Game on Tree: Visual Hallucination Mitigation via Coarse-to-Fine View Tree and Game Theory

Xianwei Zhuang, Zhihong Zhu, Zhanpeng Chen, Yuxin Xie, Liming Liang, Yuexian Zou*
School of ECE, Peking University, China
{xwzhuang,zhihongzhu,troychen927,yuxinxie,limingliang}@stu.pku.edu.cn,
zouyx@pku.edu.cn

Abstract

Large Vision-Language Models (LVLMs) may produce outputs that are unfaithful to reality, also known as visual hallucinations (VH), which hinders their application in multimodal understanding and decision-making. In this work, we introduce a novel plug-and-play train-free decoding algorithm named Game and Tree based Hallucination Mitigation (GTHM), designed for mitigating VH. GTHM is inspired by empirical observations that the fuzziness of multi-granularity view perception exacerbates VH. Based on this, GTHM leverages visual information to construct a coarse-to-fine visual view tree (CFTree) that organizes visual objects, attributes, and relationships in a hierarchical manner. Additionally, we innovatively model the optimal visual-token matching process on the CFTree as the cooperative game. Specifically, we define the Tree-based Shapley Value (TSV) for each visual view on the CFTree to assess its significant contribution to the overall visual understanding, thereby determining the optimal visual granularity. Subsequently, we utilize the TSV as guidance to implement adaptive weight contrastive decoding to achieve vision-aware decoding. Extensive experiments on four popular benchmarks confirm the effectiveness of our GTHM in alleviating VH across different LVLM families without additional training or post-processing. Our code is published at https://github.com/mengchuang123/GTHM.

1 Introduction

With the development of large language models (LLMs), large vision-language models (LVLMs) have made significant progress in model architecture, training methods, and data diversity (Liu et al., 2023c; Gong et al., 2023; Li et al., 2023a; Maaz et al., 2023; Zhang et al., 2023a; Zhu et al., 2023). LVLMs excel at converting complex visual patterns into coherent linguistic representations, leading to significant performance improvements in visual question answering (Bai et al., 2023; Dai et al., 2023; Liu et al., 2023b) and cross-modal understanding tasks (Xie et al., 2024; Zhuang et al., 2024e; Xin and Zou, 2023). However, LVLMs may produce outputs that are not faithful to reality, known as visual hallucinations (VH) (Gunjal et al., 2023; Li et al., 2023c; Liu et al., 2023a; Lovenia et al., 2023), which can impact their reliability and applicability across various domains. Recent research indicates that even more complex and powerful LVLMs cannot avoid VH (Dai et al., 2022; Li et al., 2023c; Guan et al., 2023).

In the context of LVLMs, extensive efforts have been dedicated to mitigating VH and enhance the reliability and fidelity of LVLM outputs. Current methods for alleviating VH generally can fall into three categories: post-processing (Zhou et al., 2023; Huang et al., 2023) and self-correction (Yin et al., 2023) techniques, fine-tuning based on instruction (Liu et al., 2023a; Yu et al., 2023), and decoding strategy approaches (Chuang et al., 2023; Leng et al., 2023; Chen et al., 2024). Despite some progress, these approaches still exhibit several limitations, including: (1) the potential requirement for additional datasets and training, or the incorporation of extra post-processing pipelines or more powerful external LVLMs (Zhou et al., 2023; Liu et al., 2023a; Yu et al., 2023); (2) a predominant focus on object hallucinations, often neglecting other visual elements such as relationships and attributes (Leng et al., 2023); (3) the necessity for time-consuming sampling processes for visual localization (Chen et al., 2024). Therefore, there remains an urgent need for more efficient methods to mitigate VH and achieve trustworthy LVLMs.

Empirically, as shown in Figure 1, we observe that providing misaligned or improperly granular views in the decoding of LVLMs can lead the model to focus more on language priors, thus ex-

*Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17984–18003
November 12-16, 2024

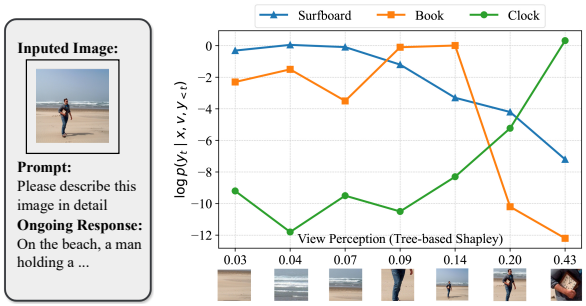


Figure 1: Experiment using LLaVa-1.5 shows that the inability to perceive the optimal view during decoding can exacerbate VH and produce incorrect tokens, i.e., 'Surfboard' and 'Book', where the lower proposed tree-based Shapley values represent poorer view perception.

acerbating VH. We provide more analysis in Section 2. Consequently, we derive a critical insight: assisting LVLMs in dynamically perceiving visual views at different granularities during autoregressive decoding will help LVLMs understand specific visual objects and relationships, thereby alleviating VH. Based on the above observations, we innovatively propose a framework for VH mitigation termed GTHM. Specifically, GTHM facilitates VH mitigation through several novel strategies:

First, we organize VH trigger words, including entities, attributes and relations, into a coarse-to-fine visual view tree (CFTree), which serves as structured data for optimal token querying to assist LVLMs in paying attention to visual context in autoregressive decoding. Our CFTree comprises three coarse-to-fine hierarchical levels: event, relation, and entity, which enhance the efficiency of perceiving optimal visual regions without the need for time-consuming sampling. **Second**, we adopt a game-theoretic perspective to achieve the perception of optimal visual views. Specifically, the search for optimal visual views is modeled as a cooperative game, where we innovatively define the tree-based Shapley values (TSV) to assess the contribution of each visual view from CFTree in achieving the overall visual-token match. **Third**, we implement adaptive contrastive encoding based on game scores to reduce unfaithful tokens. We perform contrastive decoding guided by TSV values for tokens with different significance scores.

