [14, 33] such as DINO-HEAL [14] and TCD [33] still struggle to understand the temporal causality, as illustrated in Fig. 1. As a result, enabling VideoLLMs without training cost to produce descriptions that are especially temporally faithful, together with spatial fidelity, remains a challenging open problem.

In this paper, we propose *Self-Diagnostic Contrastive Decoding (SEASON)*, a training-free method that enhances both temporal and spatial faithfulness in VideoLLMs. To specifically achieve temporal faithfulness while dynamically assessing which types of hallucination each token may suffer, we propose *temporal homogenization* to construct temporally-negative video to expose the spurious temporal correlations lying within VideoLLMs, and a *self-diagnostic mechanism* to identify each token’s hallucination tendency based on its preceding context. More specifically, the former produces a temporal negative that is temporally incoherent yet remains spatially consistent. The latter identifies the potential hallucination type of the current token by analyzing the frame-attention divergences between the original video and these specialized negatives (temporal and spatial). By diagnosing the hallucination tendency of each token and contrasting with an appropriate negative, *SEASON* enables VideoLLMs to produce textual responses that are temporally faithful while maintaining spatial fidelity, without any additional training cost.

In summary, our contributions are as follows:

- We present *SEASON*, a self-diagnostic contrastive decoding framework that adaptively enhances temporal and spatial faithfulness for each output token in a training-free manner.

- *SEASON* introduce temporally-homogenized videos as negatives that amplify spurious temporal correlations, yielding a temporally hallucination-focused distribution for contrastive decoding.
- We develop the hallucination diagnostician that estimates token-level hallucination tendency via frame-attention divergence and assigns a corresponding contrastive penalty.

2. Related Work

2.1. Mitigating Spatial Hallucination in Visual Large Language Models

Visual Large Language Models (VLLMs) [1, 20, 24, 25] have demonstrated remarkable capabilities in vision-language tasks. However, VLLMs often suffer from spatial hallucination, generating spatially inconsistent descriptions due to over-reliance on language priors rather than actual visual evidence [10, 18]. To mitigate these errors, research has explored two directions. Training-based approaches improve visual-language alignment through higher-quality datasets [19] or human feedback [31]. While effective, these methods require large-scaled annotated datasets and costly re-training.

On the other hand, training-free approaches [11, 12, 27, 35] modify the decoding process of MLLMs without updating parameters. For example, VCD [12] reduces hallucination by counteracting language priors, while MARINE [35] leverages auxiliary guidance signals to correct false information. Although efficient and widely applicable, these methods for hallucination mitigation are designed for a static image and thus cannot capture temporal dependencies within video. As directly extending to video domain, they struggle to capture temporal dynamics in video and often encounter the challenge of temporal hallucination [14, 26].

2.2. Mitigating Temporal Hallucination in Video Large Language Models

The issue of hallucinations in Video Large Language Models (VideoLLMs) has emerged as a critical research area, catalyzing the development of dedicated benchmarks [14, 26, 33]. Building upon these, several works including training-based and training-free have been proposed to advance this research area. Training-based methods [3, 6, 8, 17, 23, 29] primarily employ reinforcement learning [29], preference optimization [8, 17, 23], or pre-training from scratch [3]. For example, ArrowRL [29] encourages divergent interpretations between forward and reversed videos, while RRPO [23] utilizes sub-sequence-level refined rewards and a token-wise regularizer. Despite promising results, they require costly retraining, auxiliary reward models, and high-quality preference data, which limits their scalability and model-agnostic deployment.

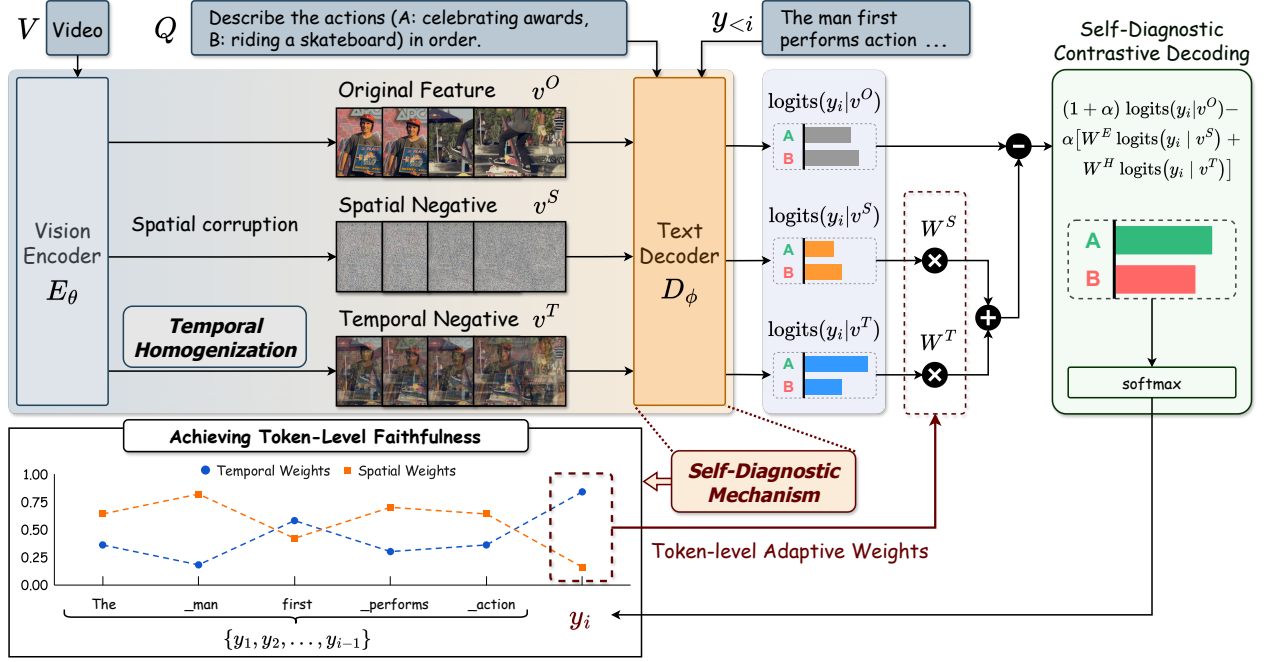


Figure 2. **Overview of SEASON.** Given the input video (V) and the question (Q), our proposed *SEASON* contrasts the original video representations (v^O) against our introduced spatial (v^S) and temporal (v^T) negatives to jointly achieve temporal and spatial faithfulness. Specifically, we design v^T via the proposed “Temporal Homogenization”, focusing on introducing temporal ambiguity while preserving spatial semantics. The “Self-Diagnostic Mechanism” computes token-level adaptive weights (W^S, W^T) by measuring attention divergence, dynamically steering the final decoding to penalize spatial or temporal hallucinations.

In contrast, training-free methods bypass these costs and be able to applied to various models at inference time, offering an attractive alternative. For example, DINO-HEAL [14] leverages saliency maps from DINOv2 [22] to re-weight visual features for improving object motion understanding, while TCD [33] contrasts token predictions between original and frame-dropped videos. Nevertheless, these methods still struggle to understand complex temporal relationships, notably event causality. To address this issue, we propose *SEASON* to employ temporally-homogenized video as contrasted negatives, encouraging VideoLLM to understand the temporal relationship within video to achieve temporal faithfulness.

