AVCD: Mitigating Hallucinations in Audio-Visual Large Language Models through Contrastive Decoding

Chaeyoung Jung* **Youngjoon Jang*** **Joon Son Chung**
Korea Advanced Institute of Science and Technology (KAIST)

Abstract

Hallucination remains a major challenge in multimodal large language models (MLLMs). To address this, various contrastive decoding (CD) methods have been proposed that contrasts original logits with hallucinated logits generated from perturbed inputs. While CD has shown promise in vision-language models (VLMs), it is not well-suited for AV-LLMs, where hallucinations often emerge from both unimodal and cross-modal combinations involving audio, video, and language. These intricate interactions call for a more adaptive and modality-aware decoding strategy. In this paper, we propose Audio-Visual Contrastive Decoding (AVCD)—a novel, training-free decoding framework designed to model trimodal interactions and suppress modality-induced hallucinations in AV-LLMs. Unlike previous CD methods in VLMs that corrupt a fixed modality, AVCD leverages attention distributions to dynamically identify less dominant modalities and applies attentive masking to generate perturbed output logits. To support CD in a trimodal setting, we also reformulate the original CD framework to jointly handle audio, visual, and textual inputs. Finally, to improve efficiency, we introduce entropy-guided adaptive decoding, which selectively skips unnecessary decoding steps based on the model's confidence in its predictions. Extensive experiments demonstrate that AVCD consistently outperforms existing decoding methods. Especially, on the AVHBench dataset, it improves accuracy by 2% for VideoLLaMA2 and 7% for video-SALMONN, demonstrating strong robustness and generalizability. Our code is available at https://github.com/kaistmm/AVCD.

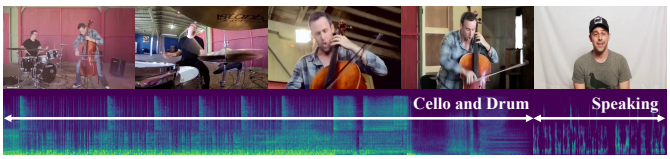
1 Introduction

Large language models (LLMs) have achieved remarkable success, enabling AI systems to perform a wide range of text-based tasks such as problem-solving, summarization, translation, and human interaction [1, 7, 41, 52, 57, 58, 72]. To extend these capabilities beyond language, recent advances have introduced multimodal large language models (MLLMs), which integrate visual and auditory inputs alongside text. By incorporating multiple modalities, MLLMs enhance the model's ability to understand and solve complex tasks that require multimodal reasoning [11, 35, 43, 50, 64, 69].

However, despite these advancements, hallucination remains a persistent challenge. It refers to the generation of biased or factually incorrect information that does not faithfully reflect the given input, posing a significant barrier to the reliable deployment of both LLMs and MLLMs [25, 37, 40, 53, 55, 63, 71]. In vision-language models (VLMs), hallucinations frequently occur when the model misinterprets visual inputs and generates responses that fail to align with the actual visual content. This issue is largely attributed to an over-reliance on statistical biases [2, 3, 17, 33] inherited from pretraining and the dominant influence of language priors [4, 20, 45, 61]. As a result, models often produce responses based on familiar language patterns instead of accurately reflecting the visual input, a problem that has been extensively addressed in previous studies [6, 18, 28, 56].

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



AVCD
The video shows a man playing the cello in a room. The **sound of the cello** is deep and rich, and it is accompanied by the **sound of the drums** in the background. The man is wearing a hat and **a shirt with a bird on it**, and he is smiling. The video also shows a man playing the drums in a room with a window. The sound of the drums is loud and clear, and it is accompanied by the sound of the cello in the background.

Question
Describe the given video and audio in detail.

*Visual mismatch
*Audio-visual mismatch

Base
The video shows a man playing the cello **in front of a piano**. **The cello and piano are both sounding at the same time**. The man is wearing **a black shirt** and hat, and he is smiling. The video also shows a man playing drums in a room with a red wall. The man is wearing a plaid shirt and jeans. He is sitting down and playing the drums with his hands.

Figure 1: **Hallucination mitigation with Audio-Visual Contrastive Decoding (AVCD).** Inaccurate visual and audio-visual information is highlighted in red and blue, respectively, and corrected during inference via AVCD, enabling the production of precise details such as *'a shirt with a bird on it'*.

To address these challenges, contrastive decoding (CD) has recently been applied to VLMs. These methods perturb a fixed single modality and obtain the logits from the distorted information, which are then contrasted with those generated from the unaffected input. This balances modality reliance and effectively reduces hallucinations. As VLMs have two modalities (vision and language), one branch of research perturbs the language information, while another distorts the visual information for CD. For language perturbation, ICD [56] introduce distorted instructions to generate corrupted outputs. For vision perturbation, CODE [26] replaces visual input tokens with self-generated descriptions, while VCD [28] introduces Gaussian noise into image inputs. Similarly, SID [24] retains only the least informative visual tokens to generate distorted outputs. However, CD has not yet been explored in AV-LLMs, where hallucinations pose a particularly difficult challenge due to the complexity of multimodal interactions [27]. Figure 1 shows that *Base* decoding can lead to hallucinations not only from single visual inputs but also from audio-visual combinations, highlighting the need for more sophisticated modeling of cross-modal interactions to mitigate such errors.

In this paper, we propose Audio-Visual Contrastive Decoding (AVCD), training-free decoding framework tailored for AV-LLMs. The core innovation of AVCD lies in how it generalizes the idea of CD, which has been primarily explored in VLMs. Unlike prior methods that perturb a fixed modality, AVCD dynamically identifies the less dominant modalities, guided by the model's attention distribution across the three modalities. AVCD then applies attentive masking to selectively perturb the less dominant modalities and contrast the resulting biased logits with those derived from the original. This encourages the model to rely on more balanced multimodal evidence, thereby reducing hallucinations that may arise from over-dependence on any single modality or from misaligned cross-modal cues. Furthermore, we reformulate the original CD to jointly handle audio, visual, and textual modalities. This allows the model to consider the more intricate hallucination patterns that can emerge from combinations of multiple modalities, which prior methods overlook. Finally, to enhance inference efficiency, we also introduce an entropy-guided adaptive decoding mechanism. Specifically, when the model exhibits high confidence in its predictions, additional decoding steps are skipped. This strategy significantly reduces the computational overhead associated with performing multiple forward passes, while preserving the benefits of contrastive reasoning. By combining dominance-aware attentive masking, trimodal-aware contrastive formulation, and efficient inference, AVCD effectively mitigates modality-induced hallucinations in AV-LLMs. As shown in Figure 1, AVCD enhances complex cross-modal interactions, resulting in reduced hallucinations.

We validate the effectiveness of AVCD through extensive experiments across a range of MLLMs, including evaluations involving image [38], video [11, 35], and audio-visual inputs [11, 50]. The results consistently demonstrate that AVCD mitigates hallucinations during inference, regardless of the input modality. Specifically, we evaluate AVCD on the AVHBench dataset [51], a benchmark specifically designed to assess hallucinations in audio-visual settings. When applied to two widely used AV-LLMs, AVCD achieves a 2% relative improvement in accuracy on VideoLLaMA2 [11] and 7% on video-SALMONN [50]. These improvements underscore AVCD's robustness and its effectiveness in enhancing the reliability of AV-LLMs for trimodal understanding.

