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 [34] and DINO-HEAL [15]. 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 [34]: 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 [15]: We searched over two key components: normalization usage and DINO model variants.

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

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

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

Training-Based Baselines For training-based baselines: ArrowRL [30], TPO [18], and RRPO [24], 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 [15] benchmark.

Table 7. **Per-sample inference latency** comparison on the VidHalluc [15] benchmark. Results are reported in seconds.

Models	BQA	MCQ	STH	TSH
LLaVA-OV-7B [14]	0.80	0.84	0.89	1.04
+TCD [34]	0.94	1.00	1.06	1.20
+DINO-HEAL [15]	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 [34]	0.90	1.03	1.07	1.24
+DINO-HEAL [15]	0.80	0.91	0.88	1.19
+SEASON (Ours)	1.19	1.33	1.43	1.58

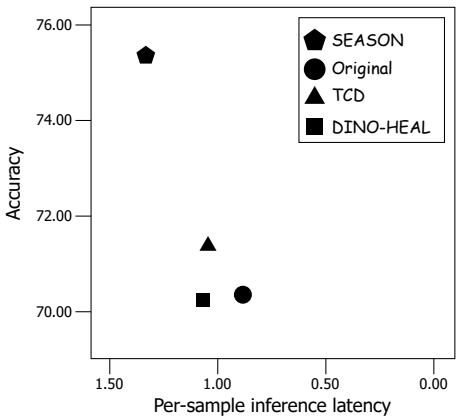


Figure 7. Corresponding accuracy and latency of applying TCD [34], DINO-HEAL [15], and *SEASON* with LLaVA-OV-7B [14] on VidHalluc [15].

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 [15] for the purpose of investigating the sensitivity of our method to the hyperparameters α and β.

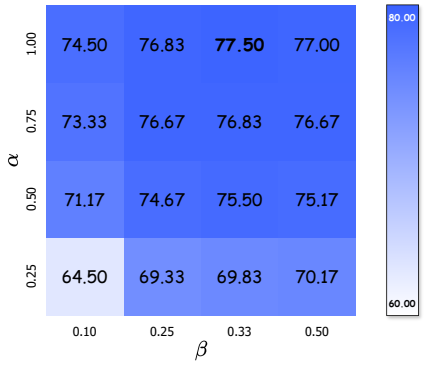


Figure 8. Analysis of α and β. The values represent the Accuracy of LLaVA-OV-7B [14] on the TSH subtask in VidHalluc [15].

As observed in Fig. 8, our method demonstrates consistent performance across a wide range of configurations. The performance on the TSH subtask in VidHalluc [15] 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 [14], 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	VideoHallucer	AVG
	TSH	TH	
Early Layers	71.2	49.0	60.1
Middle Layers	74.2	53.0	63.6
Late Layers	77.0	54.5	65.8
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 Experiments

C.1. Additional Quantitative Evaluation

In Tab. 9, we evaluate *SEASON* on LVBench, by scaling the number of input frames (**#F**) from 8 to 64 with Qwen2.5-VL-7B [4]. We observe consistent accuracy gains with a moderate increase in latency on long video understanding benchmarks.

Table 9. Evaluation of accuracy (%) and efficiency on LVBench.

Models	#F	ER	EU	KIR	TG	Rea	Sum	sec/sample
Qwen2.5-VL-7B [4]	64	40.2	42.7	45.7	38.6	47.3	32.8	6.10
***+SEASON* (Ours)**	64	**42.4**	**44.1**	**47.1**	**40.0**	**48.3**	**37.9**	9.58

In Tab. 10, we evaluate *SEASON* on the MLVU reasoning task using Qwen3-VL-8B-Thinking, where we still observe improved performance by *SEASON*, demonstrating the preservation of CoT reasoning capability.

Table 10. Evaluation of accuracy (%) on MLVU reasoning task.