3. Method

3.1. Problem Formulation

Given an input video $V = \{f_1, f_2, \dots, f_{|V|}\}$ consisting of $|V|$ frames and an associated textual query Q , a VideoLLM (parameterized by vision encoder E_θ and text decoder D_ϕ) aims to generate a textual response $y = \{y_1, \dots, y_N\}$ that accurately answers queries about the visual content. A hallucination occurs when the generated response in y is not presented in the video V . Formally, we define a **spatial hallucination** as objects or attributes described in y that are visually absent within individual frames f_i , and a **temporal hallucination** as the descriptions in y that contradict the actual temporal structure (e.g., event order and causality)

presented in video V .

Therefore, we propose *SEASON*, a self-diagnostic contrastive decoding approach for adaptively suppressing temporal and spatial hallucinations during inference. As illustrated in Fig. 2, *SEASON* first mitigates the temporal hallucination by providing a strong temporal negative signal to contrast (Sec. 3.2). Furthermore, we introduce a self-diagnosing contrastive decoding strategy that diagnoses each token’s hallucination tendency and then applies an adaptive contrastive penalty against the corresponding (temporal or spatial) negatives (Sec. 3.3).

3.2. Mitigating Temporal Hallucination via Temporal Homogenization

To mitigate temporal hallucination, we aim to explicitly expose and penalize a VideoLLM’s reliance on spurious temporal correlations. To achieve this, we introduce *temporal homogenization*, a novel augmentation that constructs negatives which are temporally incoherent yet spatially consistent. We specifically focus on this design because simply following Leng et al. [12] by adding Gaussian noise to video frames (which produces spatial negative s^v , Fig. 2) corrupts both spatial and temporal structures simultaneously. This yields a *temporally-easy* negative, as it provides suboptimal contrast by allowing the model to reject this negative based on the obvious spatial corruption rather than focusing on the intended temporal inconsistency.

In contrast, our temporal negative v^T (Fig. 2) maxi-

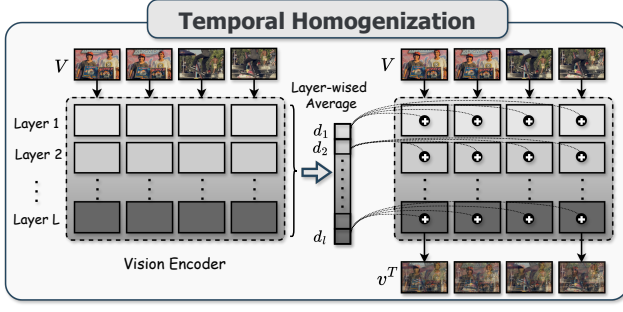


Figure 3. Illustration of Temporal Homogenization. This constructs the temporal negative v^T by computing a layer-wise average of frame features ($d_1, \dots, d_L$) and progressively re-injecting this global context back into each frame’s representation within the vision encoder. The resulting representation would be temporally ambiguous while preserving per-frame structure information.

mally preserves spatial fidelity, which ensures the resulting contrastive signal reflects only temporal inconsistencies, yielding a *temporally-hard* negative. This is achieved by temporally aggregating all the frame features in a layer-wise manner and respectively re-injecting them back into each frame’s representation, neutralizing temporal variation while retaining spatial semantics. See the discussion below.

Temporal Homogenization for Hallucination Exposure.

To expose the VideoLLM’s potential temporal hallucination, we propose the “temporal homogenization” to produce the temporal negatives v^T . v^T is an augmented video representation designed to be temporally incoherent yet spatially consistent and be constructed in Fig. 3. The core idea to achieve this is to homogenize the temporal information across the frames in a layer-wise manner while keeping per-frame spatial structure intact.

Specifically, we define the temporal negatives v^T as the set of final-layer homogenized frame features from vision encoder (E_θ),

$$v^T = \{h_{L,t}\}_{t=1}^{|V|}. \quad (1)$$

Each $h_{L,t}$ is obtained through a progressive and layer-wise homogenization process. At every layer l and frame t , the homogenized feature $h_{l,t}$ is defined as a linear combination of the frame feature from the corresponding global context d_l and the pre-homogenization feature $h'_{l,t}$. This feature $h'_{l,t}$ is computed from the previous layer’s output $h_{l-1,t}$, which has already been recurrently mixed with global context from preceding layers.

$$h_{l,t} = (1 - \beta)h'_{l,t} + \beta d_l, \text{ where } h'_{l,t} = E_\theta^{(l)}(h_{l-1,t}). \quad (2)$$

Here, d_l denotes the mean of the frame features in l -th layer pre-computed from a standard forward pass on V ($d_l = \frac{1}{|V|} \sum_{t=1}^{|V|} h'_{l,t}$), $\beta \in [0, 1]$ is a hyperparameter used to control the degree of temporal homogenization, and $h_{0,t}$ are the patch embeddings of frame f_t .

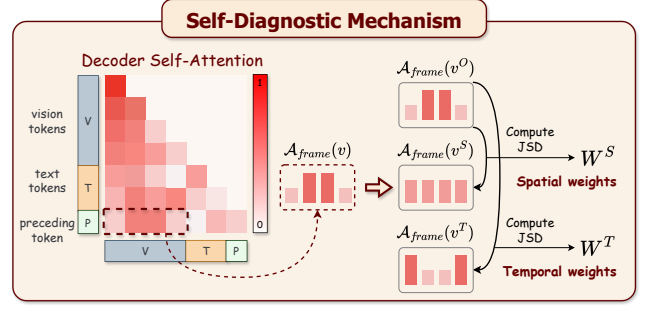


Figure 4. Illustration of the Self-Diagnostic Mechanism. This process extracts the frame-level attention distribution (A_{frame}) from the preceding token. It computes JSD divergence between the attention distributions of the original video (v^O) and the negatives (v^S, v^T), outputting the adaptive spatial (W^S) and temporal (W^T) diagnostic weights to penalize spatial or temporal hallucination for each output token.

Mitigating Temporal Hallucination via Contrastive Decoding.

With the obtained temporal negatives v^T , we aim to contrast the temporally hallucinated output distribution induced by v^T against those from the original v^O . This enable us to achieve the temporal faithfulness. Thus, we impose a visual contrastive decoding to eliminate the temporal priors lying in the text decoder (D_ϕ). Formally, given a textual query Q and the video representations v^O (original) and v^T (temporal negative), the contrastive distribution p_{SEASON^T} is formulated as:

$$p_{SEASON^T}(y_i) = \text{softmax}[(1 + \alpha) \text{logits}(y_i|v^O, Q, y_{<i}) - \alpha \text{logits}(y_i|v^T, Q, y_{<i})], \quad (3)$$

where α controls the contrastive strength ($\alpha = 0$ reduces to regular decoding). The resulting outputs y_i are generated from a distribution explicitly purified of temporal hallucination.

Building on the formulation (Eq. 3) for achieving temporal faithfulness, we next introduce our full self-diagnostic contrastive decoding strategy, a unified framework that also mitigates spatial hallucination by incorporating the spatial negatives. The effectiveness of our core temporal negatives, v^T , will also be experimentally verified in Table 3.