2 Related Work

Vision-language models. Recent advancements in MLLMs have significantly improved vision-language integration, enhancing both multimodal reasoning and interaction capabilities [8, 23, 30,

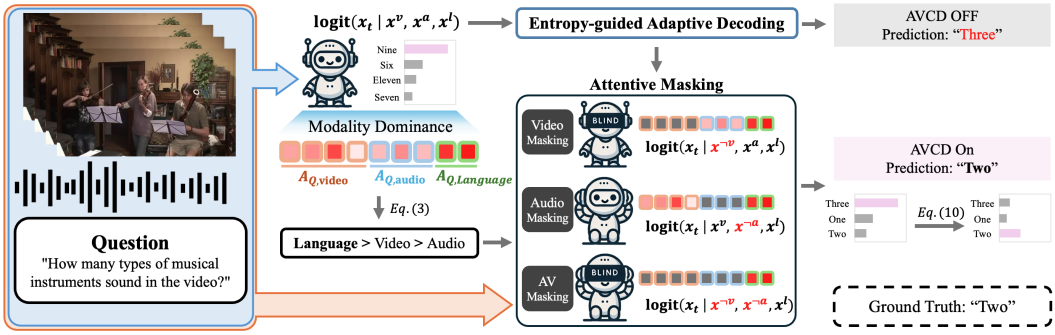


Figure 2: **Overall AVCD pipeline.** Given an audio-visual input and a question, the model generates predicted logits along with a stacked modality dominance score D_M, computed by summing the attention values of the final query token across modalities from the attention map A_{Q_K} (Eq. (3)). To improve efficiency, CD is skipped when the model's prediction has high confidence (i.e., low entropy). Otherwise, once a dominant modality is identified (e.g., language > video > audio), AVCD applies all possible masking combinations across the less dominant modalities (Audio, Video, and Audio-Visual) for CD. An attentive masking strategy is used to perturb the less dominant modalities, and CD is performed using Eq. (10). This process promotes balanced multimodal reasoning by enhancing the influence of weaker modalities (e.g., audio and video) while maintaining efficient inference.

66, 70, 75]. LLaVA [39] extends the LLaMA [54] framework, a foundational text-based LLM, by incorporating visual inputs to enable multimodal reasoning and instruction-following. Flamingo [5] introduces a few-shot learning framework that bridges visual and textual information for contextual reasoning. Expanding upon this approach, InstructBLIP [14] enhances BLIP-2 [30] through instruction tuning, refining image-based dialogue and multimodal understanding. Expanding beyond static images, Video-ChatGPT [43] extends GPT [7]-based models to engage with video content, enabling interactive video analysis. More recently, VideoLLaVA [35] builds on LLaVA, adding temporal reasoning to improve video-based dialogue and comprehension.

Audio-visual large language models. With the advancement of VLMs, AV-LLMs have emerged [10, 11, 12, 21, 42, 47, 65, 68, 69, 73], expanding multimodal capabilities to include audio perception. By jointly understanding visual and auditory inputs, AV-LLMs enable more context-aware reasoning, making them well-suited for complex real-world applications. VideoLLaMA [69] extends LLaMA by integrating both visual and audio inputs, enhancing video-based reasoning. VideoLLaMA2 [11] refines this approach with improved temporal modeling and multi-frame integration, resulting in more accurate multimodal comprehension. Building on this, video-SALMONN [50] unifies speech, audio, language, and video to enable context-aware multimodal interactions across multiple sensory inputs. Recently, Meerkat [12] improves audio-visual interaction by aligning signals at multiple levels using special modules before decoding. These improvements represent a major step forward in multimodal AI, uniting language, vision, and audio for a more comprehensive understanding.

Mitigating hallucinations in LLMs. A widely adopted approach is RLHF [46], which optimizes models using human preference data to enhance response reliability. Similarly, inference-time interventions with human labels [31] incorporate real-time feedback to guide model predictions. Other strategies involve fine-tuning with hallucination-targeted datasets [19, 36] or post-hoc revisors [74] that refine outputs and correct errors. However, these approaches heavily depend on human supervision and further training, posing scalability challenges in real-world applications.

To address the challenges posed by supervised methods, recent research has explored non-training, non-human intervention techniques as more scalable alternatives [16, 22, 44]. In parallel, contrastive decoding (CD) has been proposed as a training-free approach that suppresses unreliable predictions by subtracting the logits of a weaker model from those of a stronger one [32]. This logit-level operation leads to more accurate and reliable outputs. Building on this, DoLA [13] improves model performance by conducting contrastive comparisons between the early, underdeveloped layers of the transformer and the later, more refined layers. This contrast helps the model identify and refine the outputs more effectively. Additionally, CD-based techniques have been extended to VLMs. For example, CODE [26] perturbs output quality by employing self-descriptions and comparing them with the original VLM predictions, ensuring better consistency and accuracy. ICD [56] generate corrupted outputs by introducing distorted instructions for CD. Furthermore, VCD [28] tackles

hallucinations in VLMs by injecting noise into input images and contrasting the generated responses with their original outputs. SID [24] recently addresses vision-related hallucinations by preserving the least important visual tokens in the attention map and contrasting the resulting biased outputs. These approaches help mitigate hallucinations and enhance the reliability of VLMs.

However, despite their promise, their validation has been largely confined to VLMs, leaving the complex interactions among audio, video, and language in AV-LLMs underexplored. Addressing the challenges of multimodal reasoning in this setting requires more sophisticated strategies. To address this, we reformulate the existing CD framework to generalize across a wide range of MLLMs.

3 Method

3.1 Preliminaries

Formulation for audio-visual large language models. For video data, a visual encoder such as CLIP [48] extracts frame-level features and converts them into a sequence of M fixed-length visual tokens, denoted as $\mathbf{x}^v = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$. Similarly, an audio encoder like BEATs [9] processes audio signals into a sequence of N fixed-length audio tokens, represented as $\mathbf{x}^a = \{\mathbf{x}_{M+1}, \mathbf{x}_{M+2}, ..., \mathbf{x}_{M+N}\}$. Text inputs are tokenized into L textual tokens using a tokenizer, forming $\mathbf{x}^l = \{\mathbf{x}_{M+N+1}, \mathbf{x}_{M+N+2}, ..., \mathbf{x}_{M+N+L}\}$. These modality-specific token sequences are then concatenated to create a unified representation $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{K}$, where $K = M + N + L$ is the total number of tokens. This unified representation is used as the input to the LLM decoder. Given the input sequence $\mathbf{x}$, the model generates a response autoregressively, predicting the next token based on previous generated tokens and modality-specific information:

$$\mathbf{y}_t \sim p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l, \mathbf{y}_{<t}) \propto \exp\left(\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l, \mathbf{y}_{<t})\right). \tag{1}$$

Here, $\mathbf{y}_t$ represents the token generated at timestep t, while $\mathbf{y}_{<t}$ refers to the sequence of previously generated tokens. At each timestep, the decoder first computes hidden states, which are then mapped to the vocabulary dimension through a linear projection layer. This produces a logit vector, where each element corresponds to a raw score for a token in the vocabulary. Finally, the softmax function transforms these raw scores into a probability distribution over the possible next tokens.

Contrastive decoding for vision-language models. Existing multimodal CD methods aim to mitigate hallucinations, which arise when models excessively rely on language while neglecting visual inputs [18, 28, 56]. To counteract this, a common sampling strategy is employed:

$$\text{logit} = (1+\alpha)\text{logit}(\mathbf{y}|\mathbf{x}^v, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{y}|\mathbf{x}^{\neg v}, \mathbf{x}^l), \tag{2}$$

where α controls the contrastive effect, and $\mathbf{x}^{\neg v}$ represents biased or corrupted visual information (e.g., through noise injection, augmentation, or masking) [28, 59, 60]. Corrupting a less dominant modality makes logits rely more on the unaffected modality. Subtracting these biased logits from the original reduces the dominant modality's influence, enabling more balanced multimodal reasoning.

To prevent the suppression of correct predictions, an adaptive plausibility constraint is applied to the CD-enhanced outputs, ensuring that logits significantly deviating from the original distribution are excluded. This approach, originally proposed by [32], is also adopted in our method and remains consistent with its application in various CD-based studies [13, 24, 28, 56]. A detailed explanation of the adaptive plausibility constraint can be found in Supp. B.

3.2 Audio-Visual Contrastive Decoding

In this section, we introduce Audio-Visual Contrastive Decoding (AVCD), a decoding strategy tailored for AV-LLMs. As illustrated in Figure 2, AVCD begins by generating an initial prediction from the full multimodal input while recognizing the dominant modality via attention distributions. An entropy-guided adaptive decoding mechanism then determines whether additional decoding is necessary. If the prediction is confident (i.e., low entropy), further decoding steps are skipped for efficiency. Otherwise, attentive masking is applied to the less dominant modalities to generate perturbed outputs. Finally, we apply a reformulated CD, specifically designed for trimodal settings, by contrasting the perturbed and original outputs. This enhances cross-modal understanding while maintaining efficient inference. A full description of the AVCD algorithm is provided in Supp. E.

