

894 **Supplementary Material: AVCD**

895 This supplementary material complements the main paper by providing the following sections. To

896 support code reproducibility, we include the source code along with a README file.

897	<b>A Detailed Proof of AVCD</b>	<b>1</b>
898	A.1 Proof of Eq. (7) . . . . .	1
899	A.2 Detailed proof of Eq. (10) . . . . .	1
900	<b>B Adaptive Plausibility Constraint</b>	<b>1</b>
901	<b>C Experiments on Image-LLMs</b>	<b>2</b>
902	<b>D Further Discussions</b>	<b>2</b>
903	D.1 Modality dominance analysis . . . . .	2
904	D.2 Impacts of $\alpha$ . . . . .	3
905	D.3 Impacts of masking ratios . . . . .	4
906	<b>E Algorithm of AVCD</b>	<b>4</b>
907	<b>F Further Qualitative Results</b>	<b>4</b>
908	<b>G Computational Resource</b>	<b>4</b>
909	<b>H Limitations and Future Works</b>	<b>4</b>
910	<b>I Social Impact</b>	<b>4</b>

## 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  $\delta$  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  $\text{logit}^v$  and  $\text{logit}^a$ , dealing with the video and audio modalities in CD, respectively. Applying the logarithm function to Eq. (4) yields the following:

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

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

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

where  $\alpha^v$  controls the degree of contrastive influence.

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

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

where  $\alpha^a$  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 [31], 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 [36] model evaluated on the POPE [32] dataset.** AVCD outperforms both the original model’s decoding (*Base*) and VCD [27], demonstrating its strong generalization capability across AV-, video-, and image-LLMs.

Method	Random		Popular		Adversarial	
	Acc. ↑	F1 Score ↑	Acc. ↑	F1 Score ↑	Acc. ↑	F1 Score ↑
<i>MSCOCO</i>						
<i>Base</i>	82.93	80.87	81.13	79.27	81.10	77.60
VCD [27]	85.53	84.04	83.63	82.32	80.87	80.13
AVCD	<b>86.03</b>	<b>84.87</b>	<b>84.23</b>	<b>83.24</b>	<b>81.27</b>	<b>80.70</b>
<i>AOKVQA</i>						
<i>Base</i>	84.03	83.22	80.20	80.00	74.23	75.33
VCD [27]	85.90	<b>85.46</b>	82.00	82.15	76.17	<b>77.71</b>
AVCD	<b>85.97</b>	84.46	<b>83.90</b>	<b>82.57</b>	<b>81.63</b>	76.20
<i>GQA</i>						
<i>Base</i>	83.60	82.79	77.90	78.11	75.13	79.20
VCD [27]	<b>85.97</b>	<b>85.54</b>	79.27	80.01	76.53	77.97
AVCD	<b>85.97</b>	84.46	<b>83.90</b>	<b>82.57</b>	<b>81.63</b>	<b>80.58</b>

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

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

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  $\beta$  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 [27] and SID [23].

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

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

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 [36] and compare the performance of VCD [27] 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 [49] dataset, and for the video-LLM, we use the MSVD-QA [60] 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

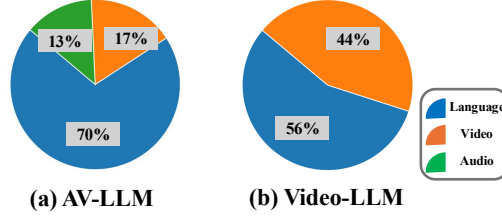


Figure A.1: **Modality dominance analysis using VideoLLaMA2 [11]**. The analysis is conducted on the AVHBench [49] dataset for audio-visual inputs (AV-LLM) and the MSVD-QA [60] dataset for video-only inputs (Video-LLM).

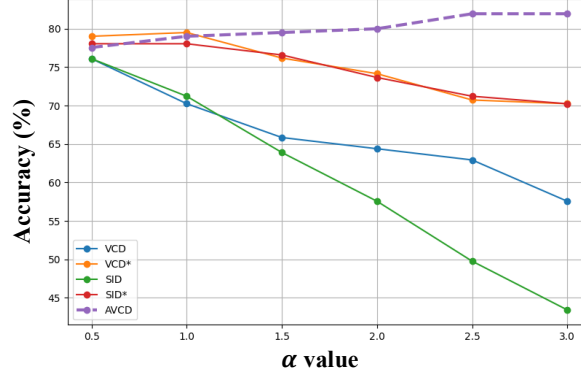


Figure A.2: **Ablation study on  $\alpha$  values.** We evaluate several CD methods originally designed for image-LLMs, including VCD [27] and SID [23], along with our proposed AVCD, which is specifically designed for AV-LLMs. We vary the  $\alpha$  value from 0.5 to 3. VCD\* and SID\* denote extended 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.

963 13%, respectively, revealing a strong bias toward the language tokens. Figure A.1 (b) presents  
 964 the dominance between language and video in the video-LLM setting, showing a more balanced  
 965 distribution of 56% vs. 44%.