3.3. Achieving Token-Level Faithfulness via Self-Diagnostic Contrastive Decoding

In Sec. 3.2, the VideoLLM is able to mitigate temporal hallucination via contrastive decoding. To jointly mitigate spatial hallucination together, we propose a “self-diagnostic mechanism” to assess the risk of each token being temporally or spatially hallucinated. Our insight is that the generation of a token is highly dependent on its preceding context, which serves as an indicator to reflect the potential hallucination tendencies. To achieve our goal, we interpret the frame-level attention divergence between preceding text token (y_{i-1}) and video representation ($v \in [v^O, v^S, v^T]$) as

a measure of the current token (y_i)’s reliance on temporal or spatial cues. For this interpretation, the required spatial negative (referenced in Fig. 2) is created by adding the Gaussian noise to the original video representation v^O . We now explain this mechanism and how it is integrated with contrastive decoding below:

Detecting Hallucination Tendency via Self-Diagnostic Mechanism. As illustrated in Fig. 4, the self-diagnostic mechanism identifies each token’s hallucination tendency by observing how the attention pattern of its preceding token changes when temporal cues are removed (i.e., the changes between original video and temporal negatives, v^O and v^T). The core idea is that tokens that rely on temporal consistency will exhibit strong attention shifts when the video is temporally homogenized, whereas tokens grounded in static objects remain unaffected. Consequently, the divergence between the frame-level attention distributions of the original video representations, spatial, and temporal negatives (denoted as v^O , v^S , and v^T , in Fig. 2, respectively) serves as a self-diagnostic signal, indicating which types of hallucination (temporal or spatial) a token is prone to and how penalties should be allocated during decoding.

To quantify this behavior, we derive the frame-level attention distribution $\mathcal{A}_{frame}(v)$ from the text decoder (D_ϕ)’s multi-head attention. Specifically, $\mathcal{A}_{frame}(v)$ is denoted as the normalized attention that the preceding token y_{i-1} assigns to each video frame, calculated by giving the attention matrix of layer j (denoted as $A_j \in \mathbb{R}^{n,n}$) obtained by summing over all heads, preceding text token y_{i-1} , and $v_{t,k}$ (the k -th visual token extracted from frame f_t in the video representation v ($v \in [v^O, v^S, v^T]$)). This is formulated as:

$$\mathcal{A}_{frame}(v) = \text{softmax}_t \left[\sum_k \left(\sum_{j \in J} A_j \right) (y_{i-1}, v_{t,k}) \right]. \quad (4)$$

To diagnose the nature of current token (y_i)’s hallucination tendency, we compare the $\mathcal{A}_{frame}(v)$ scores computed on original video representations, spatial, and temporal negatives (v^O , v^S , and v^T). Specifically, we compute the divergence between these frame-level attention distributions ($\mathcal{A}_{frame}(v^O)$ vs. $\mathcal{A}_{frame}(v^S)$ and $\mathcal{A}_{frame}(v^O)$ vs. $\mathcal{A}_{frame}(v^T)$) to determine which type of hallucination the current token is prone to. Formally, let w_S and w_T denote the tendency to be spatially or temporally hallucinated, which are calculated via Jensen-Shannon divergences (JSD) to measure the relative degree of attention shift as follows:

$$\begin{aligned} w_S &= \frac{D_S}{D_S + D_T}, \quad w_T = \frac{D_T}{D_S + D_T}, \\ D_S &= \text{JSD}(\mathcal{A}_{frame}(v^O), \mathcal{A}_{frame}(v^S)), \\ D_T &= \text{JSD}(\mathcal{A}_{frame}(v^O), \mathcal{A}_{frame}(v^T)). \end{aligned} \quad (5)$$

Here, D_S and D_T measure the relative degree of attention

shift. A larger D_T signifies a temporal hallucination tendency (as the token relies heavily on temporal cues), while a larger D_S signifies a spatial hallucination tendency.

By this method, the derived adaptive weights (w_S, w_T) thus serve as the core diagnostic signals, quantifying the token’s specific hallucination tendency and establishing the basis for the subsequent contrastive decoding stage.

Self-Diagnostic Contrastive Decoding. To mitigate temporal alongside spatial hallucinations of VideoLLM without any further training cost, we integrate the self-diagnostic mechanism into contrastive decoding. The adaptive diagnostic weights (w_S, w_T) derived above are thus incorporated to dynamically balance the contrastive penalties at each generation step.

Given the current decoding step i , we first obtain three logit distributions by feeding the textual context ($y_{<i}, q$) to the text decoder D_ϕ conditioned on the original video representation v^O , spatial negative v^S , and temporal negative v^T . For brevity, we denote these logits distributions as $\text{logits}(y_i|v^O)$, $\text{logits}(y_i|v^S)$, and $\text{logits}(y_i|v^T)$, respectively. Therefore, the final self-diagnostic contrastive decoding distribution p_{SEASON} that adaptively suppresses both spatial and temporal hallucinations is formulated, by combining these logits using the diagnostic weights:

$$\begin{aligned} p_{SEASON}(y_i) &= \text{softmax} \left[(1 + \alpha) \text{logits}(y_i|v^O) \right. \\ &\quad \left. - \alpha [w_S \text{logits}(y_i|v^S) + w_T \text{logits}(y_i|v^T)] \right]. \end{aligned} \quad (6)$$

This formulation acts as logit-space contrastive decoding, where the adaptive weights (w_S, w_T) determine the penalty for potential spatial or temporal hallucination tendency for each token. A larger w_T suppresses potential temporal hallucination, while a larger w_S penalizes possible spatial hallucination.

By dynamically steering the decoding direction in this manner, the model achieves per-token self-assessment and adaptively corrects potential hallucinations, ensuring both temporal and spatial faithfulness throughout the generation process.

4. Experiments

4.1. Experimental Setup

Benchmarks and Metrics. To directly assess the reduction of spatial and temporal hallucinations, we evaluate *SEASON* on three dedicated video hallucination examination benchmarks: VidHalluc [14], VideoHalluc [26], and EventHalluc [33]. To verify that our method preserves general video understanding, we further evaluate performance on two temporal understanding benchmarks: TempCompass [21] and TVBench [7], along with two conventional video understanding benchmarks: VideoMME [9]

Table 1. Evaluation of multiple hallucination examination benchmarks with different VideoLLMs as backbones. **Bold** marks the best per group; highlights indicate the top two benchmark results.

Models	Training-free	VidHalluc					VideoHalluc					EventHallusion	
		BQA	MCQ	STH	TSH	AVG	ORH	TPH	SDH	EFH	ENFH	AVG	AVG
LLaVA-OV-7B [13]	-	74.36	90.27	63.65	53.00	70.32	56.50	52.50	56.50	15.00	51.50	46.40	60.15
+TCD [33]	✓	72.44	90.06	58.57	64.33	71.35	59.50	53.50	56.00	17.50	54.00	48.10	68.46
+DINO-HEAL [14]	✓	74.29	90.36	63.18	53.00	70.21	57.00	53.50	56.50	15.50	52.00	46.90	60.15
+SEASON (Ours)	✓	73.15	90.51	60.29	77.50	75.36	63.00	55.50	56.50	19.50	48.00	48.50	69.19
QWEN2.5-VL-7B [4]	-	75.79	84.05	74.91	59.00	73.44	62.00	46.50	70.50	31.50	55.00	53.10	63.33
+TCD [33]	✓	74.60	85.57	74.37	64.67	74.80	61.00	46.50	70.50	31.00	55.00	52.80	64.79
+DINO-HEAL [14]	✓	75.86	84.21	75.86	58.67	73.65	61.00	46.50	71.50	32.00	55.50	53.30	63.57
+ArrowRL [29]	✗	76.14	87.84	70.92	57.83	73.18	60.50	55.00	68.00	33.50	55.50	54.50	68.95
+SEASON (Ours)	✓	78.08	87.32	71.92	77.67	78.75	63.50	49.50	73.50	31.50	56.00	54.80	66.01
LLaVA-Video-7B [34]	-	75.02	90.76	51.23	38.00	63.75	60.00	61.50	66.50	16.50	52.50	51.40	63.57
+TCD [33]	✓	73.50	90.41	49.64	46.00	64.89	58.00	61.00	65.50	14.50	51.50	50.10	64.30
+DINO-HEAL [14]	✓	75.27	90.81	51.51	37.67	63.81	59.50	61.00	66.50	17.00	52.50	51.30	64.30
+TPO [17]	✗	74.85	90.69	49.62	42.50	64.42	60.00	59.50	68.50	16.00	52.50	51.30	63.33
+RRPO [23]	✗	76.80	91.23	49.83	37.67	63.88	59.00	58.00	67.00	21.00	51.50	51.30	67.97
+SEASON (Ours)	✓	74.71	90.95	49.86	50.33	66.46	60.50	62.00	68.00	18.00	48.50	51.40	66.99

Table 2. Performance comparisons on benchmarks for hallucination examination, temporal, and conventional video understanding. Different VideoLLMs are applied as backbones. **Bold** marks the best per group; highlights indicate the best benchmark results.