In experiments, our GTHM achieves superior performance compared to existing methods on four popular VH benchmarks and across three LVLM families. In summary, the contributions of this paper are three-fold: (1) We propose a training-free, plug-and-play framework called GTHM, based on our CFTree, for efficient VH mitigation. Our GTHM is inspired by empirical observations that the fuzziness of multi-granularity view perception exacerbates VH. (2) We innovatively model view search for decoding optimal tokens as a cooperative game, and achieve adaptive game-augmented contrastive decoding via the proposed TSV for effectively mitigating VH. (3) Extensive experiments and comprehensive evaluations confirm that our GTHM significantly outperforms existing methods.

2 Preliminaries and Motivation

Problem Formulation. We consider a general LVLM, symbolized as θ, which is designed with an architecture that integrates a vision encoder, a vision-text interface, and a text decoder. Initially, visual information v undergoes processing through the vision encoder to produce a visual embedding, which is then modified by the vision-text interface to align with the textual query x. The combined data serves as input to the text decoder, which autoregressively generates a textual output y as:

$$y_t \sim p_\theta(y_t|v, x, y_{<t}), \tag{1}$$

$$p_\theta(y_t|v, x, y_{<t}) \propto \exp\left(\mathrm{f}_\theta(y_t|v, x, y_{<t})\right), \tag{2}$$

where, y_t represents the t-th token of y, while $y_{<t}$ refers to the sequence of tokens generated prior to t-th step. The function f_θ is the logit distribution function. When the generated token y_t does not align with the input image v, VH occurs, which can distort the actual visual content.

Hallucination Analysis. As demonstrated by the empirical results in Figure 2, we utilize the proposed TSV (*Def.* in 2)to measure visual perception and analyze the generation of VH from the perspective of multi-granularity visual views. Based on the results, we have made the following observations: (1) Captions without VH have higher TSV values. Inaccurate visual perception (lower TSV) corresponds to higher VH. (2) The longer the degree of sentence generation, the lower the TSV value, and the higher the likelihood of VH occurring. This is consistent with the discovery that excessive reliance on long-history tokens may also lead to the direction of VH (Zhou et al., 2023). (3) Similar to Favero et al. (2024), we use the prompt dependency measure (PDM) (*Def.* in Favero et al. (2024)) to measure the degree of VH. We observe that outputs with higher TSV values correspond to

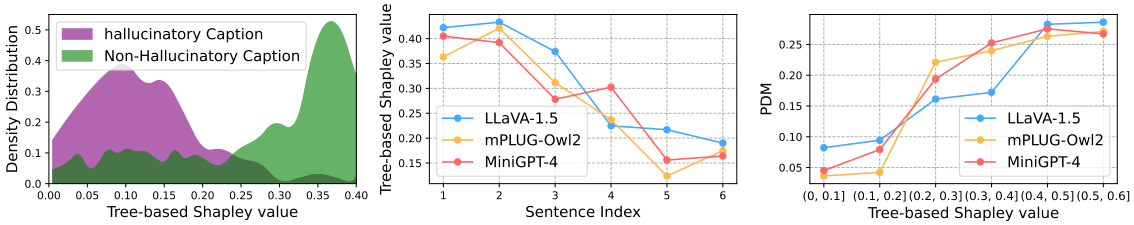


Figure 2: Analysis of the outputs of different LVLMs on sampled 500 images in the MSCOCO validation set, consisting of (1) TSV distribution for different outputs of LLaVA-1.5, (2) the relationship between sentence length and TSV, and (3) the relationship between prompt dependency measure (PDM) (Favero et al., 2024) and TSV.

higher PDM, indicating that higher TSV values provide greater visual benefits, thereby reducing VH. These analyses suggest that if LVLM fails to accurately focus on the visual view related to the token y_t, it tends to maximize the likelihood based on the textual prompts x and historical tokens $y_{<t}$. This results in the generation of tokens that align with the statistical probability of the language model but contradict the actual visual information. These motivated us to assist LVLM in dynamically perceiving visual views under different granularities and understanding specific visual objects and relationships during autoregressive decoding.

3 Methodology

3.1 Coarse-to-Fine Visual View Tree (CFTree)

Construction of our CFTree. As shown in Figure 3, we organize the CFTree $\mathcal{T}$ into a three-level hierarchical structure comprising event, relation, and entity layers to facilitate the structuring of multi-granularity views, which unify all visual elements prone to VH. The event layer serves as the root of the CFTree and represents the global visual scene, i.e., the inputted entire image. The entity layer consists of all leaf nodes of the CFTree, where each node represents the finest granularity of view. The relation layer represents the combinations of views depicted by each pair of leaf nodes. Initially, we adopt RAM (Zhang et al., 2023b) and GroundingDINO (Liu et al., 2023d) to extract all entity tags and their bounding boxes from the image, forming the leaf nodes of the CFTree. Subsequently, we pair the entities and expand their bounding boxes to form relation nodes (parent nodes). Finally, we connect the three layers to complete the CFTree, which presents a structured view from coarse to fine granularity.

Definition 1 ***(View Paths)****: Given a CFTree $\mathcal{T}$, the visual view path $\mathcal{P}(v_i)$ of a node $v_i \in \mathcal{T}$ is defined as the set of nodes within the path from the root node v_0 to v_i.*

Nodes on the view path $\mathcal{P}(v_i)$ (see Def. 1) represent a range of perspectives from coarse to fine for the same view v_i. Intuitively, the views that optimally match the token y_t are most likely to be found within the same view path $\mathcal{P}$.

