

---

# Distort Time to Improve Video Temporal Reasoning (Appendix)

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Adaptive Plausibility Constraint

2 **Contrastive Decoding in Video Large Language Models.** Given a text query  $\mathbf{x}$  and a video input  
3  $\mathbf{V}$ , the model generates two output distributions: one conditioned on the original  $\mathbf{V}$  and the other  
4 on the distorted video input  $\mathbf{V}'$ , which is derived by applying pre-defined distortion (e.g., adding  
5 noise to visual features as the simplest case) to  $\mathbf{V}$ . Then, a new contrastive probability distribution  
6 is computed by leveraging the differences between two original distributions. The new contrastive  
7 distribution  $p_{\text{vtd}}$  is formulated as:

$$p_{\text{vtd}}(\mathbf{y} \mid \mathbf{V}, \mathbf{V}', \mathbf{x}) = \text{softmax}[(1 + \alpha)\text{logit}_{\theta}(\mathbf{y} \mid \mathbf{V}, \mathbf{x}) - \alpha\text{logit}_{\theta}(\mathbf{y} \mid \mathbf{V}', \mathbf{x})], \quad (1)$$

8 where larger  $\alpha$  indicate a stronger amplification of the differences ( $\alpha = 0$  reduces to regular decoding).

9 **Adaptive Plausibility Constraint.** Eq. 1 rewards texts favored by the response with original video  
10 inputs and penalizes texts favored by the response with distorted video inputs. However, the response  
11 with distorted video inputs is not always mistaken. Although video inputs are distorted, they may  
12 still preserve useful information, which can lead to correct answers. Therefore, penalizing all texts  
13 from response with distorted video inputs indiscriminately would penalize these correct answers, and  
14 conversely reward implausible answers. To tackle this issue, we follow Li et al. [4] to introduce the  
15 plausibility constraint.

16 Adaptive plausibility constraint is contingent upon the confidence level associated with the output  
17 distribution with original video inputs:

$$\mathcal{V}_{\text{head}}(\mathbf{y}_{<t}) = \left\{ \mathbf{y}_t \in \mathcal{V} : p_{\theta}(\mathbf{y}_t \mid \mathbf{V}, \mathbf{x}, \mathbf{y}_{<t}) \geq \beta \max_{\mathbf{w}} p_{\theta}(\mathbf{w} \mid \mathbf{V}, \mathbf{x}, \mathbf{y}_{<t}) \right\} \quad (2)$$

$$p_{\text{vtd}}(\mathbf{y}_t \mid \mathbf{V}, \mathbf{V}', \mathbf{x}) = 0, \quad \text{if } \mathbf{y}_t \notin \mathcal{V}_{\text{head}}(\mathbf{y}_{<t}) \quad (3)$$

18 where  $\mathcal{V}$  is the output vocabulary of LVLMs and  $\beta$  is a hyperparameter in  $[0, 1]$  for controlling the  
19 truncation of the next token distribution. Larger  $\beta$  indicates more aggressive truncation, keeping only  
20 high-probability tokens.

21 Combining the video contrastive decoding and the adaptive plausibility constraint, we obtain the full  
22 formulation:

$$\mathbf{y}_t \sim \text{softmax}[(1 + \alpha)\text{logit}_{\theta}(\mathbf{y}_t \mid \mathbf{V}, \mathbf{x}, \mathbf{y}_{<t}) - \alpha\text{logit}_{\theta}(\mathbf{y}_t \mid \mathbf{V}', \mathbf{x}, \mathbf{y}_{<t})] \quad (4)$$

23 subject to  $\mathbf{y}_t \in \mathcal{V}_{\text{head}}(\mathbf{y}_{<t})$

## 24 B Technical Details of Video Temporal Distortion

### 25 B.1 Disrupting Moving Content in Remaining Frames

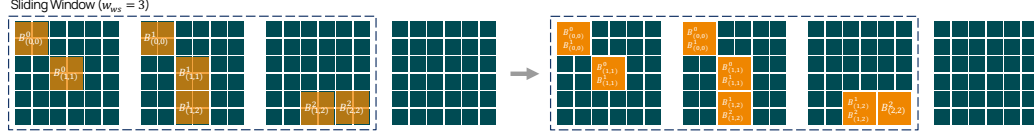


Figure 1: Disrupt moving content in remaining frames within a sliding window. **Left:** Marked dynamic blocks. **Right:** Fusion results of marked dynamics.

26 **Downsampling.** Note that in the downsampling stage, i.e., when we downsample each frame from  
 27 size  $(H, W)$  to  $(\frac{H}{w_{bs}}, \frac{W}{w_{bs}})$ , we do not really reduce token numbers. As illustrated in Fig. 1 (left), we  
 28 actually replace all image tokens within one “downsampled” region with the mean of the tokens  
 29 included. For example, the value of each of the four token within  $B^0_{(0,0)}$  is the mean of the four  
 30 tokens.

## 31 C Experiments

### 32 C.1 Experimental Configuration

33 In **TempCompass** [5], we use the following hyperparameters:  $\alpha = 1$ ,  $\beta = 0.2$ ,  $w_{fdr} = 0.2$ ,  
 34  $w_{tdr} = 0.4$ ,  $w_{ws} = 8$ ,  $w_{cfr} = 0.3$ ,  $w_{bs} = 3$ ,  $w_{fpr} = 0.5$ ,  $w_{momentum} = 0.8$ . In **EventHallusion** [6],  
 35 we use the following hyperparameters:  $\alpha = 1$ ,  $\beta = 0.2$ ,  $w_{fdr} = 0.5$ ,  $w_{tdr} = 0.5$ ,  $w_{ws} = 8$ ,  
 36  $w_{cfr} = 0.3$ ,  $w_{bs} = 3$ ,  $w_{fpr} = 0.8$ ,  $w_{momentum} = 0.8$ . In **Video-MME** [1] and **MLVU** [7], we use  
 37 the following hyperparameters:  $\alpha = 1$ ,  $\beta = 0.2$ ,  $w_{fdr} = 0.6$ ,  $w_{tdr} = 0.8$ ,  $w_{ws} = 8$ ,  $w_{cfr} = 0.3$ ,  
 38  $w_{bs} = 3$ ,  $w_{fpr} = 0.8$ ,  $w_{momentum} = 0.5$ .