Recognizing dominant modality via attention distributions. Transformer-based MLLMs process cross-modal information through an attention mechanism that integrates tokens from different modalities (audio, video, and language) via queries, keys, and values [11, 35, 43, 50]. The attention map captures the model's focus at each decoding step, providing a direct measure of modality influence. By analyzing the attention distribution of the final query token (A_{Q_K}) in the attention map, inspired by [24, 49], we quantify modality dominance. Higher attention weights assigned to a specific modality indicate that the modality has a stronger influence on the model's prediction.

Formally, the dominance score $D_{\mathcal{M}}$ for a given modality $\mathcal{M}$ (where $\mathcal{M}$ represents the set of indices corresponding to video, audio, or language tokens) is computed as:

$$D^j_{\mathcal{M}} = \sum_{i \in \mathcal{M}} A_{Q^j_K, i}, \quad D_{\mathcal{M}} = \frac{1}{J} \sum_{j=1}^{J} D^j_{\mathcal{M}}, \tag{3}$$

where $D^j_{\mathcal{M}}$ represents the dominance score of modality $\mathcal{M}$ at layer $j \in \{1, 2, ..., J\}$ and $A_{Q_K, i}$ is the attention weight assigned by the last query token to i-th key token from modality $\mathcal{M}$. The dominant modality is identified with the highest average $D_{\mathcal{M}}$ across layers. This formulation adaptively measures modality importance at the attention level, enabling systematic identification.

Attentive masking via zeroing out. Once the dominant modality is identified by analyzing the attention distribution of the final query token (A_{Q_K}), we perform masking on combinations of the less dominant modalities. Specifically, we apply an attentive masking strategy using a threshold defined by the top $P\%$ of the mean stacked A_{Q_K} across transformer layers. Tokens in the non-dominant modalities that exceed this threshold are set to zero, intentionally suppressing informative signals in those modalities to produce logits biased toward the dominant modality for CD. Unlike existing methods that directly distort inputs [28, 59, 60], which may introduce undesired noise due to deviations from the trained input distribution, our attentive masking strategy preserves the structure of the learned input space. This design allows AVCD to avoid the noise from directly perturbing inputs, focusing instead on mitigating hallucinations that arise from cross-modal reasoning [24].

Figure 3 demonstrates the adaptive behavior of the model under the attentive masking strategy for specific modalities. When the video modality is masked, the model compensates by leveraging audio cues, and in the absence of audio, it prioritizes visual semantics. When the text prompt is masked, the model shifts its reliance to visual elements in the video, which can lead to undesired output predictions, such as "3d art." These results demonstrate how our attentive masking strategy minimizes reliance on the masked modality, enabling the model to leverage the remaining modalities for inference.

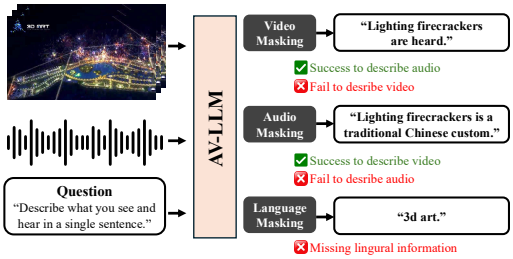


Figure 3: **Analysis of the attentive masking strategy.** By masking a specific modality, its influence is reduced, allowing the model to focus on the remaining modalities when generating outputs.

A reformulated contrastive decoding for trimodal integration. Existing CD methods operates on language and vision modalities. However, AV-LLMs generate the next token based on the joint distribution $p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l)$, incorporating an additional audio modality. Assuming the language modality is dominant while the vision and audio modalities are less dominant, we can reformulate the CD method to account for the additional modality by modeling the probabilities in Eq. (2) as follows:

$$p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \sum_{a' \in \{a, \neg a\}} \frac{1}{2} p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{a'}, \mathbf{x}^l), \tag{4}$$

$$p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^l) = \sum_{a' \in \{a, \neg a\}} \frac{1}{2} p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{a'}, \mathbf{x}^l), \tag{5}$$

where $\{a, \neg a\}$ denote the intact and corrupted audio states, respectively. Including the audio modality in the original distribution $p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l)$ may interfere with accurately modeling the relationship between language and vision, represented by $p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l)$. To mitigate this issue, we intentionally corrupt the audio information, which encourages the model to rely more on the language and vision modalities. By averaging the original and corrupted distributions, we simply reduce the impact of the

audio, making the model less sensitive to the audio modality. To convert the distribution into logit form, we apply the logarithm to Eq. (4) as follows:

$$\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \log\left(\frac{1}{2}p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \frac{1}{2}p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right). \tag{6}$$

However, directly adding probabilities inside the logarithm is undesirable since the model outputs logits rather than probabilities. To address this, we apply Taylor expansion for logarithm function:

$$\log(\frac{A+B}{2}) \simeq \frac{\log(A) + \log(B)}{2}, \tag{7}$$

with the error in the approximation being second order in the relative difference between A and B. The detailed derivation of Eq. (7) is provided in Supp. A.

Based on our derivation, we expect the difference between the original logits and those from corrupted signals to be minimal. To validate the effectiveness of our attentive masking strategy, we compute the exact approximation error (defined in Supp. A.1) between the logits from original and corrupted signals on 100 samples from the AVHBench dataset [51], after applying the adaptive plausibility constraint. As shown in Table 1, our method produces smaller errors than VCD across multiple masked modality combinations, indicating that it introduces less distortion when masking modalities. This result supports the mathematical approximation that allows moving the addition operation outside the logarithm. Accordingly, we can approximate $\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l)$ as:

Table 1: **Approximation error under the adaptive plausibility constraint.** On the AVHBench [51] dataset, AVCD produces a smaller deviation from the original logits compared to VCD [28].

Decoding	Masked Modality		
	A↓	V↓	AV↓
VCD	0.083	0.015	0.073
AVCD (Ours)	**0.032**	**0.015**	**0.037**

$$\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \frac{1}{2}\left(\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right). \tag{8}$$

Similarly, by applying the logarithm to Eq. (5) and substituting the results into Eq. (2), we derive the extended CD that leverages the visual modality:

$$\begin{aligned}\text{logit} &\propto (1+\alpha^v)\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) - \alpha^v \log p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^l) \\ &\propto (1+\alpha^v)\left(\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right) \\ &- \alpha^v\left(\log p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right),\end{aligned} \tag{9}$$

where α^v controls the degree of contrastive influence. Same procedures can be applied to the audio modality. To simultaneously address hallucinations in both video and audio, we sum the logits, applying CD to each modality as follows:

$$\begin{aligned}\text{logit}_{\text{AVCD}} = (2+\alpha^v+\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + (1-\alpha^v+\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) \\ + (1+\alpha^v-\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l) - (\alpha^v+\alpha^a)\text{logit}(\mathbf{y}|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l),\end{aligned} \tag{10}$$

where α^v and α^a control the contrastive strength in the video and audio domains, respectively.

Entropy-guided adaptive decoding. As described in Eq. (10), using AVCD can significantly increase computational cost. To mitigate this, we propose an entropy-guided adaptive decoding (EAD) that selectively applies AVCD based on the entropy of the original logit distribution. Entropy serves as a confidence measure, enabling the model to bypass unnecessary forward passes for high-confidence tokens. If the entropy is low, indicating high confidence, standard decoding is applied. Otherwise, AVCD refines uncertain predictions. This adaptive mechanism enhances inference efficiency by reducing unnecessary computation while preserving the benefits of CD for ambiguous cases.

4 Experiments

This section presents a thorough evaluation of the proposed AVCD method across multiple datasets and a diverse set of MLLMs. For systematic analysis, the MLLMs are categorized into three groups based on their input modalities: AV-LLMs, video-LLMs, and image-LLMs. Results for image-LLMs and additional experiments are included in Supp. C and Supp. D, respectively.