3.2 CFTree-based Game Modeling

We employ cooperative game theory to search for the optimal visual context for decoding on the CFTree, which offers a hierarchical and meaningful reference derived from the original input image v_i. We start by introducing notation about the game theory, and then propose our modeling method.

Preliminaries. Cooperative game theory fundamentally explores how agents set of players $\mathcal{U}$ with a set function $f(\cdot)$ collaborate to maximize shared outcomes and allocate rewards based on individual contributions(Grabisch and Roubens, 1999; Sun et al., 2020). The Shapley value (Kuhn et al., 1953; Shapley, 1988) is a classical game solution with underlying axiom systems for the unbiased estimation of the contribution of each player. Formally, given a player groups $\mathcal{U}$, a subset coalition $\mathcal{S} \subseteq \mathcal{U}$ and a set function $f(\cdot)$ used to evaluate coalition game scores, the Shapley value $\phi(i|\mathcal{U})$ of the player $i \in \mathcal{U}$ is defined as:

$$\phi(i \mid \mathcal{U}) = \sum_{\mathcal{S} \subseteq \mathcal{U} \setminus \{i\}} p(\mathcal{S})[f(\mathcal{S} \cup \{i\}) - f(\mathcal{S})],$$
$$p(\mathcal{S}) = \frac{|\mathcal{S}|!(|\mathcal{U}| - |\mathcal{S}| - 1)!}{|\mathcal{U}|!}, \tag{3}$$

where, $p(\mathcal{S})$ is the likelihood of $\mathcal{S}$ being sampled. In our work, we estimate the saliency of visual understanding for each CFTree node based on the modified Shapley value to obtain the optimal view.

Game Modeling. Our objective is to quantitatively assess the contribution of view v_i to the

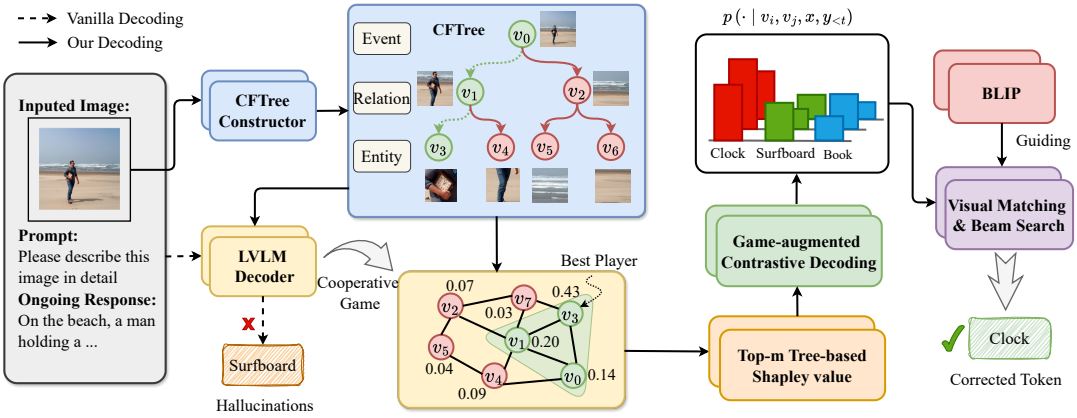


Figure 3: The illustration of the proposed GTHM framework, consisting of (1) the coarse-to-fine view tree designed to organize visual elements of different granularities, (2) game modeling with the proposed TSV to evaluate visual perception level, and (3) the adaptive game-augmented contrastive decoding focuses on more faithful decoding.

LVLM's understanding of visual scenes. Consequently, we model the view node v_i within $\mathcal{T}$ as players and the node set $\mathcal{T}$ as whole coalitions in a cooperative game. Inspired by (Deng et al., 2024), which uses CLIP to evaluate the matching scores between tokens and visual information, we employ BLIP (Li et al., 2022b) to compute the similarity scores between views and textual tokens. Thus, we define the set function for the game as

$$f(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{v_i \in \mathcal{S}} \text{BLIP}(v_i, p \oplus y_t^i), \quad y_t^i \sim p_\theta(y_t^i | v_i, x, y_{<t}), \tag{4}$$

where $\mathcal{S}$ is a player coalition, $p \oplus y_t^i$ denotes merging prompts and the token generated by LVLM.

Definition 2 *(Tree-based Shapley value): Given a CFTree $\mathcal{T}$ and a player $v_i \in \mathcal{T}$, we we consider $[\mathcal{P}(v_i)]$ as a single hypothetical player, which is the union of the players in $\mathcal{P}(v_i)$ (Def. in 1), the tree-based Shapley value is defined as:*

$$\phi^{\mathcal{T}}(v_i) = \phi([\mathcal{P}(v_i)] \mid \mathcal{T} \backslash \mathcal{P}(v_i) \cup \{[\mathcal{P}(v_i)]\}) - \sum_{v_j \in \mathcal{P}(v_i)} \phi(v_j \mid \mathcal{T} \backslash \mathcal{P}(v_i) \cup \{v_j\}), \tag{5}$$

where ϕ is vanilla Shapley value in Eq. 3.

Intuitively, our TSV is equivalent to the total benefit of including the entire visual path of view v_i minus the benefits of the views within $\mathcal{P}(v_i)$. We utilize our TSV to measure the significance of coarse-to-fine visual information within v_i on the LVLM's perception of visual semantics, i.e.,, assessing the consistency between the next token and the actual view v_i. Formally, our TSV are akin to Shapley interactions, but we incorporate constraints based on the tree structure. Additional properties and theoretical analysis are provided in the appendix.

