

# MAD: Modality-Adaptive Decoding for Mitigating Cross-Modal Hallucinations in Multimodal Large Language Models

## Supplementary Material

### 7. Impact of $\gamma$

In our modality-adaptive weighting mechanism, the temperature parameter  $\gamma$  is a crucial hyperparameter that controls the contrastive strength between modality-specific distributions and the full distribution. To determine the appropriate  $\gamma$  value, we conduct a systematic ablation study by varying  $\gamma$  from 0.5 to 3.0 with intervals of 0.5. For each  $\gamma$  value, we evaluate the performance of our MAD approach on AVHBench [23] and CMM [24] benchmarks using VideoLLaMA2-AV [51] and Qwen2.5-Omni [38] as the base model.

#### 7.1. Results and Analysis

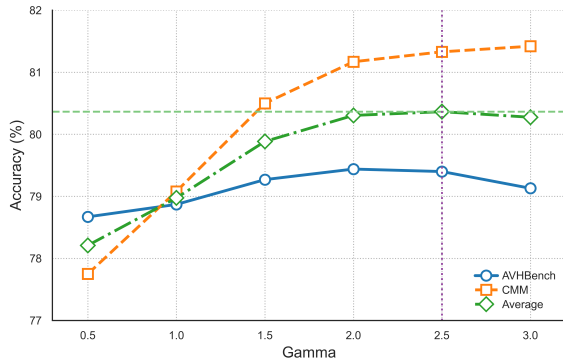


Figure 4. Impacts of  $\gamma$  in VideoLLaMA2-AV

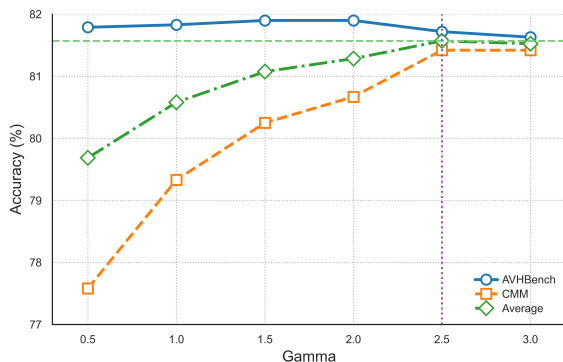


Figure 5. Impacts of  $\gamma$  in Qwen2.5-Omni

Based on this analysis, we use  $\gamma = 2.5$  as the default value across all experiments. This value demonstrates consistently robust performance across different models and benchmarks.

### 8. Qualitative Analysis of Modality-Adaptive Weights

In the main paper, we present a quantitative analysis demonstrating that the extracted modality weights  $w_m$  accurately reflect the modality relevance implied by each task. Here, we provide additional details on the experimental setup and present qualitative examples to further illustrate the behavior of our modality-adaptive weighting mechanism.

#### 8.1. Dataset Construction

We randomly sample 100 videos from VideoMME [50] and construct 300 questions categorized into three types based on their modality dependency:

- **Visual-related questions** (100 questions): Questions that require only visual information to answer.
- **Audio-related questions** (100 questions): Questions that require only audio information to answer.
- **Audio-visual-related questions** (100 questions): Questions that require both audio and visual information to answer.

#### 8.2. Question Examples

Table 5 presents representative examples of questions from each category along with their corresponding computed modality weights.

#### 8.3. Analysis

As shown in Table 5, our modality-adaptive weighting mechanism successfully assigns higher weights to the relevant modality for each question type. For visual-related questions,  $w_v$  is consistently higher, indicating strong reliance on video information. Conversely, audio-related questions exhibit higher  $w_a$  values. For audio-visual questions requiring multimodal reasoning,  $w_{av}$  receives the highest weight, demonstrating the model’s ability to recognize the need for integrated multimodal understanding.

### 9. Modality Weight Distribution Analysis Across Tasks

To validate that our modality-adaptive weighting mechanism appropriately adjusts weights according to task characteristics, we analyze the distribution of modality weights across different question types in AVHBench and CMM benchmarks. All experiments are conducted using VideoLLaMA2-AV as the base model.

Table 5. Representative examples of questions and their modality weights.