Table 2: **Results on AV-LLMs.** We evaluate two representative AV-LLMs across three datasets. For decoding, we compare the original model's decoding (*Base*), VCD [28] extended with audio via Eq. (2), and VCD*, which incorporates audio using our proposed formulation in Eq. (10) along with adaptive dominant modality recognition. AVCD consistently outperforms all other decoding methods across the benchmarks, demonstrating the effectiveness of both our trimodal CD formulation.

Model	Decoding	Datasets		
		MUSIC-AVQA [29]	AVHBench [51]	AVHBench-Cap
		Acc. (%) ↑	Acc. (%) ↑	Score ↑
VideoLLaMA2 [11]	*Base*	81.30±0.09	70.52	2.84±0.01
	VCD [28]	77.66±0.03	65.18	2.86±0.01
	VCD*	81.49±0.03	69.18	3.00±0.01
	AVCD	**81.58**±0.03	**72.15**	**3.03**±0.01
video-SALMONN [50]	*Base*	48.50±0.06	58.19	1.83±0.01
	VCD [28]	41.57±0.08	60.61	2.41±0.01
	VCD*	49.00±0.11	60.66	2.44±0.01
	AVCD	**49.73**±0.06	**62.18**	**2.47**±0.02

Figure 4: **Qualitative results on AV-LLM and video-LLM using VideoLLaMA2 [11].** AVCD effectively leverages all modalities by mitigating the issue of certain modalities being ignored.

4.1 Experimental Settings

Dataset and evaluation. We evaluate AV-LLMs on MUSIC-AVQA [29], which tests synchronized audio-visual reasoning with question-answering (QA) pairs derived from the MUSIC dataset. We also use AVHBench [51], designed to assess audio-visual hallucinations using adversarial QA pairs. We treat the entire dataset as the test set and use the initially released portion as the validation set, which is employed to evaluate inference speed as reported in Figure 5. Since audio-video captioning in AVHBench follows a distinct evaluation protocol, we assess it separately. For video-LLMs, we use MSVD-QA [62], which involves questions about objects, actions, and events in short video clips, using the first 1,000 test examples. We also use ActivityNet-QA [67], a more challenging benchmark requiring reasoning over long videos with complex temporal understanding.

Evaluation protocol. We evaluate model accuracy using a GPT-3.5-based framework, following the protocol of VideoLLaMA2 [11], except for AVHBench, which can be evaluated without GPT assistance. GPT-3.5 is used to perform binary verification of response correctness. To assess consistency, we report the mean and standard deviation across multiple runs.

Baselines. For AV-LLMs, we evaluate the audio-visual variants of VideoLLaMA2 [11] and video-SALMONN [50], both of which integrate visual and auditory cues to support multimodal understanding. Notably, video-SALMONN performs fusion at the feature level rather than the token level; therefore, we treat the fused audio-visual input as a single modality when comparing it against language. For video-LLMs, we assess VideoLLaMA2 in its video-only configuration and Video-LLaVA [35], an extension of LLaVA that incorporates temporal context across video frames.

Implementation details. In our experiments, the dominance-aware attentive masking method is applied to all transformer layers except the final layer. Based on the attention map, we mask the locations with the top 50% highest values (refer to Table A.3 in the Supp. D for more details). Based on our analysis that the modality dominance between video and audio is relatively balanced (see Figure A.1 in the Supp. D), we set α^v and α^a to be equal. To determine their optimal values, we randomly select 100 samples from each dataset and vary the value from 0.5 to 3.0 in increments of 0.5. As a result, we set it to 2.5 for the AVHBench dataset, and to 0.5 for all other datasets. Furthermore, we set the entropy threshold to $\tau = 0.6$ for entropy-guided adaptive decoding.

4.2 Experimental Results

Audio-visual LLMs. In Table 2, we compare three different types of decoding methods using two models, VideoLLaMA2 and video-SALMONN. We use the term *Base* to refer to the original decoding method. We first extend VCD [28] by incorporating the audio modality, following the original CD formulation. This baseline extension, however, leads to performance degradation when applied directly to AV-LLMs. In contrast, VCD*, which integrates audio using our proposed AVCD formulation (Eq. (10)) with dominant modality estimation, improves performance over the VCD method in most cases. Finally, AVCD, which further introduces a dominance-aware attentive masking strategy in place of the Gaussian noise used in VCD, consistently achieves the best results. This demonstrates the effectiveness of our approach in processing audio-visual inputs. Notably, on AVHBench, a benchmark for audio-visual hallucinations, AVCD improves accuracy by around 2% for VideoLLaMA2 and 7% for video-SALMONN relative to the base method. In addition to simple QA tasks on MUSIC-AVQA and AVHBench, AVCD also excels in audio-video captioning.

Video-LLMs. To evaluate whether AVCD generalizes beyond AV-LLMs, we apply it to a video-LLM that processes only visual and textual inputs, without audio. Specifically, we test AVCD on the VideoLLaMA2 model (Table. 3). The results show that AVCD consistently improves decoding performance across all settings, while VCD—originally designed for image-LLMs—performs worse than the base method. These findings indicate that AVCD not only mitigates hallucinations in AV-LLMs, but also serves as a generalizable decoding strategy for multi-modal models beyond its original scope. Additionally, AVCD surpasses VCD in image-LLMs as well, despite VCD being tailored for such models (see Table A.1 in Supp. C for details).

Table 3: **Results on video-LLMs.** AVCD surpasses both Base and VCD across all experiments.

Model	Decoding	Datasets	
		MSVD-QA [62]	ActivityNet-QA [67]
		Acc. (%) ↑	Acc. (%) ↑
VideoLLaMA2 [11]	*Base*	74.43±0.31	47.19±0.55
	VCD [28]	71.30±0.57	45.65±0.04
	AVCD	**75.20**±0.42	**48.22**±0.04
Video-LLaVA [35]	*Base*	70.20±0.20	47.48±0.02
	VCD [28]	71.80±0.25	47.25±0.02
	AVCD	**72.16**±0.24	**48.03**±0.14

Qualitative results. To qualitatively illustrate the effect of AVCD, Figure 4 presents QA pairs generated by AV-LLMs and video-LLMs using the VideoLLaMA2 model. In the AV-LLM example, the video and audio are misaligned, as the sound corresponds to a violin instead of a guitar. This requires complex audio-visual reasoning, where base decoding and VCD fail to correctly interpret the relationship between the modalities. However, AVCD successfully resolves this issue by properly understanding the modality interactions. In the case of the video-LLM, the video primarily depicts a snowy scene, but the prompt contains misleading words such as "warm". Since language tends to dominate over visuals, both base decoding and VCD are misled, producing incorrect responses. In contrast, AVCD overcomes this bias by appropriately adjusting the influence of each modality, leading to the correct answer. Further qualitative examples are provided in Supp. F.

4.3 Discussions

Revisiting conventional CD. We extend the conventional approach (Eq. (2)) of performing single-instance CD to handle multiple instances from two less dominant modalities simultaneously in AV-LLMs. As shown in rows 2 to 4 of Table 4, conventional CD performs comparably or better than base decoding, confirming its effectiveness for single-instance CD. By simply adapting this approach for AV-LLMs, where language is dominant, the following formulation is derived.

Table 4: **Ablation on CD with Eq. (2) and Eq. (11).** AVCD effectively extends existing CD to trimodal configurations.

Decoding	Masked Modality			Acc ↑
	A	V	AV	
Base				70.52
w/ Eq. 2	✓			71.88
w/ Eq. 2		✓		70.16
w/ Eq. 2			✓	72.07
w/ Eq. 11	✓	✓	✓	70.94
AVCD	✓	✓	✓	**72.15**

$$\begin{aligned}\text{logit} = (1+3\alpha)\text{logit}(\mathbf{x}|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{x}|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) \\ - \alpha\text{logit}(\mathbf{x}|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{x}|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l).\end{aligned} \tag{11}$$

In the fifth row of Table 4, we report the performance of CD using Eq. (11). Although there is a slight improvement compared to the base model, it is evident that the performance is lower than when using Eq. (2) with either audio or AV masking. This approach over-emphasizes audio-visual contrasts, as audio-induced hallucinations are partially resolved when CD is applied to the AV domain, and video-induced hallucinations are also partially mitigated. On the other hand, in the sixth row

of Table 4, AVCD effectively addresses imbalance by assigning a positive coefficient to each domain's output, achieving better performance than both Eq. (2)-based and Eq. (11)-based contrastive decoding, highlighting its effectiveness.