Optimal Visual View-aware Context Candidates. Based on the proposed TSV, we calculate the salience scores for all nodes (views) in the CFTree. A higher salience score for a view indicates a superior visual-token alignment when tokens are generated guided by that granularity of view. We select the top-m views as candidates for the optimal visual context. Following prior work (Chen et al., 2024), we obtain the part-of-speech (POS) tags (Honnibal and Montani, 2017) for the currently generated tokens and calculate the TSV only for those tokens corresponding to objects, attributes, and relations to optimize time efficiency. In practice, we adopt the depth-first search algorithm to traverse CFTree to obtain the TSV of all views, detailed in Alg. 1.

3.3 Vision-aware Contrastive Decoding

Following the approach described in Section 3.2, we calculate the TSV for all views within the CFTree. We then compute the difference in TSVs between each pair of views as $d(v_i, v_j) = |\phi^{\mathcal{T}}(v_i) - \phi^{\mathcal{T}}(v_j)|$, and identify the top-m pairs of visual contexts with the most TSVs discrepancies. Subsequently, we amplify the informational contrast between the visual contexts by contrasting decoding probability distributions of each pair (vi, vj) and redistributing the output in log space (Li et al., 2022c). Unlike vanilla contrastive decoding strategies that employ a constant factor, our proposed adaptive visual contrastive decoding strategy utilizes the ratio of TSVs as a contrast

factor to redistribute probabilities:

$$p\left(\cdot \mid v_i, v_j, x, y_{<t}\right) \propto \exp[(1+\lambda_\phi)\cdot \\ \mathrm{f}_\theta\left(\cdot \mid v_i, x, y_{<t}\right) - \lambda_\phi \cdot \mathrm{f}_\theta\left(\cdot \mid v_j, x, y_{<t}\right)], \tag{6}$$

where, $\lambda_\phi = \lambda \frac{\phi^\tau(v_j)}{\phi^\tau(v_i)}$ It is an adaptive scaling factor and λ is a trade-off hyperparameter, $\phi^\tau(v_i) > \phi^\tau(v_j)$ and f_θ is the logit function. Through game-augmented contrastive decoding in Eq. 6, we identify m candidates for optimal visual view contexts. Ultimately, we employ the BLIP to compare the global visual-text similarity between the current text sequence $y_{\leq t}$ and the original image v. and integrate with beam search following (Chen et al., 2024) to obtain the optimal token.

3.4 Theoretical Analysis of our GTHM

Theorem 1 *(Game Theory) Our tree-base Shapley value (Def. in Eq. 5) satisfies the following axioms:* ***Linearity***, ***Symmetry***, ***Dummy***, *and* ***Recursivity***.

Intuition: This result indicates that the visual perception scores derived from our TSV possess properties not found in conventional visual-text contrastive methods. In summary, our TSV introduces an appealing aspect: **Coarse-to-fine receptive field**: TSV measures the fair contribution of a set of view paths to total benefits (i.e., the visual perception of LVLMs), rather than operating on a single view. This also enhances the robustness to biases in view elements during decoding.

Theorem 2 *(Information Theory) The CFTree consists of coarse-to-fine three hierarchical levels: event X_e, relationship X_r, and entity X_a. The level X_r depends on X_e, and the level X_a depends on X_r. By using our GTHM, additional information $\mathcal{I}(X_e, X_r) + \mathcal{I}(X_r, X_a)$ is introduced compared to vanilla decodings. This also reduces the overall decoding prediction error P_ϕ of X_r and X_a under the constraints of the given tree hierarchy:*

$$P_\phi \approx \frac{\mathcal{H}(X_r, X_a \mid X_e)}{\log(|\mathcal{X}|-1) + \log e}, \tag{7}$$

where $\mathcal{I}(\cdot)$ denotes mutual information, $\mathcal{H}(\cdot)$ represents the calculation of information entropy, and $\mathcal{X}$ is the value space of all random variables.

Intuition: According to Theorem 2, we can analyze that GTHM inherently introduces additional hierarchical and structural constraints. By performing decoding in our CFTree, it reduces the decoding prediction error of X_r and X_a, thereby reducing the VH of relationships, attributes, and entities.

4 Experiments

Benchmarks. Following common settings (Leng et al., 2023; Chen et al., 2024; Yin et al., 2023), We evaluate the effectiveness of our GTHM in VH mitigation on four popular benchmarks: (1) quantitative metrics CHAIR (Rohrbach et al., 2018) on MSCOCO dataset (Lin et al., 2014); (2) the Polling-based Object Probing Evaluation (POPE) (Li et al., 2023c) on the MSCOCO dataset; (3) general-purposed Multimodal Large Language Model Evaluation (MME) benchmark (Fu et al., 2023); (4) qualitative evaluation benchmark LLaVABench (Liu et al., 2023b).

CHAIR evaluates how frequently objects mentioned in a caption do not actually appear in the provided labels, thereby quantifying OH (Rohrbach et al., 2018). CHAIR quantifies hallucination at the sentence and instance levels through CHAIR_s and CHAIR_i metrics respectively, where higher values indicate more severe VH. POPE (Li et al., 2023c) employs a streamlined methodology to measure VH, which provides three distinct sampling methods: *random*, *popular*, and *adversarial* options. The MME benchmark (Fu et al., 2023) is a versatile tool crafted to evaluate and compare multimodal LLMs quantitatively.