Models	Training-free	Hallucination Examination				Temporal Understanding			Conventional Video Understanding		
		VidHalluc	VideoHalluc	EventHallusion	AVG	TempCompass	TVBench	AVG	VideoMME	MVBench	AVG
Proprietary models											
GPT-4o [1]	-	81.2	53.3	91.9	75.5	73.8	39.9	56.9	71.9	49.1	60.5
Gemini 1.5 Pro [24]	-	72.8	37.8	80.4	63.8	67.1	47.6	57.4	75.0	60.5	67.8
LLaVA-OV-7B [13]	-	70.3	46.4	60.2	59.0	68.3	42.8	55.6	50.6	54.4	52.5
+TCD [33]	✓	71.4	48.1	68.5	62.6	68.5	42.9	55.7	50.4	54.3	52.4
+DINO-HEAL [14]	✓	70.2	46.9	60.2	59.1	68.3	42.9	55.6	50.6	54.4	52.5
+SEASON (Ours)	✓	75.4	48.5	69.2	64.3	69.3	43.0	56.1	50.7	54.8	52.7
QWEN2.5-VL-7B [4]	-	73.4	53.1	63.3	63.3	72.9	46.7	59.8	49.9	61.8	55.8
+TCD [33]	✓	74.8	52.8	64.8	64.1	73.4	46.1	59.8	49.4	60.1	54.7
+DINO-HEAL [14]	✓	73.7	53.3	63.6	63.5	73.2	46.9	60.1	49.9	61.9	55.9
+ArrowRL [29]	✗	73.2	54.5	69.0	65.5	72.6	49.4	61.0	49.6	59.8	54.7
+SEASON (Ours)	✓	78.7	54.8	66.0	66.5	73.7	47.7	60.7	50.4	62.3	56.4
LLaVA-Video-7B [34]	-	63.8	51.4	63.6	59.6	69.7	45.2	57.5	53.0	58.4	55.7
+TCD [33]	✓	64.9	50.1	64.3	59.8	70.0	45.2	57.6	53.0	57.5	55.3
+DINO-HEAL [14]	✓	63.8	51.3	64.3	59.8	69.9	45.5	57.7	53.2	58.3	55.8
+TPO [17]	✗	64.4	51.3	63.3	59.7	69.5	45.1	57.3	53.0	58.2	55.6
+RRPO [23]	✗	63.9	51.3	68.0	61.1	70.3	45.2	57.8	53.8	58.1	56.0
+SEASON (Ours)	✓	66.5	51.4	67.0	61.6	71.1	46.4	58.8	53.4	57.9	55.6

and MVBench [16]. Most subtasks are evaluated using QA accuracy following their official protocols. Key exceptions include the STH subtask in VidHalluc, which combines a classification score and a description score. For EventHallusion [33] and TempCompass [21], we follow [6, 29] and reproduce only the deterministic subtasks to avoid reliance on third-party LLM-based evaluators and reduce evaluation ambiguity. Please refer to the Appendix for more experiments and related details.

Base Models and Baselines. We apply our training-free framework to three open-source VideoLLMs (LLaVA-OV-7B [13], Qwen2.5-VL-7B [4], and LLaVA-Video-7B [34]) to demonstrate broad applicability. For comparison, we include two training-free hallucination mitigation methods designed for VideoLLMs, TCD [33] and DINO-HEAL [14]. In addition, as a reference for state-of-the-

art performance, we report results from three training-based methods (ArrowRL [4], TPO [17], and RRPO [23]) that aim to enhance temporal reasoning ability of VideoLLMs. To isolate the contribution of our novel components, we also perform a comparison of each negative in Tab. 5.

Implementation Details. SEASON is applied purely during inference without any retraining or fine-tuning. We utilize 8 frames for inference, and all experiments are conducted on the same model backbones with identical settings across baselines to ensure a fair comparison. For the self-diagnostic mechanism, we select attention layers $J = [20, 21, 22, 23]$ (Eq. 4) based on empirical analysis. Hyperparameters (contrastive strength α and homogenization degree β) are tuned via grid search to systematically explore critical settings. Additional experimental details are provided in the Appendix.

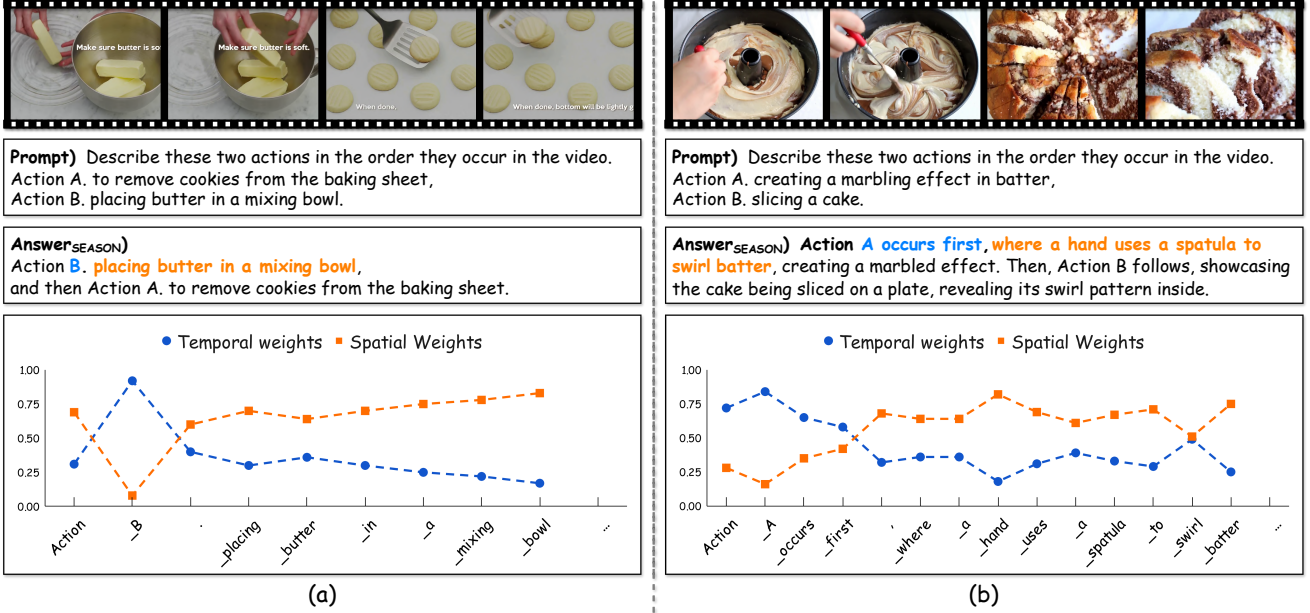


Figure 5. **Qualitative visualization of SEASON’s self-diagnostic mechanism.** Qualitative visualization of SEASON’s self-diagnostic weights (W^T and W^S). In the generated text (the x-axis in the line plot), blue tokens are identified as relying on visual temporal cues; SEASON thus contrasts them against the temporal negative (v^T) to ensure token-level temporal faithfulness. For instance, tokens critical for temporal ordering like “B” (in (a)), as well as “A” and “first” (in (b))) clearly receive high temporal weights (W^T) to ensure the sequence is correct. On the other hand, orange tokens rely on visual spatial cues and are contrasted against the spatial negative (v^S). This is evident as tokens describing objects and interactions, such as “placing butter...mixing bowl” in (a) and “hand...swirl batter” in (b), are assigned high spatial weights (W^S). Both (a) and (b) are samples from Vidhalluc [14].