Efficient inference via entropy-guided adaptive decoding. In Figure 5, we show the trade-off between inference speed and accuracy for VideoLLaMA2 on the AVHBench validation set under varying EAD thresholds (τ). The x-axis shows average inference time per generated token, and the y-axis indicates accuracy. Higher τ reduces additional forward passes, speeding up inference but limiting performance gains. Note that FlashAttention [15] is not applied, as AVCD relies on full attention weights.

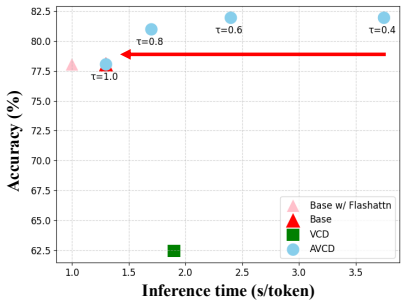


Figure 5: **Comparison across entropy thresholds (τ).** τ controls over the trade-off between inference speed and accuracy. At $\tau = 0.8$, it achieves faster inference than VCD while outperforming *Base* decoding in accuracy.

We observe that accuracy remains stable when the entropy threshold τ is below 0.6, but gradually declines as τ increases to 0.8 and 1.0. At $\tau = 1.0$, most additional decoding steps are skipped, resulting in faster inference and accuracy comparable to that of base decoding without FlashAttention. In contrast, setting $\tau = 0.8$ yields faster inference than VCD (2.25 vs. 2.5) while still improving accuracy over both the base method (78.05% $\rightarrow$ 80.98%) and VCD. These results show that our EAD strategy successfully balances speed and performance. They also demonstrate that AVCD enables efficient inference, despite requiring multiple full forward passes.

Evaluation of dominant modality recognition. To evaluate the effectiveness of our adaptive dominant modality recognition, we compare our method against a fixed strategy that assumes a single modality is always dominant. We use VideoLLaMA2 for the experiments and evaluate on AVHBench for audio-video-text inputs and MSVD-QA for video-text inputs. As shown in Table 5, treating language as the dominant modality matches the adaptive strategy, whereas fixing audio or vision leads to a clear drop in accuracy. This suggests that when all three modalities are present, the model identifies language as the dominant modality. This observation is supported by our analysis of the final-layer attention distribution (see Figure A.1 in our Supp. D), where 70% of the attention is directed toward language modality.

Table 5: **Ablation study on dominant modality recognition.** We compare fixed dominant modality settings with our adaptive recognition strategy. AVCD consistently outperforms static configurations, demonstrating the effectiveness of dynamic modality selection.

Dominant Modality	Dataset	
	AVHBench [51]	MSVD-QA [62]
Audio	68.67	-
Vision	67.79	70.70±0.14
Language	**72.15**	74.20±0.14
Adaptive	**72.15**	**75.20**±0.42

In contrast, for video-text input pairs, the attention is more evenly split between video and text (44% vs. 56%), indicating that no single modality clearly dominates. In such cases, adaptively selecting the dominant modality leads to better performance than relying on a fixed choice, such as always prioritizing video or language. These results demonstrate that, when diverse modality pairs are provided, determining the dominant modality based on attention distribution offers a flexible and effective alternative to fixed strategies, maintaining or even improving model performance.

5 Conclusion

In this work, we propose Audio-Visual Contrastive Decoding (AVCD), a training-free and generalizable inference-time decoding framework for AV-LLMs. AVCD extends CD to the trimodal setting by introducing three key components: (1) dominance-aware attentive masking to identify and perturb less dominant modalities based on attention distributions, (2) a trimodal contrastive formulation that captures hallucination patterns emerging from complex cross-modal interactions, and (3) an entropy-guided adaptive decoding mechanism that improves inference efficiency. Together, these components enable AVCD to reduce modality-induced hallucinations effectively, while maintaining computational efficiency. Our approach is broadly applicable to MLLMs beyond AV-LLMs, offering a plug-and-play solution to hallucination mitigation during inference.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT, RS-2025-02263977, Development of Communication Platform supporting User Anonymization and Finger Spelling-Based Input Interface for Protecting the Privacy of Deaf Individuals).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv*, 2023.

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proc. CVPR*, 2020.

[3] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proc. EMNLP*, 2016.

[4] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proc. CVPR*, 2018.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Proc. NeurIPS*, 2022.

[6] Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proc. WACV*, 2022.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. NeurIPS*, 2020.

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv*, 2023.

[9] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *Proc. ICML*, 2023.

[10] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Proc. NeurIPS*, 2023.

[11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv*, 2024.

[12] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *Proc. ECCV*, 2024.

[13] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *Proc. ICLR*, 2024.

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proc. NeurIPS*, 2023.

[15] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv*, 2023.

[16] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proc. ICML*, 2024.

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proc. CVPR*, 2017.

[18] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proc. CVPR*, 2024.

[19] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proc. AAAI*, 2024.

[20] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Loddon Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. 2022 ieee. In *Proc. CVPR*, 2022.

[21] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proc. CVPR*, 2024.

[22] 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 *Proc. CVPR*, 2024.

[23] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. In *Proc. NeurIPS*, 2023.

[24] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *Proc. ICLR*, 2025.

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[26] Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In *Proc. NeurIPS*, 2024.

[27] Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv*, 2024.

[28] 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 *Proc. CVPR*, 2024.

[29] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proc. CVPR*, 2022.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. ICML*, 2023.

[31] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Proc. NeurIPS*, 2024.

[32] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proc. ACL*, 2023.

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proc. EMNLP*, 2023.

[34] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv*, 2024.

[35] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proc. EMNLP*, 2024.

[36] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. In *In Proc. CoRR*, 2023.

[37] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv*, 2024.

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proc. CVPR*, 2024.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. NeurIPS*, 2023.

[40] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv*, 2023.

[41] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023.

[42] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv*, 2023.

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proc. ACL*, 2024.

[44] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proc. ACL*, 2023.

[45] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proc. CVPR*, 2021.

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, 2022.

[47] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv*, 2023.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

[49] Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, and Jinwoo Shin. Hierarchical context merging: Better long context understanding for pre-trained llms. In *Proc. ICLR*, 2024.

[50] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. In *Proc. ICLR*, 2024.

[51] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *Proc. ICLR*, 2025.

[52] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv*, 2022.

[53] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv*, 2024.

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv*, 2023.

[55] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv*, 2023.

[56] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Proc. ACL*, 2024.

[57] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *Proc. ICLR*, 2022.

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. NeurIPS*, 2022.

[59] Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv*, 2024.

[60] 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*, 2024.

[61] Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. Overcoming language priors in visual question answering via distinguishing superficially similar instances. In *Proc. ACL*, 2022.

[62] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proc. CVPR*, 2016.

[63] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv*, 2024.

[64] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv*, 2021.

[65] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *Proc. ECCV*, 2024.

[66] 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 *Proc. CVPR*, 2024.

[67] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proc. AAAI*, 2019.

[68] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv*, 2024.

[69] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proc. EMNLP*, 2023.

[70] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *Proc. ICLR*, 2024.

[71] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv*, 2023.

[72] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv*, 2023.

[73] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv*, 2023.

[74] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *Proc. ICLR*, 2024.

[75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Proc. ICLR*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We highlight our contributions in both the abstract and the final two paragraphs of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:We list our limitations in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We include the detailed proof in Supplementary Materials A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the details of our experiments and we will release the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code open access.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include experiments setting and the details in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report both the mean and standard deviation for the GPT-assisted evaluation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the computational resources in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We follow the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include that in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not pose any foreseeable risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly use open source assets following their license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We don't release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We don't use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We don't include human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: In this paper, the core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Supplementary Material: AVCD

This supplementary material complements the main paper by providing the following sections. To support code reproducibility, we include the source code along with a README file.