Baselines. We adopt regular greedy decoding and beam search decoding methods, and various state-of-the-art (SOTA) decoding methods as baselines, including DoLa (Chuang et al., 2023), OPERA (Huang et al., 2023), VCD (Leng et al., 2023), Woodpecker (Yin et al., 2023), LURE (Zhou et al., 2023), and HALC (Chen et al., 2024). We implement these baselines based on HALC and compare them with our GTHM under the same benchmarks and environment settings.

Backbones. Following previous settings (Leng et al., 2023; Chen et al., 2024), we select popular LVLMs LLaVA-1.5 (Liu et al., 2023c), MiniGPT-4 (Chen et al., 2023) and mPLUG-Owl2 (Ye et al., 2023) as the backbones for all baselines except Woodpecker and LURE, where, Woodpecker and LURE utilize extra LLMs, i.e., ChatGPT (Brown et al., 2020) and GPT-4 (Achiam and Steven Adler, 2023), for self-correction and distillation. We investigate the hallucinations of these LVLMs under different decoding and post-processing strategies to evaluate the effectiveness of our GTHM.

Settings. In the experiment, we adopt Grounded-SAM (Ren et al., 2024)[1] with

[1] https://github.com/IDEA-Research/

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	BLEU↑	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	BLEU↑	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	BLEU↑
Greedy	22.17	7.23	16.24	29.36	11.80	14.57	26.10	8.27	15.48
Beam Search	19.45	6.25	16.44	27.94	11.20	14.88	22.33	7.79	15.86
DoLA	21.82	7.79	15.77	30.15	12.04	14.90	25.76	8.10	15.36
OPERA	22.48	8.32	15.60	29.64	11.98	14.83	22.49	7.57	15.57
VCD	21.24	7.68	16.20	30.24	12.25	14.51	26.41	9.35	14.81
Woodpecker	20.33	7.30	**17.20**	27.93	10.80	**15.62**	26.97	9.19	**16.57**
LURE	19.65	6.71	16.56	27.41	10.60	15.16	21.45	7.79	15.66
HALC	14.33	6.17	15.92	19.43	8.66	14.73	19.67	7.74	15.89
GTHM	**12.67**	**5.10**	16.72	**16.53**	**8.07**	14.98	**17.20**	**7.23**	16.06

Table 1: Comparison of the mean of five CHAIR evaluation results with different SOTA decoding baselines and our GTHM on MSCOCO datasets, with whole statistical results in Appendix.

RAM (Zhang et al., 2023b) to construct our three-layer CFTree. We utilize HALC (Chen et al., 2024) baseline code [2] based on HuggingFace TransformersRepository (Wolf et al., 2019) to implement our algorithm. The hyperparameter of top-m in Section 3.3 can be set to 2 to achieve a trade-off between efficiency and performance. The decoding process of LVLM and all experiments are performed on 8 A100 GPUs. More details and results are provided in the appendix.

4.1 Main Results on CHAIR and POPE

Following established evaluation protocols (Huang et al., 2023; Yin et al., 2023; Chen et al., 2024), we conduct the CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023c) evaluations on a randomly selected subset of 500 images from the MSCOCO validation set. We perform five experimental runs with different random seeds and report the statistical mean and standard deviation (more statistical results are provided in the Appendix).

CHAIR Evaluation. Following HALC (Chen et al., 2024), we set '*Please describe this image in detail.*' as the input prompt and utilize BLEU to evaluate the quality of text generation, as results are shown in Table 1. And we have more detailed observations: (1) Our GTHM significantly reduces VH at both the sentence and instance levels across different families of LVLMs. It can be observed that GTHM markedly outperforms existing decoding and post-processing baselines. For instance, using LLaVA-1.5 as the backbone, GTHM reduces sentence-level and instance-level VH by 13.1% and 21.0%, respectively, compared to the SOTA HALC. This demonstrates the efficacy and generalizability of our GTHM in alleviating VH. (2) GTHM maintains high-quality sentence generation without reorganization or reprocessing via extra LLMs. Compared to other decoding methods, our GTHM not only achieves optimal VH mitigation but also maintains a high BLEU score. Woodpecker (Yin et al., 2023) exhibits the highest sentence generation quality, attributed to its use of ChatGPT for post-processing.

POPE Evaluation. Following HALC (Chen et al., 2024), we utilize offline POPE (OPOPE) benchmark with accuracy, precision and $F_\beta = 0.2$ as metrics to evaluate VH, which replaces the live interactions of POPE with offline checks. As results shown in Table 2, we have several observations: (1) Our GTHM outperforms other SOTA baselines across most metrics. Averaging results over five random runs, GTHM consistently achieves optimal results in the majority of settings, which further demonstrates the effectiveness of GTHM in mitigating VH. (2) Our approach effectively mitigates VH across three different LVLM architectures, which illustrates the versatility and plug-and-play nature of our GTHM.

4.2 Main Results on the MME Benchmark

Following (Yin et al., 2023; Leng et al., 2023; Chen et al., 2024), we utilize object-level subsets (i.e., "existence" and "count") and attribute-level subsets (i.e., "position" and "color") to evaluate VH, with results shown in Table 3 and the whole results shown in Appendix. We can observe that: (1) GTHM significantly reduces object and attribute hallucination across a range of subsets of MME. Our GTHM consistently achieves optimal VH mitigation performance on each MME subset. This, consistent with Section 4.1, demonstrates that our