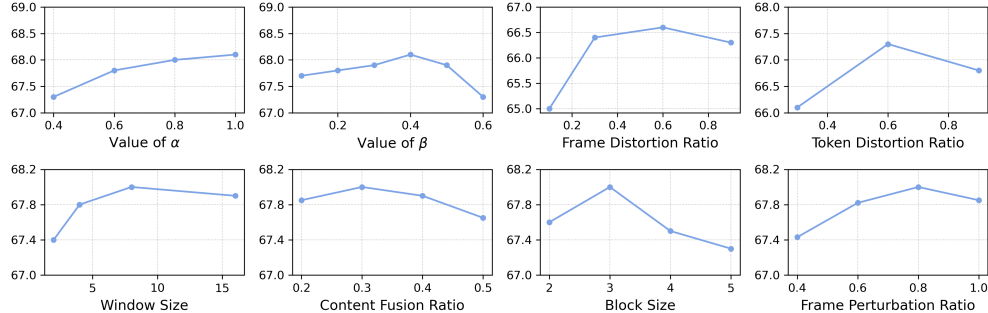


Figure 2: Sensitivity to Hyperparameter Settings on TempCompass [5].

### 39 C.2 Analysis

40 From Fig. 2, we observe that a moderate level of distortion is crucial for effective contrastive  
 41 decoding. In the ablation study of parameters most closely related to the degree of distortion—such  
 42 as Frame Distortion Ratio, Token Distortion Ratio, Content Fusion Ratio, and again Frame Distortion  
 43 Ratio—we find that setting the values too low results in limited improvements, while excessively high  
 44 values, i.e., severe distortion, lead to a relative decline in performance. This aligns with our earlier  
 45 analysis: overly severe distortion tends to randomize the model’s responses, thereby undermining  
 46 its role as a negative response to guide the generation in contrastive decoding. Only appropriately  
 47 calibrated distortion can effectively induce negative responses, thereby enhancing performance via  
 48 contrastive decoding.

## 49 D Limitations

50 When performing video distortion, our Temporal Distortion Unit relies solely on signals from the  
51 model itself—specifically, the attention maps extracted from the intermediate LLM layers—as  
52 guidance to estimate the importance of each visual token and each video frame. Compared to treating  
53 all frames equally and applying uniform random sampling, our approach represents a significant  
54 improvement. However, it is still not perfect. Attention maps do not always accurately reflect the  
55 true importance of each visual token, and relying on them often yields only coarse-grained results.  
56 To more precisely assess the importance of visual representations, future work may explore more  
57 accurate and robust methods beyond attention-based guidance.

58 Moreover, our current study is limited to Video LLMs, with distortion applied only to the visual  
59 representations. In practice, many videos come with accompanying subtitles, and models often  
60 take both video and subtitle inputs. An interesting future direction would be to distort both modal-  
61 ities—applying not only visual distortion but also video-aware distortion to subtitles. This would be  
62 challenging and different from the purely text-based distortion strategies employed in existing works  
63 on contrastive decoding for LLMs.

## 64 E Extended Discussion

65 We are among the first to explore video temporal understanding from the perspective of language and  
66 image priors, and to enhance it using contrastive decoding with video temporal distortion.

67 Recent works [3, 2] have applied contrastive decoding to mitigate hallucinations in image understand-  
68 ing with MLLMs. For example, VCD [3] introduces random noise to distort the original image, while  
69 SID [2] prunes important tokens based on attention guidance. TCD [6] alleviates event hallucination  
70 in videos by randomly dropping frames.

71 SID [2] adopts a similar strategy to estimate token importance and removes the most important  
72 tokens—this is conceptually similar to the second step of our video temporal distortion. However,  
73 there are notable differences in how attention maps are utilized and how the pruning is applied.  
74 Specifically, SID uses attention maps from the  $k$ -th layer to assess token importance and then prunes  
75 the most important tokens starting from the  $(k + 1)$ -th layer.

76 In contrast, our approach aggregates attention maps from all layers, from shallow to deep, to compute  
77 a more accurate importance score. Furthermore, while SID performs pruning from intermediate  
78 layers, we input the distorted video representations directly at the first layer, ensuring that the dropped  
79 information is effectively masked from the very beginning. This design allows our method to better  
80 mask information that should be dropped.

## 81 References

- 82 [1] Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou,  
83 Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of  
84 multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- 85 [2] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-  
86 introspective decoding: Alleviating hallucinations for large vision-language models. In *Proceedings*  
87 *of the International Conference on Learning Representations (ICLR)*, 2025. Published as a conference paper  
88 at ICLR 2025.
- 89 [3] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing.  
90 Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In  
91 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- 92 [4] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettle-  
93 moyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings*  
94 *of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, Canada, 2023.  
95 Main conference long paper.
- 96 [5] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou.  
97 Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

- 98 [6] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing  
99 event hallucinations in video llms. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial*  
100 *Intelligence (AAAI)*. Association for the Advancement of Artificial Intelligence, 2025. To appear.
- 101 [7] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, et al. Mlvu: A comprehensive benchmark for multi-task long  
102 video understanding. *arXiv preprint arXiv:2406.04264*, 2024.