A Detailed Proof of AVCD

A.1 Proof of Eq. (7)

Since A and B represent probabilities as defined in Eq. (6), they are positive numbers and let $A = M + \delta$ and $B = M - \delta$, where M is the mean and δ is a small perturbation with $|\delta| \ll M$. Applying the Taylor expansion of the logarithm function, we have

$$\begin{aligned} \log(M+\delta) &\simeq \log M + \frac{\delta}{M} - \frac{\delta^2}{2M^2}, \\ \log(M-\delta) &\simeq \log M - \frac{\delta}{M} - \frac{\delta^2}{2M^2}. \end{aligned} \tag{12}$$

Now, taking the average of $\log A$ and $\log B$, we get the right-hand side term of Eq. (7):

$$\frac{\log(M+\delta) + \log(M-\delta)}{2} \simeq \log M - \frac{\delta^2}{2M^2}. \tag{13}$$

Given that the left-hand side of Eq. (7) is $\log M$, the resulting error scales with the square of the difference between A and B.

A.2 Detailed proof of Eq. (10)

Considering the language modality is dominant, CD can be extended to mitigate hallucinations in AV-LLMs by leveraging logit^v and logit^a, dealing with the video and audio modalities in CD, respectively. Applying the logarithm function to Eq. (4) yields the following:

$$\begin{aligned} &\log p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^l) \\ &= \log\left(\frac{1}{2}p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \frac{1}{2}p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right) \\ &\simeq \frac{1}{2}\left(\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right). \end{aligned} \tag{14}$$

Similarly, applying the logarithm to Eq. (5) and substituting the results into Eq. (2), we derive:

$$\begin{aligned} \text{logit}^v &\propto (1+\alpha^v)\log p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^l) - \alpha^v \log p(\mathbf{x}_t|\mathbf{x}^{\neg v}, \mathbf{x}^l) \\ &\propto (1+\alpha^v)(\log(p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)) \\ &- \alpha^v(\log p(\mathbf{x}_t|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{x}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l)), \end{aligned} \tag{15}$$

where α^v controls the degree of contrastive influence.

Following the same procedure for CD in the audio domain, we derive the corresponding logit as:

$$\begin{aligned} &\text{logit}^a \propto \\ &(1+\alpha^a)(\log p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{x}_t|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) \\ &- \alpha^a(\log p(\mathbf{x}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l) + \log p(\mathbf{x}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l)), \end{aligned} \tag{16}$$

where α^a regulates the strength of CD. Therefore, by summing Eq. (15) and Eq. (16) to account for both non-dominant modalities, we obtain Eq. (10) as the final logit expression.

B Adaptive Plausibility Constraint

CD penalizes model outputs that rely on distorted inputs, thereby promoting a more reliable generation process. However, a critical challenge arises when such penalization leads to incorrect outputs. In particular, overly strict penalties can inadvertently suppress valid outputs that align with linguistic norms and commonsense reasoning, while promoting low-probability tokens that degrade overall quality.

To address this issue, we incorporate an *adaptive plausibility constraint*, inspired by [32], which dynamically truncates the candidate logits by retaining only those tokens whose probabilities under

Table A.1: **Results on an image-LLM using the LLaVA-1.5 [38] model evaluated on the POPE [33] dataset.** AVCD outperforms both the original model's decoding (*Base*) and VCD [28], demonstrating its strong generalization capability across AV-, video-, and image-LLMs.

Method	Random		Popular		Adversarial	
	Acc. ↑	F1 Score ↑	Acc. ↑	F1 Score ↑	Acc. ↑	F1 Score ↑
MSCOCO						
Base	82.93	80.87	81.13	79.27	81.10	77.60
VCD [28]	85.53	84.04	83.63	82.32	80.87	80.13
AVCD	**86.03**	**84.87**	**84.23**	**83.24**	**81.27**	**80.70**
AOKVQA						
Base	84.03	83.22	80.20	80.00	74.23	75.33
VCD [28]	85.90	**85.46**	82.00	82.15	76.17	**77.71**
AVCD	**85.97**	84.46	**83.90**	**82.57**	**81.63**	76.20
GQA						
Base	83.60	82.79	77.90	78.11	75.13	79.20
VCD [28]	**85.97**	**85.54**	79.27	80.01	76.53	77.97
AVCD	**85.97**	84.46	**83.90**	**82.57**	**81.63**	**80.58**

the original model exceed a predefined plausibility threshold after CD. The constraint is formally defined as follows:

$$\mathcal{V}_{\text{head}}(\mathbf{y}_{<t}) = \{\mathbf{y}_t \in \mathcal{V} \mid p_\theta(\mathbf{y}_t \mid \mathbf{x}, \mathbf{y}_{<t}) \geq \beta \max_{\mathbf{w}} p_\theta(\mathbf{w} \mid \mathbf{x}, \mathbf{y}_{<t})\}, \tag{17}$$

where $\mathcal{V}$ represents the model's output vocabulary, and $\beta \in [0, 1]$ is a hyperparameter that determines the level of truncation. A higher β enforces more aggressive truncation, limiting the output space to high-confidence logits from the original distribution. For AVCD, we set $\beta = 0.1$, following the configuration used in VCD [28] and SID [24].

Given this constraint, we redefine the contrastive probability distribution as:

$$\text{logit}_{\text{AVCD}}(\mathbf{y}_t \mid \mathbf{x}, \mathbf{y}_{<t}) = -\inf, \quad \text{if } \mathbf{y}_t \notin \mathcal{V}_{\text{head}}(\mathbf{y}_{<t}). \tag{18}$$

This formulation removes tokens that deviate significantly from the original output distribution, reducing the risk of generating implausible content. By integrating this constraint with CD, we refine the final token sampling process:

$$y_t \sim \text{softmax}[\text{logit}_{\text{AVCD}}], \quad \text{subject to } y_t \in \mathcal{V}_{\text{head}}(y_{<t}).$$

This mechanism narrows the candidate pool to retain only the most probable token and prevents the model from inadvertently favoring improbable tokens due to contrasting distorted inputs.

C Experiments on Image-LLMs

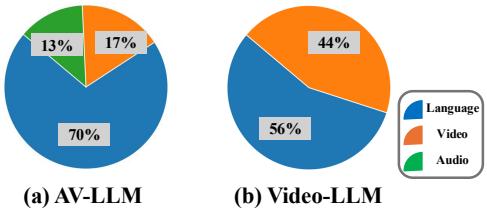
To demonstrate the broad applicability of AVCD, we also evaluate it on an image-LLM. We use the LLaVA-1.5 [38] and compare the performance of VCD [28] and AVCD. As shown in Table A.1, the results show that AVCD consistently outperforms VCD on most datasets, even though VCD was originally designed for image-LLMs. This indicates that AVCD is effective not only for audio-visual tasks but also for image-based tasks.

D Further Discussions

D.1 Modality dominance analysis

We conduct a detailed analysis of modality dominance in VideoLLaMA2 [11] using both its audio-visual and video-only variants. Following the methodology described in the main paper, we compute attention weights based on the final token and calculate the average dominance over 200 samples. For the AV-LLM, we use the AVHBench [51] dataset, and for the video-LLM, we use the MSVD-QA [62] dataset.

Figure A.1 (a) shows the modality dominance among language, video, and audio in the AV-LLM setting. Language accounts for 70% of the attention, while video and audio contribute 17% and



(a) AV-LLM (b) Video-LLM

Figure A.1: **Modality dominance analysis using VideoLLaMA2 [11].** The analysis is conducted on the AVHBench [51] dataset for audio-visual inputs (AV-LLM) and the MSVD-QA [62] dataset for video-only inputs (Video-LLM).

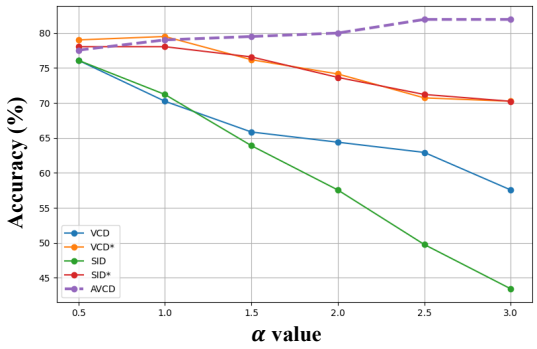


Figure A.2: **Ablation study on α values.** We evaluate several CD methods originally designed for image-LLMs, including VCD [28] and SID [24], along with our proposed AVCD, which is specifically designed for AV-LLMs. We vary the α value from 0.5 to 3. VCD* and SID* denote extended versions of the original methods, adapted using our formulation in Eq. (10). This formulation consistently improves performance when applied to all methods, demonstrating its generalizability across decoding strategies. Moreover, VCD* consistently outperforms SID* across settings, highlighting its robustness beyond image-language domains.