Category	Question	$w_v$	$w_a$	$w_{av}$
Visual	Is the singer a black woman?	0.58	0.17	0.25
	Does the man with white hair wear glasses?	0.72	0.02	0.26
	Does the character’s surroundings get brighter when the character moves?	0.84	0.15	0.01
Audio	Can you hear seagulls?	0.22	0.53	0.25
	What kind of music is playing in the background?	0.16	0.59	0.25
	Is the background music upbeat and fast-paced?	0.09	0.73	0.18
Audio-Visual	Is there music when she appears on stage?	0.27	0.16	0.56
	Does the narration match the subtitles?	0.22	0.11	0.67
	Does the character move when there is a clicking sound?	0.39	0.11	0.50

### 9.1. Weight Analysis on AVHBench

AVHBench is specifically designed to evaluate cross-modal hallucinations by presenting questions where one modality can mislead the model’s understanding of another modality. Figure 6 shows the distribution of modality weights ( $w_v$ ,  $w_a$ ,  $w_{av}$ ) across different task categories.

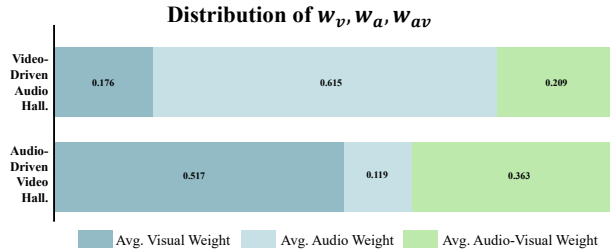


Figure 6. Analysis on AVHBench

**Video-Driven Audio Hallucination (V→A).** In this task category, visual information can mislead audio-related understanding. Despite the potential interference from video, the questions fundamentally require audio comprehension. Our analysis reveals that the model correctly identifies audio as the primary modality, assigning the highest proportion to  $w_a$ . This demonstrates that MAD successfully prioritizes the relevant modality even in the presence of misleading cross-modal information.

**Audio-Driven Video Hallucination (A→V).** Conversely, this task involves audio information potentially interfering with video understanding. The weight distribution shows a predominant  $w_v$  proportion, indicating that the model appropriately recognizes video as the essential modality for answering these questions.

This symmetric behavior across V→A and A→V tasks validates the adaptability of our weighting mechanism.

### 9.2. Weight Analysis on CMM

The CMM benchmark categorizes cross-modal hallucinations into three types based on the dominant modality that causes interference. Figure 7 presents the modality weight distributions for each category.

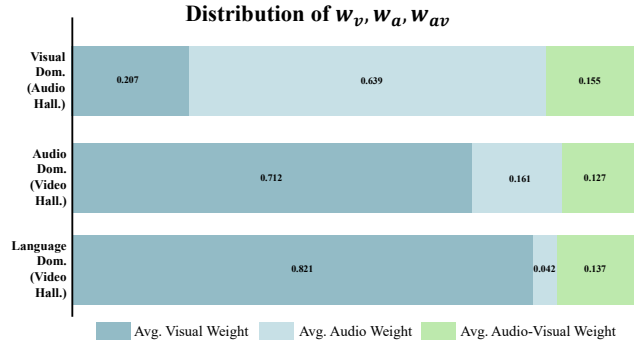


Figure 7. Analysis on CMM

**Visual Dominance (Visual Dom.).** This category addresses cases where visual information is over-relied upon, leading to the neglect of audio information. Our analysis shows that the model assigns higher weight to  $w_a$ , effectively preventing audio hallucinations by emphasizing the underrepresented audio modality.

**Audio Dominance (Audio Dom.).** In contrast, this category involves audio information inappropriately influencing video understanding. The weight distribution demonstrates increased  $w_v$  proportion, which helps suppress video hallucinations caused by audio interference. This behavior mirrors the Visual Dom. pattern but in the opposite direction, further confirming the mechanism’s adaptability.

**Language Dominance (Language Dom.).** This category predominantly contains video-related questions that are susceptible to language bias, such as ”Did you see the shape

of the wheel is circular in the video?” These questions can be answered through linguistic priors without genuine visual understanding. The model compensates for this by assigning substantially higher weight to  $w_v$ , forcing greater reliance on actual visual evidence and thereby mitigating language-based shortcuts.