966 In AV-LLM, language is the dominant modality for all 200 samples, which explains why fixing the  
 967 dominant modality to language yields the same performance as selecting it adaptively, as shown  
 968 in Table 4 of the main paper. In contrast, video-LLM exhibits sample-specific variation in dominant  
 969 modality. This highlights the benefit of adaptively selecting the dominant modality, which results in  
 970 noticeable performance gains, as evidenced in Table 4.

## 971 D.2 Impacts of $\alpha$

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

979 As shown in Figure A.2, we compare the performance of VCD [27], SID [23], their enhanced versions  
 980 VCD\* and SID\* (which incorporate Eq. (10)), and AVCD across different  $\alpha$  values. AVCD achieves  
 981 the best performance when  $\alpha = 2.5$ , while the performance of the other models drops sharply as  
 982  $\alpha$  increases. Notably, the models that incorporate Eq. (10) exhibit a relatively smaller performance  
 983 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  $\alpha$  value.

### D.3 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.2, we experiment with masking ratios of 25%, 50%, 75%, and 100% (i.e., full masking). 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.2: **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
<b>50</b>	<b>81.95</b>
75	80.00
100	80.00

## 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.

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

---

**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  $\alpha^v, \alpha^a$ , Entropy Threshold  $\tau$

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

```
1: Initialize empty output sequence  $\mathbf{y} \leftarrow \emptyset$ 
2: while EOS token  $\notin \mathbf{y}$  do
3:   Compute original logits and dominance score:
     logit,  $D_{\mathcal{M}} \leftarrow \text{LM}(\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l, \mathbf{y}_{<t})$ 
4:   Compute entropy:
      $p_t \leftarrow \text{softmax}(\text{logit})$ 
      $H_t \leftarrow -\sum p_t \log p_t$ 
5:   if  $H_t < \tau$  then
6:     Append top token:  $\hat{y}_t \leftarrow \text{argmax logit}$ 
7:     continue
8:   end if
9:   Apply modality masking based on  $D_{\mathcal{M}}$ :
10:  Visual-masked logits:
     logitv  $\leftarrow \text{LM}(\mathbf{x}^{-v}, \mathbf{x}^a, \mathbf{x}^l, \mathbf{y}_{<t})$ 
11:  Audio-masked logits:
     logita  $\leftarrow \text{LM}(\mathbf{x}^v, \mathbf{x}^{-a}, \mathbf{x}^l, \mathbf{y}_{<t})$ 
12:  Both-masked logits:
     logitv,a  $\leftarrow \text{LM}(\mathbf{x}^{-v}, \mathbf{x}^{-a}, \mathbf{x}^l, \mathbf{y}_{<t})$ 
13:  Compute AVCD-adjusted logits:
     
$$\begin{aligned} \text{logit}_{AVCD} = & (2 + \alpha^v + \alpha^a)\text{logit} \\ & + (1 - \alpha^v + \alpha^a)\text{logit}^v \\ & + (1 + \alpha^v - \alpha^a)\text{logit}^a \\ & - (\alpha^v + \alpha^a)\text{logit}^{v,a} \end{aligned}$$

14:  Append top token:  $\hat{y}_t \leftarrow \text{arg max logit}_{AVCD}$ 
15:  Append  $\hat{y}_t$  to  $\mathbf{y}$ 
16: end while
17: return  $\mathbf{y}$ 
```

---

1028 by avoiding the need for computationally intensive retraining, which is particularly beneficial when  
1029 scaling AV-LLMs to real-world deployment scenarios. Overall, AVCD facilitates broader and more  
1030 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 [27] 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.

**Prompt:** Is the athlete wearing pant?

*\*Visual mismatch*

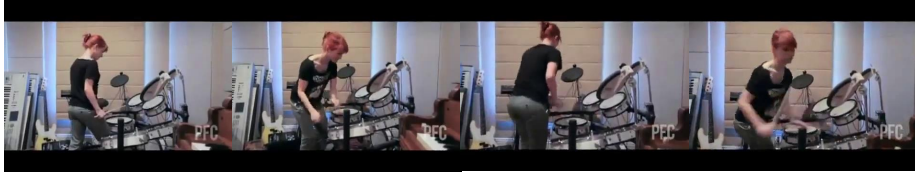


**Base**  
The athlete is **wearing a pant**.

**VCD**  
The athlete is **wearing red pant**.

**AVCD**  
**No**, the athlete is not wearing pant.  
He is wearing shorts.

**Prompt:** Does the person in black wear a hat?



**Base**  
Yes, the person in black **is wearing a hat** in the video.

**VCD**  
Yes, the person in black **is wearing a hat**.

**AVCD**  
**No**, the person in black does not wear a hat.

**Prompt:** Is the person in white wearing glasses?



**Base**  
Yes, the person in white **is wearing glasses**.

**VCD**  
Yes, the person in white **is wearing glasses**.

**AVCD**  
**No**, the person in white is not wearing glasses.

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 [36].**

**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 [36].** AVCD generates more accurate image-based descriptions compared to the *base* decoding and VCD [27].