Grounded-Segment-Anything

[2] https://github.com/BillChan226/HALC

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	Acc.↑	Prec.↑	$F_{\beta=0.2}$↑	Acc.↑	Prec.↑	$F_{\beta=0.2}$↑	Acc.↑	Prec.↑	$F_{\beta=0.2}$↑↑
Greedy	74.24	97.48	92.45	69.43	96.96	91.09	73.06	96.50	92.18
Beam Search	72.35	97.79	91.80	69.32	97.09	90.39	72.35	97.10	92.34
DoLA	**74.47**	97.01	93.17	**70.50**	97.18	91.20	73.26	96.77	92.26
OPERA	72.42	96.82	91.14	69.88	96.60	91.43	71.56	97.27	92.10
VCD	73.75	96.92	92.78	69.02	96.66	90.79	71.60	97.53	93.10
Woodpecker	71.10	94.82	91.06	70.17	97.40	91.67	72.69	97.71	93.25
LURE	71.98	97.34	92.16	70.68	97.15	91.21	72.41	96.85	93.05
HALC	72.19	97.44	93.01	69.83	97.88	91.76	72.65	97.13	93.20
GTHM (Ours)	74.01	**98.30**	**93.78**	70.30	**97.90**	**92.15**	**73.48**	**97.80**	**93.49**

Table 2: Comparison of the mean of five OPOPE results on MSCOCO dataset with different decoding baselines under the '*random*' setting. Higher accuracy (Acc.), precision (Prec.), and F-score ($F_{\beta=0.2}$) indicate better performance. Bold indicates the best results. More results including *Popular* and *adversarial* settings are provided in Appendix.

Decoding	LLaVA-1.5				MiniGPT-4			
	Object-level↑		Attribute-level↑		Object-level↑		Attribute-level↑	
	Existence	Count	Position	Color	Existence	Count	Position	Color
Greedy	170.00	121.33	115.00	152.33	140.00	91.67	72.00	121.00
DoLa	172.33	120.00	106.67	150.00	135.00	92.33	68.33	121.00
OPERA	165.00	116.00	104.00	149.00	142.67	90.00	70.00	120.00
VCD	180.33	131.67	125.00	155.00	145.00	96.67	73.33	129.00
LURE	167.67	118.00	108.00	138.67	145.00	82.00	70.00	114.67
HALC	185.00	138.00	126.67	158.33	150.00	102.00	75.00	135.00
GTHM (Ours)	**191.67**	**147.67**	**135.00**	**165.00**	**160.00**	**110.00**	**82.00**	**138.67**

Table 3: Results on the hallucination subset of MME. The best performances within each setting are bolded. The whole results, including those on the mPLUG-Owl2, are presented in Appendix.

method can achieve comprehensive performance gains while ensuring efficient text generation quality. (2) LVLMs exhibit relatively lower evaluation scores on positional and counting hallucinations, indicating potential limitations in their ability to understand visual positions and perform counting reasoning.

4.3 More Analysis and Ablation Experiments

We conduct ablation experiments on our CFTree and game strategy using CHAIR on MSCOCO to evaluate the effectiveness of the components of our proposed method in detail. Specifically, we evaluate the effectiveness of the components by removing or modifying the specific settings of CFTree and game strategies, as results shown in Table 4.

Effect of the CFTree. As shown in Groups 1 and 4 in Table 4, removing any layer of CFTree (relation and entity layers) will result in a significant decrease in performance. We observe that removing the entity layer results in a greater performance decline than removing the relation layer. This may be because LVLMs exhibit a higher degree of hallucination at the entity level than at the relation level. Moreover, by replacing our CFTree with a randomly selected view approach, a significant performance drop can be seen, even falling below most baselines. This further illustrates that incorrect perception of visual views during autoregressive decoding exacerbates VH.

Effect of our Proposed TSV. As shown in Groups 2 and 4 in Table 4, we replace our TSV with BLIP similarity scores for views and tokens, as well as with standard Shapley values, to conduct ablation studies. We observe that our TSV achieves better view-aware evaluation than the other two methods, resulting in optimal VH mitigation performance. This is because our TSV leverages the hierarchical organization of views in our CFTree, facilitating the comprehensive evaluation of token and view path correlations and thus yielding superior view evaluation outcomes.

Effect of our Game-augmented Contrastive Decoding. To assess the effectiveness of our game-augmented adaptive contrastive decoding, we conduct ablation studies by removing Eq. 6 and re-

Group	Settings	LLaVA-1.5			MiniGPT-4		
		$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	BLEU↑	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	BLEU↑
1	Random Sample Views w/o CFTree	22.30	7.81	16.06	31.68	12.89	14.60
	Our CFTree w/o Relation Layer	15.31	6.59	15.83	20.46	9.22	14.35
	Our CFTree w/o Entity Layer	17.35	7.11	16.62	25.15	10.24	14.52
2	BLIP Similarity Score Replaces TSV	16.28	6.44	16.30	23.42	9.70	15.10
	Vanilla Shapley value Replaces TSV	15.20	6.06	16.13	21.74	8.56	**15.69**
3	w/o Game-augmented Contrastive Decoding	19.24	6.86	15.99	26.51	11.02	15.03
	Fix Factor 0.05 Replace $\phi^\tau(v_j)/\phi^\tau(v_i)$ in Eq.6	14.18	5.85	16.37	19.45	8.89	15.21
4	Our Full GTHM	**12.67**	**5.10**	**16.72**	**16.53**	**8.07**	14.98

Table 4: Ablation experiments on the CHAIR benchmark with the best results highlighted in bold.

placing the adaptive factor with a fixed factor. The results are shown in Table 4. We observe that removing game-augmented contrastive decoding significantly increases VH. We observe that when the adaptive factor is replaced with a fixed amplification factor of $\alpha = 0.05$, the ability to alleviate VH experiences a slight decrease. The ablation study results further validate the rationality and effectiveness of our game-augmented contrastive decoding.