### 9.3. Discussion

The consistent patterns observed across both benchmarks demonstrate that our modality-adaptive weighting mechanism successfully captures task-specific modality requirements. The model autonomously adjusts weights to emphasize the most relevant modality while suppressing potentially misleading cross-modal influences, validating the effectiveness of our approach in addressing cross-modal hallucinations.

## 10. Robustness Analysis of Modality Query Prompts

To evaluate the robustness of our modality-adaptive weighting mechanism, we conduct ablation studies with various modality query prompts. While our main paper uses the prompt  $X_m = \text{“To answer this question, which modality is needed (audio, video, or both)”}$ , we investigate whether alternative prompt formulations affect the model’s ability to generate appropriate modality weights.

### 10.1. Alternative Prompt Formulations

We design the following alternative modality query prompts that maintain similar semantic intent but differ in phrasing:

- $X_{m'_1}$ : “Identify which modality is required to answer the question (audio, video, or both)”
- $X_{m'_2}$ : “Given this question, select the necessary modality for reasoning (audio, video, or both)”
- $X_{m'_3}$ : “Which modality does this question require (audio, video, or both)”
- $X_{m'_4}$ : “State the modality relevant for answering this question (audio, video, both)”

### 10.2. Experimental Results

Figure 8 presents the performance statistics across different modality query prompts on both AVHBench and CMM benchmarks. The results demonstrate remarkable consistency across prompt variations, with standard deviations of only 0.26% on AVHBench and 0.31% on CMM. The narrow range between minimum and maximum performance values (0.59% and 0.83% respectively, indicated by the vertical bars) further confirms that our approach is highly robust to variations in prompt formulation.

This robustness suggests that the model’s modality weight generation relies on semantic understanding of the modality dependency concept rather than specific prompt

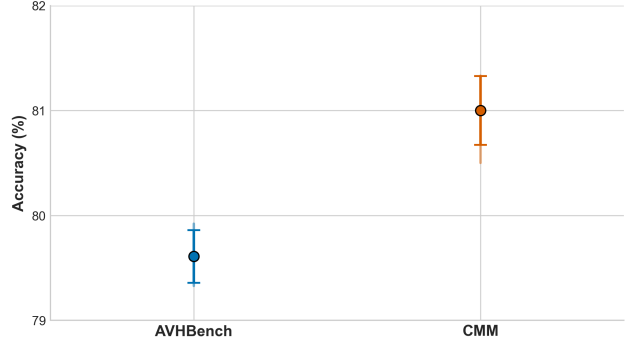


Figure 8. Performance consistency across different modality query prompts on AVHBench and CMM benchmarks using VideoLLaMA2-AV. Circles denote mean  $\pm$  standard deviation; vertical bars show minimum and maximum accuracy across all five prompt variants.

engineering or surface-level pattern matching. The consistent performance across diverse phrasings validates that MAD captures a fundamental principle of modality-task alignment, making it a principled and reliable approach for addressing cross-modal hallucinations.

## 11. Computational Analysis

Our MAD method extracts modality-adaptive weights through self-assessment prompts and applies weighted contrastive decoding during inference. To analyze the computational overhead introduced by our approach, we compare the decoding latency (ms/token)↓ with other contrastive decoding-based methods. We conduct all experiments on 8 NVIDIA RTX A6000 GPUs using VideoLLaMA2-AV and Qwen2.5-Omni.

Table 6. Computational efficiency comparison among CD-based methods.

Method	VideoLLaMA2-AV	Qwen2.5-Omni
VCD <sub>extend</sub>	3564.46	4431.78
AVCD	4811.13	9489.81
MAD (Ours)	3571.64	6700.56

## 12. Qualitative Results

We provide quantitative results demonstrating the effectiveness of MAD in Fig 9 and Fig 10. Here, we provide qualitative examples showing how our method mitigates cross-modal hallucinations.



**Video:** A situation where a metal duck target plate is being shot with a gun

**Sound:** The sound of a bullet hitting a metal duck

**Question:** Please describe what you can hear and see in detail.

### Base

**\*Vision-driven audio hallucination**

The video shows a metal box with four metal ducks on it, sitting on the ground.