4.2. Quantitative Evaluation

Hallucination Benchmark Evaluation. As shown in Tab. 1, SEASON consistently achieves the best training-free performance across all three hallucination examination benchmarks for each backbone, while remaining competitive with or even surpassing training-based methods. For example, on QWEN2.5-VL-7B [4], SEASON improves the overall VidHalluc [14] score by +5.3% over the base model and +5.6% over the training-based baseline (ArrowRL [29]). The improvement is particularly pronounced in mitigating temporal hallucinations, which our method is designed to target (Sec. 3.2). For instance, on the TSH sub-task of VidHalluc [14], SEASON boosts performance by up to +24.5%/+18.7%/+12.3% over each backbones.

General Video Understanding Evaluation. Crucially, as shown in Tab. 2, these gains in faithfulness do not come at the cost of general comprehension. SEASON improves performance on the two temporal understanding benchmarks while maintaining performance on the two conventional video understanding benchmarks (e.g., +1.4%/+1.2% on TempCompass [21] and TVBench [7] for LLaVA-Video-7B [34]). This result indicates that our self-diagnostic mechanism (Sec. 3.3) effectively penalizes suspected hallucinations without over-suppressing correct tokens.

Table 3. Ablation of SEASON’s temporal negative (v^T) design. We compare our **homogenized** strategy against alternatives (Average, Shuffled, Reverse) on temporal hallucination examination and temporal understanding benchmarks.

Models	Temporal Hallu.		Temp Compass	TV Bench	AVG
	TSH	TPH			
LLaVA-OV-7B [13]	53.0	52.5	68.3	42.8	54.2
+Average	<u>72.7</u>	<u>55.5</u>	69.3	42.2	<u>59.9</u>
+Shuffled	52.3	54.5	69.7	43.6	55.0
+Reverse	58.3	58.5	69.1	44.3	57.6
+Homogenized (ours)	77.7	<u>55.5</u>	<u>69.5</u>	42.9	61.4
QWEN2.5-VL-7B [4]	59.0	46.5	<u>72.9</u>	46.7	56.3
+Average	<u>74.7</u>	44.0	71.6	43.9	58.6
+Shuffled	68.5	<u>50.0</u>	<u>72.9</u>	47.5	<u>59.7</u>
+Reverse	67.3	51.5	71.6	44.9	58.8
+Homogenized (ours)	79.3	47.0	74.1	47.5	62.0

4.3. Qualitative Evaluation

To illustrate how SEASON mitigates hallucinations, Fig. 5 provides two qualitative examples. We visualize the token-wise adaptive weights computed by our self-diagnostic mechanism (Eq. 5). When the model generates temporal-related words (e.g., “A occurs first” in (b)), the temporal weight w_T (blue) is high, indicating a strong contrast against the temporal negative v^T . In contrast, when generating tokens for objects or static attributes (e.g., “butter” and “mixing bowl” in (a)), the spatial weight w_S (orange) is high, activating a contrast against the spatial negative v^S .

Table 4. Ablation of *SEASON*’s key components (v^S , v^T) across various hallucination examination benchmarks.

Models	VidHalluc	VideoHalluc	EventHalluc	AVG
LLaVA-OV-7B [13]	70.3	46.4	60.2	59.0
+Spatial Neg. (v^S)	75.1	46.7	69.2	63.7
+Temporal Neg. (v^T)	74.2	47.5	69.9	63.9
+SEASON	75.4	48.5	69.2	64.4
QWEN2.5-VL-7B [4]	73.4	53.1	63.3	63.3
+Spatial Neg. (v^S)	77.4	53.4	63.3	64.7
+Temporal Neg. (v^T)	77.2	51.9	68.0	66.2
+SEASON	78.7	54.8	66.0	66.5

Table 5. Ablation study of *SEASON*’s key components. We evaluate the impact of the spatial negative (v^S) and temporal negative (v^T) on temporal and overall hallucination examination tasks.

Models	VidHalluc		VideoHalluc	
	Temporal	Overall	Temporal	Overall
LLaVA-OV-7B [13]	53.0	70.3	52.5	46.4
+Spatial Neg. (v^S)	75.7	75.1	52.0	46.7
+Temporal Neg. (v^T)	77.7	74.2	55.5	47.5
+SEASON	77.5	75.4	55.5	48.5
QWEN2.5-VL-7B [4]	59.0	73.4	46.5	53.1
+Spatial Neg. (v^S)	77.0	77.4	42.0	53.4
+Temporal Neg. (v^T)	79.3	77.7	47.0	52.9
+SEASON	77.7	78.7	49.5	54.8

4.4. Ablation Study

Analysis on Temporal Negative. To validate the design of our *temporally-hard* negative (Sec. 3.2), we conduct an ablation study in Tab. 3. We compare against three simpler *temporally-easy* negatives: Average, Shuffled, and Reverse. The results clearly demonstrate the superiority of our approach. Our Homogenized method achieves the highest overall Average score for both backbones (e.g., +7.2% on LLaVA-OV-7B [13] and +5.7% on QWEN2.5-VL-7B [4]).

Ablation on Key Components. We analyze the contribution of our spatial (v^S) and temporal (v^T) negatives in Tab. 4 and 5. From Tab. 4, while applying each of the two introduced negatives individually improves performance, their combination in *SEASON* achieves the highest AVG score (e.g., +5.4% on LLaVA-OV-7B [13]). In Tab. 5, the baseline of solely applying v^T achieves the best temporal-only results on VidHalluc [14]. On the other hand, *SEASON* surpasses this baseline on VideoHalluc [26]. We attribute this discrepancy to VideoHalluc’s Yes/No VQA data format, which is susceptible to compliance bias (preference for “Yes”), as also empirically verified by Wang et al. [26]. Despite this, *SEASON* consistently achieves the best overall (spatial and temporal) mitigation across both benchmarks, and ranks among the top two for temporal-only mitigation.

Analysis of Self-diagnostic Layers. We analyze the sensitivity of our self-diagnostic mechanism (Sec 3.3) to the selected attention layers J (Eq. 4). In Fig. 6, we plot the performance in VidHalluc [14] over three backbone models

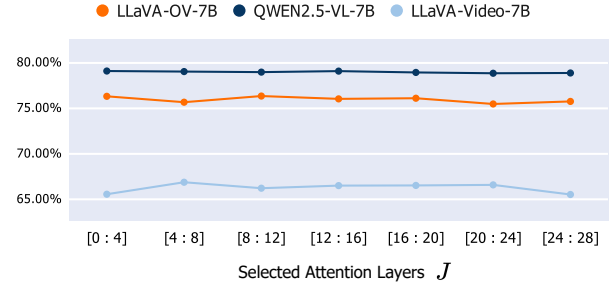


Figure 6. Ablation study on the selected attention layers (J) for *SEASON*, evaluated on the VidHalluc benchmark [14]. The performance remains robust and stable, demonstrating insensitivity to the specific layers chosen for aggregation.