Methods	Avg. Latency (ms/token)	
Greedy	51.53	1.00×
DoLa	54.86	1.06×
VCD	102.43	1.99×
HALC	479.28	9.30×
GTHM	192.91	3.74×

Table 5: Decoding latency on CHAIR.

4.4 Decoding efficiency analysis on CHAIR

We measure the throughput of LLaVA-1.5-7b using different strategies on the CHAIR benchmark, with the results presented in Table 5. Our analysis does not include the throughput of methods requiring post-processing, which involve calling additional tools to process the output. Our GTHM demonstrates superior decoding speed and enhances the performance of VH mitigation compared to the baseline HALC, which relies on extensive random sampling for retrieving the optimal visual view. This is attributed to our CFTree's early organization of visual views, which greatly reduces the search space and thus has higher decoding efficiency. We further calculate that the average number of CFTree nodes corresponding to each image in our method on the CHAIR benchmark is 9.21. Although the calculation of the Shapley value based on game theory is computationally intensive, our limited number of game players (i.e., the number of the nodes in CFTree) keeps the overall decoding time cost at a low level.

5 Related Work

VH and its Evaluation. The multi-modal large language model shows powerful performance in a large number of cross-modal understanding tasks (Bai et al., 2023; Zhuang et al., 2024a; Dai et al., 2023; Zhuang et al., 2024b,d,c; Xin et al., 2024), but cannot avoid VH. VH refers specifically to outputs that include inaccurate object representations or unfaithful content. This phenomenon has been observed in both early BERT-based models (Li et al., 2019) and recent LVLMs (Maaz et al., 2023; Zhang et al., 2023a; Zhu et al., 2023). In the realm of LVLMs, extensive studies have delved into the evaluation and detection of VH (Li et al., 2023c; Wang et al., 2023; Lovenia et al., 2023). One of the most widely adopted benchmarks for assessing VH is CHAIR (Rohrbach et al., 2018), motivated by observations that existing metrics like CIDEr (Vedantam et al., 2014) may misrepresent the presence of VH. POPE (Li et al., 2023c) evaluates VH through a binary classification framework, using precision, recall, and accuracy. Furthermore, the HALC (Chen et al., 2024) proposes offline POPE (OPOPE) to improve the evaluation of VH. We incorporate these metrics along with the BLEU score (Papineni et al., 2002) to comprehensively assess the effectiveness of our GTHM.

VH Mitigation. Various strategies have been developed to mitigate VH. Current methods for alleviating VH generally can fall into three categories: post-processing (Zhou et al., 2023; Huang et al., 2023) and self-correction (Yin et al., 2023) techniques, human feedback and fine-tuning based on instruction (Liu et al., 2023a; Yu et al., 2023), and

decoding strategy approaches (Chuang et al., 2023; Leng et al., 2023; Chen et al., 2024). However, the first two strategies may require additional datasets and training, or may require merging additional post-processing pipelines or more powerful external LVLMs (Zhou et al., 2023; Liu et al., 2023a; Yu et al., 2023). Existing decoding-based strategies may only focus on the VH of object (Chuang et al., 2023; Leng et al., 2023), ignoring elements that are prone to VH such as attributes or relationships, or require time-consuming visual region localization sampling (Chen et al., 2024). Our work focuses on designing training-free, plug-and-play efficient decoding methods to achieve VH mitigation of objects, relationships, and attributes.

Game Theory. The fundamental principle of game theory is to distribute different payoffs to game participants fairly and reasonably (Grabisch and Roubens, 1999; Sun et al., 2020). The Shapley value (Kuhn et al., 1953; Shapley, 1988) is a classical game theory solution for the unbiased estimation of the contribution of each player in a cooperation game. The game theory has been extensively applied to model interpretability (Datta et al., 2016; Yang et al., 2022; Zhang et al., 2021), explicit credit assignment (Li et al., 2021) and cross-modal feature interactions (Jin et al., 2023; Li et al., 2023b, 2022a). Considering the theoretical completeness and interpretability of the Shapley value, we innovatively model the optimal view search as a cooperative game and propose a tree-based Shapley value to estimate the optimal view.

6 Conclusions

This work proposes a training-free, plug-and-play decoding strategy to mitigate VH in LVLMs. Our method was inspired by practical observations that ambiguity in multi-granularity view perception exacerbates VH. Based on these, we construct the CFTree to organize multi-granularity views. We innovatively model the retrieving optimal multi-granularity views as a cooperative game. Subsequently, we perform adaptive contrastive decoding based on game scores to achieve bias-free distribution. Comprehensive experiments demonstrate the effectiveness of our GTHM in reducing VH across different benchmarks and LVLM families.

Limitations

Our work focuses on dynamically providing optimal visual views during the decoding process of LVLMs to enhance decoding accuracy and mitigate VH. Our visual views and CFTree are primarily constructed based on existing tools such as object detection. However, when hallucinated tokens cannot be detected or recognized by these tools, our method fails to provide suitable visual granularity for decoding, resulting in hallucinations. In future work, we will explore more effective visual dynamic perception methods to overcome these limitations.

Ethics Statement

The main research objects of this work are LVLMs, which may have uncontrollable or disloyal outputs. However, our work aims to help these LVLMs avoid these disloyal outputs and eliminate VH, which is completely in line with the ethical review. Moreover, we conduct all experiments on the public datasets, which do not contain any offensive content or information with negative social impact.

References

OpenAI Josh Achiam and et al. Steven Adler. 2023. Gpt-4 technical report.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv*, abs/2310.09478.