Models	#F	Topic Reasoning	Ego Reasoning
Qwen3-VL-8B-Thinking	32	80.2	30.2
***+SEASON* (Ours)**	32	**82.4**	**35.9**

C.2. Additional Qualitative Evaluation

In Figs. 9 to 15, we present additional qualitative results of applying TCD [34], DINO-HEAL [15], and *SEASON* with LLaVA-OV-7B [14] on TempCompass [22]. *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 [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. 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 [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. 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 girl jumping into water followed by a person kneading dough'}\n
Information B: {'subject': 'entire video', 'order': 'a person is kneading dough followed by a girl jumping into water'}\n
Information C: {'subject': 'entire video', 'order': 'a girl jumping into water while a person kneading dough'}\n

Ground Truth) C. a girl jumping into water while a person kneading dough

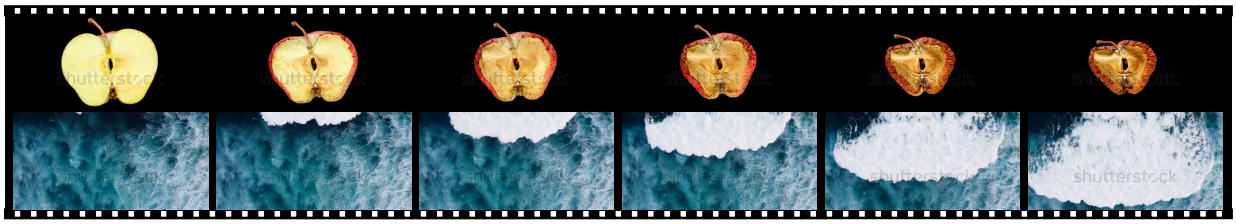
LLaVA-OV-7B) A person is kneading dough followed by a girl jumping into water.

+TCD) A person is kneading dough followed by a girl jumping into water.

+DINO-HEAL) A person is kneading dough followed by a girl jumping into water.

+SEASON) A person is kneading dough on a floured surface, while a girl in a yellow dress and denim jacket joyfully runs along a wooden dock by a serene lake, embodying a sense of freedom and connection with nature.

Figure 11. Qualitative comparison of video captions predicted by LLaVA-OV-7B [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. 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. Ensure that the generated video caption is brief.\n
Information A: {'subject': 'entire video', 'order': 'At the same time, an apple is drying out and a big wave is shown'}\n
Information B: {'subject': 'entire video', 'order': 'A big wave is shown and then an apple is drying out'}\n
Information C: {'subject': 'entire video', 'order': 'An apple is drying out and then a big wave is shown'}\n

Ground Truth) A. At the same time, an apple is drying out and a big wave is shown

LLaVA-OV-7B) An apple is drying out and then a big wave is shown.

+TCD) An apple is drying out and then a big wave is shown.

+DINO-HEAL) An apple is drying out and then a big wave is shown.

+SEASON) An apple is drying out above a turbulent blue ocean with white foam, as the apple's color changes from yellow to red and brown.

Figure 12. Qualitative comparison of video captions predicted by LLaVA-OV-7B [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. 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 [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. 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

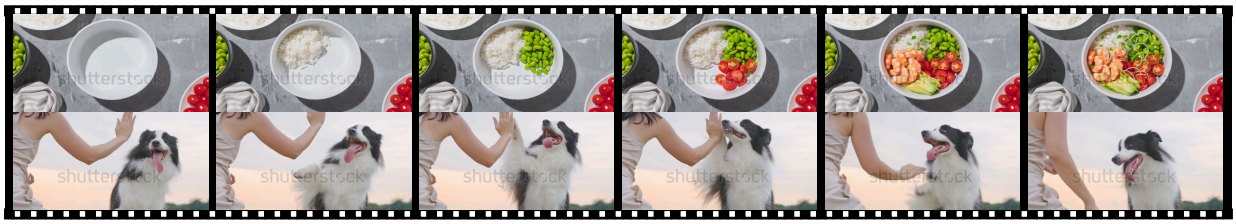
LLaVA-OV-7B) The street is filled with people and vehicles simultaneously.

+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 [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. 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 [14] with TCD [34], DINO-HEAL [15], and *SEASON* on TempCompass [22]. Note that words highlighted in **green** indicate temporal faithfulness, while those in **red** indicate temporal hallucination.