The main source of sound in the video is the **quacking of the ducks**.

Suddenly, a gun is shot, and the ducks start to move, creating a loud quacking sound.

### AVCD

The video shows a metal box with four duck cutouts on it.

The main source of sound in the video is the **quacking of ducks**.

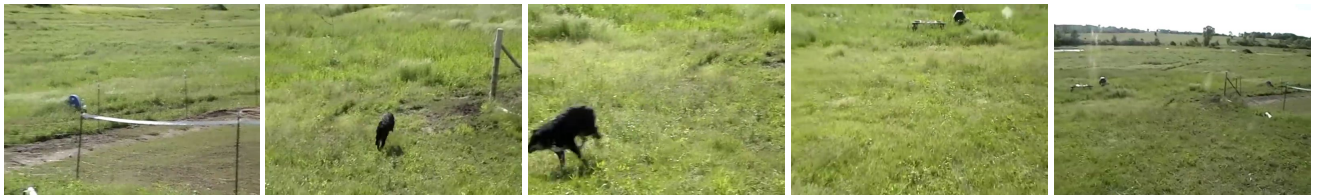
The person in the video is seen holding a metal box with a duck cutout on it.

### MAD

The video shows a metal box with four duck shapes on it, sitting on the ground or a concrete surface.

The main source of sound in the video is the **sound of shooting**, which is heard throughout the video.

It appears to be a duck-shaped target or practice aid for shooting



**Video:** A situation where a dog is running toward the camera

**Sound:** The sound of nearby sheep bleating

**Question:** Did you see any sheep?

### Base

**\*Audio-driven visual hallucination**

**Yes**

### AVCD

**Yes**

### MAD

**I did not see any sheep in the video.**

Figure 9. Qualitative Results in VideoLLaMA2-AV





**Video:** A situation where a gun is being loaded at a shooting range

**Sound:** The sound of a gun being shot by others around

**Question:** Did he pull the trigger? Please describe what you can hear and see in detail.

**Base**

**\*Audio-driven visual hallucination**

Well, from what I can see, there's a guy in a camouflage shirt. He's got a gun in his hand. He seems to be at some kind of shooting range, there's a wooden table in front of him with some stuff on it, like a bag and a couple of other things. He's holding the gun and then **he pulls the trigger**. You can hear the sound of the gun firing. It's a pretty clear shot. So, yeah, **he did pull the trigger**. What do you think about it?

**AVCD**

Well, from what I can tell, it seems like **he did pull the trigger**. I can hear a distinct click sound, which is usually what you'd hear when a trigger is pulled. And then there's a loud bang, which is likely the sound of the gun firing.

**MAD**

Well, it seems like **the man in the video is preparing for some kind of shooting activity**. He's handling a gun and there are other items on the table like a knife and a bottle. He's also wearing a green shirt. **But, I'm not sure if he actually pulled the trigger or not**. What do you think?



**Video:** A situation where a motorcycle with its engine turned off is being filmed

**Sound:** The sound of mowing grass around

**Question:** Please describe what you can hear and see in detail.

**Base**

**\*Vision-driven audio hallucination**

Well, I can see a motorcycle. It's got a sleek black and white design. The engine looks really powerful, and the wheels are big and sturdy. There's a license plate on the back. As for what I can hear, **there's a sound of a motorcycle engine revving up**. It's a bit loud, you know? It gives off that feeling of power and speed. What do you think about motorcycles? Do you have a favorite?

**AVCD**

Well, I can see a motorcycle. It's got a black seat and a shiny engine. There are some stickers on it too. And **I can hear a motorcycle engine starting up**. It's a bit of a rough start, you know? It's like the engine is trying to get going. What do you think about motorcycles? Do you ride one?

**MAD**

Well it seems like there's a motorcycle in the video. The front part of the motorcycle is visible and it has a license plate. **But hmm I can't really tell what's happening with the motorcycle or what the sound is about**. Maybe you could give me a bit more information?

Figure 10. Qualitative Results in Qwen2.5-Omni