Table 6. Evaluation of *SEASON* on the large-scale LLaVA-OV-72B [13] model across the subtasks of the VidHalluc [14].

Models	VidHalluc				
	BQA	MCQ	STH	TSH	AVG
LLaVA-OV-72B [13]	76.80	90.79	42.74	67.33	69.41
+SEASON (Ours)	78.46	90.93	40.43	76.00	71.46

when using different attention layers in our self-diagnostic mechanism. The results demonstrate that *SEASON* is highly robust to this hyperparameter. Performance for all models remains remarkably stable, regardless of whether early, middle, or final layers are selected.

Scalability to Large Scale VideoLLM. To demonstrate the scalability and general applicability of our training-free framework, we applied *SEASON* to the large-scale LLaVA-OV-72B [13] model. As shown in Tab. 6, *SEASON* provides a consistent performance boost, improving the overall AVG score on VidHalluc [14] by +2.05%. The most significant improvement comes from the temporally-focused TSH subtask, which increases by +8.67%. This result confirms that *SEASON* is a robust and scalable solution.

5. Conclusion

In this paper, we introduce Self-Diagnostic Contrastive Decoding (*SEASON*), a training-free method that mitigates temporal alongside spatial hallucinations in VideoLLMs. To address temporal hallucination, we employ temporal homogenization to produce the “temporally-hard” negatives to expose the spurious temporal correlations and a self-diagnostic Mechanism that detects the hallucination tendencies of each token by measuring attention divergence across original video and negatives. With these introduced innovations, *SEASON* adaptively enhance spatial and temporal faithfulness for each output token. Extensive experiments demonstrate that *SEASON* substantially reduces hallucinations compared to existing methods, achieving state-of-the-art results on multiple video hallucination benchmarks while preserving general video understanding capabilities.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 6
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1
- [3] Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13744–13753, 2025. 2
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7, 8, 1
- [5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 1
- [6] Jianfeng Cai, Jiale Hong, Zongmeng Zhang, Wengang Zhou, zhannianji, and Houqiang Li. Mitigating hallucination in videoLLMs via temporal-aware activation engineering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 6
- [7] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024. 5, 7
- [8] Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *arXiv preprint arXiv:2504.05810*, 2025. 2
- [9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 5
- [10] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 1, 2
- [11] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 1, 2
- [12] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1, 2, 3
- [13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7, 8, 1, 2, 3, 4, 5
- [14] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13723–13733, 2025. 1, 2, 3, 5, 6, 7, 8, 4
- [15] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1
- [16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6
- [17] Rui Li, Xiaohan Wang, Yuhui Zhang, Orr Zohar, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025. 2, 6, 1
- [18] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 1, 2
- [19] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, 2023. 2
- [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2
- [21] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 5, 6, 7, 2, 3, 4
- [22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 3
- [23] Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *arXiv preprint arXiv:2504.12083*, 2025. 2, 6, 1
- [24] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a

- family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. [1](#), [2](#), [6](#)
- [25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. [1](#), [2](#)
 - [26] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024. [2](#), [5](#), [8](#)
 - [27] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don’t miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. [1](#), [2](#)
 - [28] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. [1](#)
 - [29] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025. [2](#), [6](#), [7](#), [1](#)
 - [30] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. [1](#)
 - [31] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. [2](#)
 - [32] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. [1](#)
 - [33] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024. [1](#), [2](#), [3](#), [5](#), [6](#), [4](#)
 - [34] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. [6](#), [7](#)
 - [35] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. *arXiv preprint arXiv:2402.08680*, 2024. [1](#), [2](#)



SEASON: Mitigating Temporal Hallucination in Video Large Language Models via Self-Diagnostic Contrastive Decoding

Supplementary Material

A. Detailed Experimental Settings

A.1. Our hyperparameters (α, β)

To determine the optimal settings for *SEASON*, we performed a grid search over the hyperparameters α and β for each benchmark. We explored the following configurations:

$$(\alpha, \beta) \in \{(1.0, 0.33), (0.5, 0.25)\}$$

This resulted in a total of 2 configurations evaluated for our proposed method.

A.2. Implementation Detail of Other Baselines

Training-Free Baselines We compare our *SEASON* against two training-free baselines: TCD [33] and DINO-HEAL [14]. As the official implementations were not available at the time of writing, we re-implemented both methods strictly following the details provided in their respective papers. To ensure a fair comparison, we applied a similar grid search strategy to these baselines for each benchmark:

TCD [33]: We tuned the frame downsampling rate r and the contrastive decoding parameters (α, β) over the following search space:

$$r \in \{2, 4\}, \quad (\alpha, \beta) \in \{(1.0, 0.1), (0.5, 0.5)\}$$

This yields a total of $2 \times 2 = 4$ configurations.

DINO-HEAL [14]: We searched over two key components: normalization usage and DINO model variants.

$$\text{Normalization} \in \{\text{Enabled}, \text{Disabled}\}$$

$$\text{DINO Variants} \in \{\text{With Registers}, \text{Without Registers}\}$$

This results in $2 \times 2 = 4$ configurations.

Training-Based Baselines For training-based baselines: ArrowRL [29], TPO [17], and RRPO [23], we utilize the official pre-trained checkpoints provided by the respective authors. We strictly adhere to the model configurations specified within these checkpoints. All other implementation details and experimental settings are kept consistent with the original models to ensure the validity of the comparison.

A.3. Prompts among all Benchmarks

We follow each benchmark’s official provided prompt to implement our inference code.

B. Additional Analysis

B.1. Latency Report

We report the inference latency to evaluate the computational cost of our proposed method. All measurements were conducted on a single NVIDIA H100 80GB GPU. Tab. 7 details the average inference time per sample (seconds/sample) on the VidHalluc [14] benchmark.

Table 7. **Per-sample inference latency** comparison on the VidHalluc [14] benchmark. Results are reported in seconds.

Models	BQA	MCQ	STH	TSH
LLaVA-OV-7B [13]	0.80	0.84	0.89	1.04
+TCD [33]	0.94	1.00	1.06	1.20
+DINO-HEAL [14]	1.25	1.06	0.86	1.12
+SEASON (Ours)	1.21	1.28	1.38	1.48
QWEN2.5-VL-7B [4]	0.77	0.87	0.92	1.12
+TCD [33]	0.90	1.03	1.07	1.24
+DINO-HEAL [14]	0.80	0.91	0.88	1.19
+SEASON (Ours)	1.19	1.33	1.43	1.58

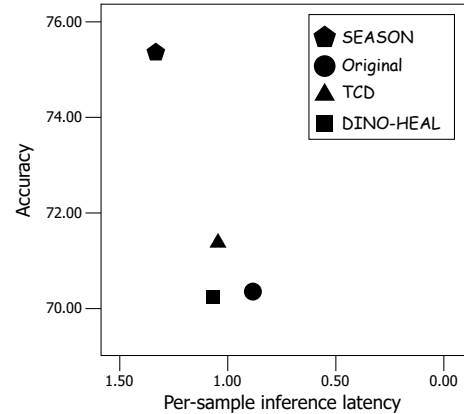


Figure 7. Corresponding accuracy and latency of applying TCD [33], DINO-HEAL [14], and *SEASON* with LLaVA-OV-7B [13] on VidHalluc [14].