13%, respectively, revealing a strong bias toward the language tokens. Figure A.1 (b) presents the dominance between language and video in the video-LLM setting, showing a more balanced distribution of 56% vs. 44%.

In AV-LLM, language is the dominant modality for all 200 samples, which explains why fixing the dominant modality to language yields the same performance as selecting it adaptively, as shown in Table 5 of the main paper. In contrast, video-LLM exhibits sample-specific variation in dominant modality. This highlights the benefit of adaptively selecting the dominant modality, which results in noticeable performance gains, as evidenced in Table 5.

D.2 Impacts of α

To evaluate the performance variation with respect to changes in α, we measure the performance while gradually adjusting the value of α. As shown in Figure A.1 (a), the dominance between video and audio is relatively balanced, which motivates us to set $\alpha = \alpha^v = \alpha^a$. We also observe that even when α is adjusted to reflect the slightly higher dominance of the video modality empirically, the performance remains largely unaffected. For instance, when using $\alpha^v = 2$ and $\alpha^a = 2.5$, the accuracy on the AVHBench validation set with VideoLLaMA2 remains at 81.95%, identical to the result obtained with the balanced setting $\alpha = \alpha^v = \alpha^a = 2.5$.

As shown in Figure A.2, we compare the performance of VCD [28], SID [24], their enhanced versions VCD* and SID* (which incorporate Eq. (10)), and AVCD across different α values. AVCD achieves the best performance when $\alpha = 2.5$, while the performance of the other models drops sharply as α increases. Notably, the models that incorporate Eq. (10) exhibit a relatively smaller performance drop, indicating that our newly defined CD formulation contributes to improved model robustness. This also demonstrates that the attentive masking strategy employed in AVCD is resilient to changes in the α value.

Table A.4: **Evaluation on the OmniBench dataset.** AVCD demonstrates robust generalization, consistently surpassing the base model. The following abbreviations indicate the evaluation categories: **Action**: Action & Activity, **Story**: Story Description, **Plot**: Plot Inference, **Object**: Object Identification & Description, **Context**: Contextual & Environmental, **Identity**: Identity & Relationship, **Text**: Text & Symbols, **Count**: Count & Quantity.

Decoding	Overall	Action	Story	Plot	Object	Context	Identity	Text	Count
Base	33.5	27.1	27.0	24.1	**61.1**	30.1	46.9	**21.4**	7.1
VCD	23.3	18.3	17.4	13.1	44.6	26.2	37.5	7.1	**14.3**
OPERA	30.8	27.9	23.0	20.7	52.1	**36.9**	34.4	14.3	7.1
VCD*	31.9	27.1	26.1	21.5	57.4	30.5	43.8	**21.4**	0.0
AVCD	**34.5**	**28.3**	**28.7**	**24.5**	60.7	30.5	**50.0**	**21.4**	7.1

D.3 Evaluation on OmniBench dataset

Following OmniBench [34], we adopt video-SALMONN as the backbone model, for which the original paper reports an overall accuracy of 35.6%. Under the same setting, we reproduce the benchmark and obtain 33.5% accuracy with Base decoding, which serves as our baseline throughout the experiments.

As summarized in Table A.4, AVCD emerges as the only method that consistently surpasses the base model across most evaluation settings. In contrast, alternative approaches frequently fail to improve upon the baseline, often yielding comparable or even degraded results. This consistent superiority underscores AVCD's robust generalization capability and resilience to dataset variability, demonstrating its effectiveness even on challenging benchmarks such as OmniBench.

D.4 Per component analysis

We conduct an ablation study on AVHBench validation set using VideoLLaMA2 as the base model in Table A.2. Baseline refers to a contrastive decoding (CD) setup with randomly selected dominant modality and Eq. (11). Dominance-aware masking improves accuracy by approximately +5%. Trimodal CD with Eq. (10), our proposed extension of CD to trimodal alignment, further improves performance by +2.9%. Entropy-guided adaptive decoding (EAD) maintains accuracy while reducing inference latency by over 30%. These results validate the complementary roles of all three AVCD components as each contributes meaningfully to either accuracy or efficiency.

Table A.2: **Per component analsysis on AVHBench.** Results show that each component—dominance-aware masking, trimodal contrastive decoding, and entropy-guided adaptive decoding—plays a complementary role in improving either accuracy or efficiency.

Method	Accuracy (%)	Inference Speed (sec/token)
Baseline	74.15	4.4
+ Dominance-aware masking	79.02	4.4
+ Eq. (10)	**81.95**	4.4
+ EAD (Ours)	**81.95**	**3.1**

D.5 Impacts of masking ratios

We investigate the impact of the masking ratio P in our attentive masking strategy, which determines the proportion of high-attention tokens to be masked. As shown in Table A.3, we experiment with masking ratios of 25%, 50%, 75%, and 100% (i.e., full masking) and evaluate performance on the AVHBench validation set. Among these, a 50% masking ratio yields the highest accuracy. When the ratio is too low (e.g., 25%), the contrast with the original model output is insufficient, limiting the effectiveness of contrastive decoding. Conversely, overly high masking ratios (75% or 100%) cause large deviations from the original logits, leading to greater error in the logarithmic approximation (see Section A.1) and thus degraded performance.

Table A.3: **Ablation study on masking ratio.** A high masking ratio significantly distorts the logit distribution, while a low masking ratio retains similarity to the original logits. A 50% masking ratio is found to be optimal.

Masking ratio P	Acc (%)
25	80.98
50	**81.95**
75	80.00
100	80.00

D.6 Analysis on hallucinations

We categorize hallucinations into three representative types as defined in AVHBench. (1) A→V: Vision-centric questions misled by irrelevant or misleading audio cues (e.g., Q: Is the ship visible in the video? *Video*: No ship visible, *Audio*: Ship sounds present.). (2) V→A: Audio-centric questions misled by visual content (e.g., Q: Is the lion making sound? *Video*: Lion visible, *Audio*: No lion sound.). (3) AV Matching: Failures to correctly judge the consistency between audio and visual modalities (e.g., Q: Are the contexts of audio and visual content matching?).

Table A.5: **Categorization of hallucinations addressed by AVCD.** AVCD consistently reduces hallucinations, particularly for V→A and AV Matching.

Category	Base	AVCD	Improvement
A→V	86.4%	86.4%	0.0%
V→A	81.3%	86.3%	+5.0%
AV Matching	64.4%	71.2%	+6.8%

We evaluate AVCD on each category using VideoLLaMA2 as the base model on the AVHBench validation set. As shown in Table A.5, while AVCD preserves the already high performance in A→V, it achieves substantial improvements in V→A and AV Matching. These results support our claim that AVCD more effectively balances the contributions of both modalities.

E Algorithm of AVCD

Algorithm. 1 shows the overall AVCD algorithm. We begin by computing the original logits and estimating the modality dominance D_M. If the entropy of the original logit distribution is sufficiently low, we directly use the original logits without applying further decoding steps. Otherwise, when language is identified as the dominant modality—as is often the case—we compute the logits from audio-masked, video-masked, and audio-visual masked signals. We then apply our reformulated CD strategy as described in Eq. (10). This process is repeated autoregressively until the end-of-sequence (EOS) token is generated.

F Further Qualitative Results

We provide additional qualitative results for the AV-LLM in Figure A.3, the video-LLM in Figure A.4, and the image-LLM in Figure A.5 and Figure A.6.

G Computational Resource

We run all experiments on a machine equipped with an AMD EPYC 7513 32-core CPU and a single NVIDIA RTX A6000 GPU. To obtain reliable inference speed measurements, all background processes unrelated to the experiment are disabled during runtime.