As expected, our method introduces moderate computational overhead compared to the base model due to additional operations (Sec. 3.2 and Sec. 3.3). However, as shown in Fig. 7, the latency remains within a reasonable range, while offering significant improvements in hallucination mitigation and preserving general video understanding (Tab. 1 and Tab. 2).

B.2. Hyperparameter Sensitivity (α , β)

In Fig. 8, we evaluated the performance on the TSH subtask in VidHalluc [14] for the purpose of investigating the sensitivity of our method to the hyperparameters α and β .

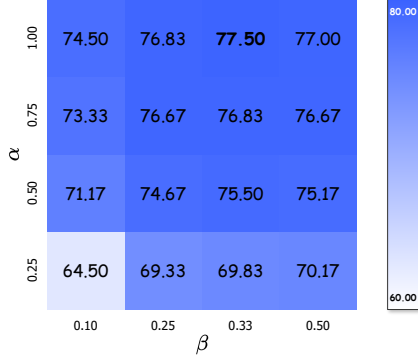


Figure 8. Analysis of α and β . The values represent the Accuracy of LLaVA-OV-7B [13] on the TSH subtask in VidHalluc [14].

As observed in Fig. 8, our method demonstrates consistent performance across a wide range of configurations. The performance on the TSH subtask in VidHalluc [14] improves as α reaching 1.00 and β reaching 0.33.

B.3. Temporal Homogenization Layers Ablation

In Sec. 3.2, we apply Temporal Homogenization at layers in the Model’s Vision Encoder (E_θ). Recall that at a given layer l and frame f_t , the homogenized feature $h_{l,t}$ is defined as a linear combination of the frame feature from the corresponding global context d_l and the pre-homogenization feature $h'_{l,t}$ ($h_{0,t}$ are the patch embeddings of frame f_t):

$$h_{l,t} = (1 - \beta)h'_{l,t} + \beta d_l, \text{ where } h'_{l,t} = E_\theta^{(l)}(h_{l-1,t}).$$

By default, this operation is applied to all layers. In Tab. 8, we vary the range of homogenization layers on LLaVA-OV-7B [13], in order to investigate the impact of layer selection in Model’s Vision Encoder (e.g., Early, Middle, and Late).

Table 8. Ablation study on the effect of applying Temporal Homogenization to different layers in Model’s Vision Encoder (E_θ).

Applied Layers	VidHalluc	VideoHalluc	AVG
	TSH	TPH	
Early Layers	71.2	49.0	60.1
Middle Layers	74.2	53.0	63.6
Late Layers	<u>77.0</u>	<u>54.5</u>	<u>65.8</u>
All Layers (default)	77.7	55.5	66.6

The results demonstrate that applying Temporal Homogenization to **All Layers** achieves the highest performance in temporal hallucination examination. Among the partial applications, Late Layers significantly outperform others.

C. Additional Qualitative Evaluation

In Figs. 9 to 15, we present additional qualitative results of applying TCD [33], DINO-HEAL [14], and *SEASON* with LLaVA-OV-7B [13] on TempCompass [21]. *SEASON* exhibits temporal faithfulness within its generated captions, which demonstrate the effectiveness of *SEASON* in mitigating temporal hallucinations.



Prompt) You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly.\nInformation A: {'subject': 'entire video', 'order': 'A girl is shown reading a book and a body of water is shown simultaneously'}\nInformation B: {'subject': 'entire video', 'order': 'A girl is shown reading a book, then a body of water is shown'}\nInformation C: {'subject': 'entire video', 'order': 'A body of water is shown, then a girl is shown reading a book'}\n

Ground Truth) B. A girl is shown reading a book, then a body of water is shown

LLaVA-OV-7B) A girl is shown reading a book and a body of water is shown **simultaneously**.

+TCD) A girl is shown reading a book and a body of water is shown **simultaneously**.

+DINO-HEAL) A girl is shown reading a book and a body of water is shown **simultaneously**.

+SEASON) A girl in a black floral dress lies on the grass, **holding an open book in her hands**, deeply engrossed in her **reading**. The serene environment around her includes tall grass and a body of water, with the sun casting a warm, golden light over the scene. **As the video progresses**, the girl continues to read, with the **body of water becoming more prominent in the background**, enhancing the peaceful and contemplative atmosphere.

Figure 9. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.



Prompt) A video and multiple pieces of information will be provided to you. One of these pieces of information matches the content of the video, while the remaining ones do not. Your objective is to pinpoint the information that is in harmony with the video and craft a suitable video caption.\nInformation A: {'subject': 'entire video', 'order': 'a white car is driving while snow particles are melting simultaneously'}\nInformation B: {'subject': 'entire video', 'order': 'snow particles are melting, then a white car is driving'}\nInformation C: {'subject': 'entire video', 'order': 'a white car is driving, then snow particles are melting'}\n

Ground Truth) B. snow particles are melting, then a white car is driving

LLaVA-OV-7B) A white car is driving **while** snow particles are melting **simultaneously**.

+TCD) A white car is driving **while** snow particles are melting **simultaneously**.

+DINO-HEAL) A white car is driving **while** snow particles are melting **simultaneously**.

+SEASON) As the video progresses, the **snow particles continue to melt**, revealing more green grass and twigs underneath. **The scene then transitions** to a highway at sunset, where **a white car drives towards** a row of wind turbines on the left side of the road. **The car moves forward**, passing by the wind turbines under a sky filled with scattered clouds, with the sun setting in the distance."

Figure 10. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.

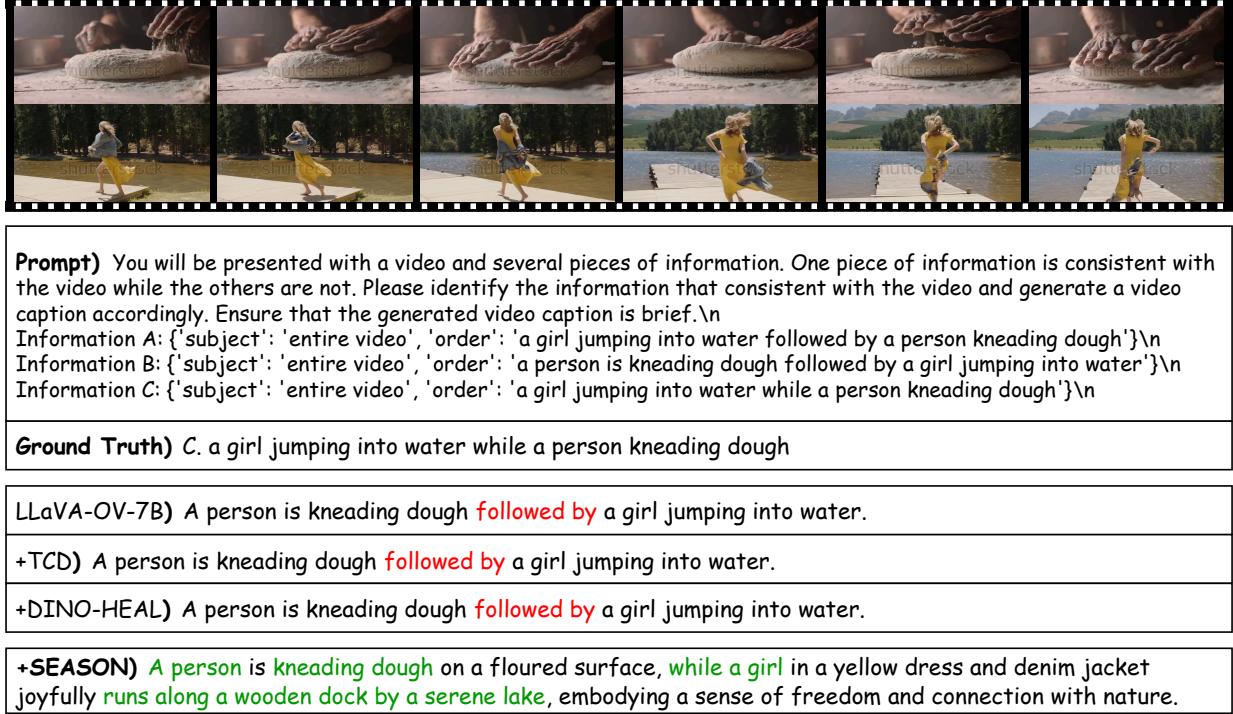


Figure 11. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.

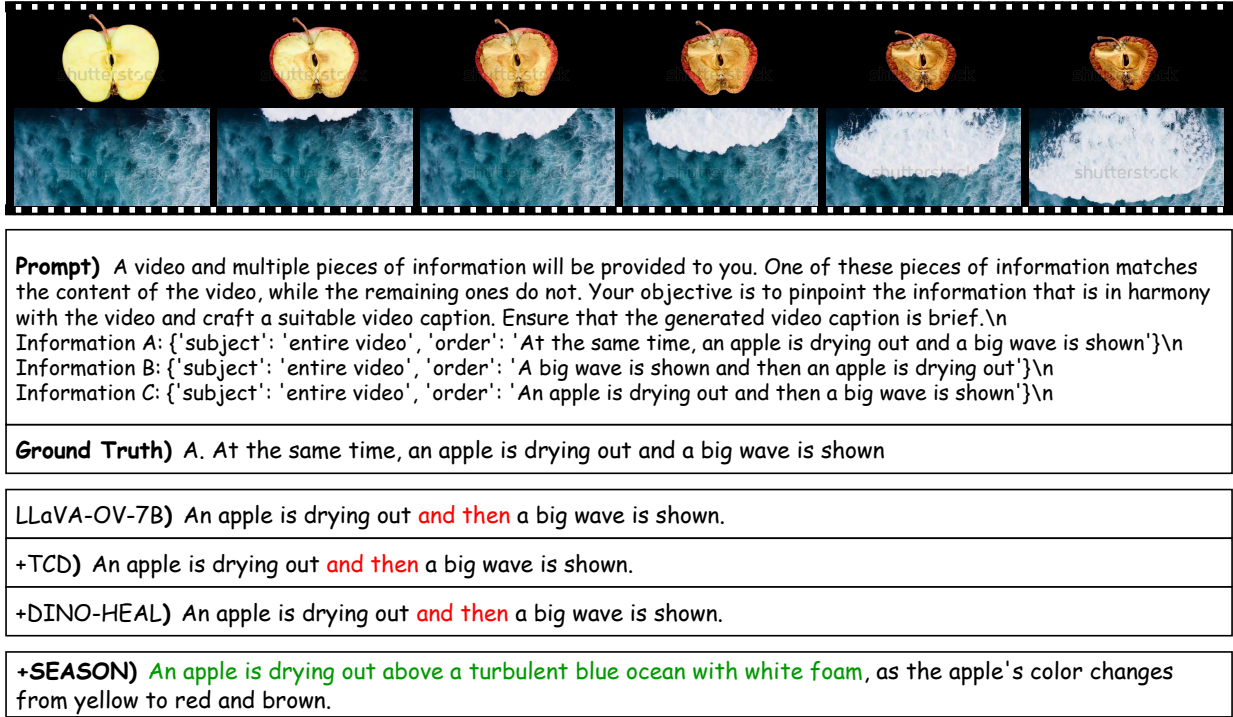
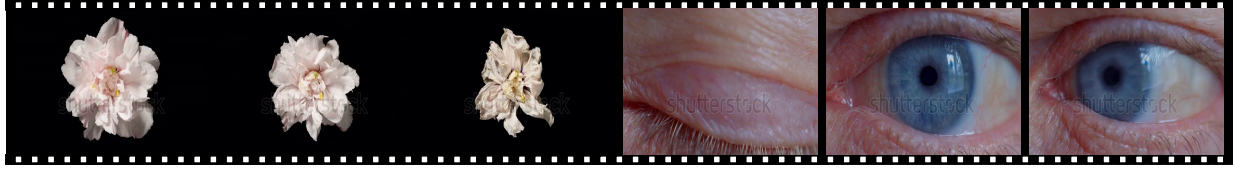


Figure 12. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.



Prompt) You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly. Ensure that the generated video caption is brief.\n
 Information A: {'subject': 'entire video', 'order': 'An eye is opening first, then a flower is wilting'}\n
 Information B: {'subject': 'entire video', 'order': 'Both events happen simultaneously'}\n
 Information C: {'subject': 'entire video', 'order': 'A flower is wilting first, then an eye is opening'}\n

Ground Truth) C. A flower is wilting first, then an eye is opening

LLaVA-OV-7B) A flower is wilting **while** an eye is opening.

+TCD) A flower is wilting **while** an eye is opening.

+DINO-HEAL) A flower is wilting **while** an eye is opening.

+SEASON) The video **begins with a stunning pink flower** in full bloom against a stark black background, **followed by a close-up of an eye opening**, showcasing the intricate details of the skin around it and the vibrant blue iris with a black pupil at the center.

Figure 13. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.



Prompt) You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly. Ensure that the generated video caption is brief.\n
 Information A: {'subject': 'the street', 'order': 'first filled with vehicles and then filled with people'}\n
 Information B: {'subject': 'the street', 'order': 'first filled with people and then filled with vehicles'}\n
 Information C: {'subject': 'the street', 'order': 'filled with people and vehicles simultaneously'}\n

Ground Truth) B. first filled with people and then filled with vehicles

LLaVA-OV-7B) The street is filled with people and vehicles **simultaneously**.

+TCD) The street is filled with people and vehicles **simultaneously**.

+DINO-HEAL) The street is filled with people and vehicles **simultaneously**.

+SEASON) From an aerial perspective, the video captures a bustling city intersection teeming with **pedestrians crossing a zebra-striped crosswalk**. **Cars occasionally pass through the intersection**, adding to the dynamic urban scene.

Figure 14. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.



Prompt) You will be presented with a video and several pieces of information. One piece of information is consistent with the video while the others are not. Please identify the information that consistent with the video and generate a video caption accordingly. Ensure that the generated video caption is brief.\n

Information A: {'subject': 'entire video', 'order': 'A dog is giving a high five to a woman, and then foods are added into a bowl'}\nInformation B: {'subject': 'entire video', 'order': 'Foods are added into a bowl, while at the same time a dog is giving a high five to a woman'}\nInformation C: {'subject': 'entire video', 'order': 'Foods are added into a bowl, and then a dog is giving a high five to a woman'}\n

Ground Truth) B. Foods are added into a bowl, while at the same time a dog is giving a high five to a woman

LLaVA-OV-7B) A dog is giving a high five to a woman, **and then** foods are added into a bowl.

+TCD) A dog is giving a high five to a woman, **and then** foods are added into a bowl.

+DINO-HEAL) A dog is giving a high five to a woman, **and then** foods are added into a bowl.

+SEASON) As the video progresses, more ingredients are added to the bowl, including shrimp, cherry tomatoes, avocado slices, and thinly sliced red onions, while the dog continues to give the high five, creating a heartwarming scene of companionship and mutual understanding.

Figure 15. Qualitative comparison of video captions predicted by LLaVA-OV-7B [13] with TCD [33], DINO-HEAL [14], and SEASON on TempCompass [21]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.