H Limitations and Future Works

While extensive experiments demonstrate that the proposed AVCD effectively mitigates hallucination in AV-LLMs at test time, it introduces additional computational overhead due to increased forward passes as the number of modalities grows. To address this, we propose an entropy-guided adaptive decoding strategy, which significantly improves inference speed. However, this approach may overlook certain types of hallucinations that occur even when the model appears confident. In future work, we aim to develop algorithms that address the potential issues caused by skipping, through a detailed analysis of cases where hallucinations occur despite low entropy.

I Social Impact

AV-LLMs are increasingly applied in real-world scenarios such as education, medical video analysis, assistive technologies, and interactive audio-visual systems. Our proposed AVCD framework contributes to this progress by enabling more accurate and robust multimodal understanding.

Algorithm 1 Audio-Visual CD (AVCD)

Require: Multimodal inputs $(\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l)$, Audio-visual Large Language Model LM, Contrastive Weights α^v, α^a, Entropy Threshold τ

Ensure: Decoded output sequence $\mathbf{y}$

Initialize empty output sequence $\mathbf{y} \leftarrow \emptyset$

while EOS token $\notin \mathbf{y}$ **do**

  Compute original logits and dominance score:

    logit, $D_{\mathcal{M}} \leftarrow \mathrm{LM}(\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l, \mathbf{y}_{<t})$

  Compute entropy:

    $p_t \leftarrow$ softmax(logit)

    $H_t \leftarrow -\sum p_t \log p_t$

  if $H_t < \tau$ **then**

    Append top token: $\hat{y}_t \leftarrow \mathrm{argmax}$ logit

    continue

  end if

  Apply modality masking based on $D_{\mathcal{M}}$:

    Visual-masked logits:

      $\text{logit}^v \leftarrow \mathrm{LM}(\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l, \mathbf{y}_{<t})$

    Audio-masked logits:

      $\text{logit}^a \leftarrow \mathrm{LM}(\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l, \mathbf{y}_{<t})$

    Both-masked logits:

      $\text{logit}^{v,a} \leftarrow \mathrm{LM}(\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l, \mathbf{y}_{<t})$

  Compute AVCD-adjusted logits:

$$\begin{aligned}\text{logit}_{AVCD} = {} & (2 + \alpha^v + \alpha^a)\text{logit} \\ & + (1 - \alpha^v + \alpha^a)\text{logit}^v \\ & + (1 + \alpha^v - \alpha^a)\text{logit}^a \\ & - (\alpha^v + \alpha^a)\text{logit}^{v,a}\end{aligned}$$

  Append top token: $\hat{y}_t \leftarrow \arg\max \text{logit}_{AVCD}$

  Append $\hat{y}_t$ to $\mathbf{y}$

end while

return $\mathbf{y}$

As a test-time decoding method, AVCD requires no additional training or model modification, making it a practical solution for enhancing existing models. This property also promotes energy efficiency by avoiding the need for computationally intensive retraining, which is particularly beneficial when scaling AV-LLMs to real-world deployment scenarios. Overall, AVCD facilitates broader and more responsible use of multimodal models through improved inference quality and efficiency.

Prompt: Describe the given video and audio in detail.

*Visual mismatch

Base
Under the soft glow of a campfire, a group of friends gathers around a table, immersed in conversation and camaraderie. The warmth of the fire dances in their eyes as they share stories and laughter, punctuating the night air with the occasional crackle and pop of the burning wood. As the evening deepens, the flickering flames cast an ever-changing light on their faces, creating an atmosphere of cozy comfort and shared joy. In the distance, the quiet hum of the wilderness echoes off the surrounding foliage, adding to the serene ambiance of this nocturnal gathering.
❌ Fail to describe the scene change and audio information

VCD
Under the calm of the night sky, **a small group gathers around a crackling campfire**, engaged in conversation and the simple joy of being together. The warmth of the flames dances off their faces, as they share stories and laughter, the embers casting a soft glow on their smiling faces. As the night grows late, the fire's light becomes a beacon, drawing in the quiet stillness of the evening, symbolizing a moment of camaraderie and connection under the vast, dark canopy above.
❌ Fail to describe the scene change and audio information

AVCD
Under the calm of a night sky, **a small group gathers around a crackling campfire**, immersed in conversation and the shared camaraderie of an outdoor adventure. The warmth of the flames dances off their faces, accentuating the outlines of their relaxed postures as they lounge **in their camping chairs**. The rhythmic popping of the fire provides a soothing backdrop to the soft murmur of their conversation.
As the night deepens, **one of the group takes on the role of cook,** reviving the fire with a **hearty breakfast**. Amidst **the sizzling and clinking of utensils**, the aroma of cooking wafts through the air, mingling with the crisp, earthy scent of the campsite. The cook's skilled hands work methodically, **flipping and stirring the food in a black cast-iron skillet**, ensuring that every morsel is cooked to perfection.

Figure A.3: **Example of hallucination in the AV-LLM using VideoLLaMA2 [11].** The video transitions from night to morning. However, except for AVCD, both the *base* decoding and VCD [28] fail to capture this scene change. Additionally, in the nighttime scene, people are seated in camping chairs around a fire, but no table is present—an aspect incorrectly described by the *base* decoding.

Figure A.4: **Example of hallucination in the video-LLM using VideoLLaMA2 [11].** AVCD effectively mitigates hallucinations caused by the original model's failure to incorporate video context.

Prompt: Describe this photo in detail.

*Visual mismatch

Base
The painting depicts a beautiful woman in profile, wearing a black veil and smiling for the camera. She appears relaxed and graceful, possibly part of a Renaissance-era portrait. Close to her face, **there is a computer mouse that stands out due to the size difference between the two subjects**. The painting itself is a stunning piece of art, capturing the woman's expression and gaze that draw the viewer in.

VCD
The image depicts a close-up of a woman with **a veil covering her face**. She is **wearing a red dress** and smiling. The painting captures a beautifully detailed portrait of the woman, emphasizing her features and expressions. The background of the painting features a mountainous landscape, which adds depth and a sense of scale to the scene. The overall composition of the image creates a captivating and intriguing visual experience for the viewer.

AVCD
The painting is a close-up portrait of **a woman wearing a veil and a Renaissance-style dress**. She has a smiling expression, and her eyes are looking directly at the viewer. The woman's features, including her lips, nose, and chin, are distinct and well-defined. In the background of the painting, there is **a mountainous landscape with a rocky cliff visible**, adding depth and context to the scene. The overall composition of the painting captures the beauty and charm of the woman, while the background landscape creates a captivating visual backdrop.

Figure A.5: **Example of hallucination in the image-LLM using LLaVA-1.5 [38].**

Prompt: Describe this photo in detail.

*Visual mismatch

Base
The image features a painting of a **Shiba Inu dog wearing a tie and glasses**, posing with a smug expression as if engaged in theatrical stage make-believe. The dog stands as the central figure while wearing a tiny hat and a **red gown**, giving it a unique appearance. Around the painting, there is a rocky landscape with several more rocks dispersed throughout the picture. **A dining table** can be seen on the right side of the painting, possibly suggesting a stage setting for this theatrical scene. **A chair is also present behind the table**, further supporting the notion that this image evokes a theatrical or performance feel.

VCD
The image features a unique painting of a dog, possibly a poodle, wearing a hat and a Renaissance-style gown. The dog is sitting on a chair, positioned in a way that it appears to be looking at the viewer. The painting gives off a whimsical and artistic vibe. In the background, there are several elements that add to the scene's ambiance. There are **two chairs**, one placed near the dog and the other in the right-hand corner of the image. There are also **two people visible**, one standing towards the right side of the scene and another near the top right corner. Additionally, there are **two boats** present in the background, one on the left side and the other towards the right.

AVCD
The image features a painting of a person wearing **a Renaissance-style outfit**, complete with a gown, and hat. The person appears to be a dog dressed up in a costume, giving the scene a humorous and creative touch. In the background, **there are several hills and a river**, which adds to the overall artistic composition of the painting. The unique combination of a costumed dog and the Renaissance-style outfit creates an interesting visual experience.

Figure A.6: **Example of hallucination in the image-LLM using LLaVA-1.5 [38].** AVCD generates more accurate image-based descriptions compared to the *base* decoding and VCD [28].