Zhaorun Chen, Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *ArXiv*, abs/2403.00425.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola:

Decoding by contrasting layers improves factuality in large language models. *ArXiv*, abs/2309.03883.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *ArXiv*, abs/2210.07688.

Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *Preprint*, arXiv:2402.15300.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefan 0 Soatto. 2024. Multi-modal hallucination control by visual information grounding. *ArXiv*, abs/2403.14003.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qianmengke Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *ArXiv*, abs/2305.04790.

Michel Grabisch and Marc Roubens. 1999. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models.

Anish Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. In *AAAI Conference on Artificial Intelligence*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Qidong Huang, Xiao wen Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Neng H. Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *ArXiv*, abs/2311.17911.

Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiang Ji, Li ming Yuan, and Jie Chen. 2023. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2482.

Harold W. Kuhn, A. W. Tucker, Melvin Dresher, Philip Wolfe, R. Duncan Luce, and H. Frederic Bohnenblust. 1953. Contributions to the theory of games.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Li Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *ArXiv*, abs/2311.16922.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726.

Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. 2023b. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11998–12008.

Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. 2021. Shapley counterfactual credits for multi-agent reinforcement learning. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. 2022a. Fine-grained semantically aligned vision-language pre-training. *ArXiv*, abs/2208.02515.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022c. Contrastive decoding: Open-ended text generation as optimization. In *Annual Meeting of the Association for Computational Linguistics*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. 2023d. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ArXiv*, abs/2310.05338.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv*, abs/2306.05424.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *ArXiv*, abs/2401.14159.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*.

Lloyd S. Shapley. 1988. A value for n-person games.

Jianyuan Sun, Hui Yu, Guoqiang Zhong, Junyu Dong, Shu Zhang, and Hongchuan Yu. 2020. Random shapley forests: Cooperative game-based random forests with consistency. *IEEE Transactions on Cybernetics*, 52:205–214.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Junyan Wang, Yi Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Mingshi Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models. *ArXiv*, abs/2308.15126.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yuxin Xie, Zhihong Zhu, Xianwei Zhuang, Liming Liang, Zhichang Wang, and Yuexian Zou. 2024. Gpa: Global and prototype alignment for audio-text retrieval. In *Interspeech 2024*, pages 5078–5082.

Yifei Xin, Xuxin Cheng, Zhihong Zhu, Xusheng Yang, and Yuexian Zou. 2024. Diffatr: Diffusion-based generative modeling for audio-text retrieval. In *Interspeech 2024*, pages 1670–1674.

Yifei Xin and Yuexian Zou. 2023. Improving audio-text retrieval via hierarchical cross-modal interaction and auxiliary captions. In *Proc. INTERSPEECH 2023*, pages 341–345.

Yu Yang, Seung Wook Kim, and Jungseock Joo. 2022. Explaining deep convolutional neural networks via latent visual-semantic filter attention. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8323–8333.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv*, abs/2311.04257.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xingguo Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *ArXiv*, abs/2310.16045.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *ArXiv*, abs/2312.00849.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. 2021. Interpreting multivariate shapley interactions in dnns. In *AAAI Conference on Artificial Intelligence*.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Siyi Liu, Yandong Guo, and Lei Zhang. 2023b. Recognize anything: A strong image tagging model. *ArXiv*, abs/2306.03514.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *ArXiv*, abs/2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

Xianwei Zhuang, Xuxin Cheng, Liming Liang, Yuxin Xie, Zhichang Wang, Zhiqi Huang, and Yuexian Zou. 2024a. PCAD: Towards ASR-robust spoken language understanding via prototype calibration and asymmetric decoupling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5235–5246, Bangkok, Thailand. Association for Computational Linguistics.

Xianwei Zhuang, Xuxin Cheng, Zhihong Zhu, Zhanpeng Chen, Hongxiang Li, and Yuexian Zou. 2024b. Towards multimodal-augmented pre-trained language models via self-balanced expectation-maximization iteration. In *ACM Multimedia 2024*.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024c. Towards explainable joint models via information theory for multiple intent detection and slot filling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19786–19794.

Xianwei Zhuang, Hongxiang Li, Xuxin Cheng, Zhihong Zhu, Yuxin Xie, and Yuexian Zou. 2024d. Kdpror: A knowledge-decoupling probabilistic framework for video-text retrieval. In *European Conference on Computer Vision*. Springer.

Xianwei Zhuang, Zhichang Wang, Xuxin Cheng, Yuxin Xie, Liming Liang, and Yuexian Zou. 2024e. MaCSC: Towards multimodal-augmented pre-trained language models via conceptual prototypes and self-balancing calibration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8077–8090, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Methods Details

We follow our baseline, i.e., HALC (Chen et al., 2024), and adopt a similar LVLM family, where the parameter size of the LVLM model is also consistent with that of HALC. λ in Eq. 6 is set to 0.02. For other experimental hyperparameters, we follow all the settings in the HALC benchmark [3]. We describe the details of calculating TSV values and the overall algorithm in our CFTree in Algorithm 1 and Algorithm 2, respectively.

A.2 Proofs of Theorem 1

Proof of Theorem 1 *Our tree-base Shapley value (Def. in Eq. 5) satisfies the following axioms:*

(1) Linearity. For a new game $w(i) = u(i) + v(i)$, $\phi_w^\tau(i) = \phi_u^\tau(i) + \phi_v^\tau(i)$;

(2) Symmetry. If $\forall \mathcal{S} \subseteq \mathcal{U} \setminus \{i, j\}$, $f(\mathcal{S} \cup \{i\}) = f(\mathcal{S} \cup \{j\})$, $\phi^\tau(i) = \phi^\tau(j)$;

(3) Dummy. If i is dummy, $\forall \mathcal{S} \subseteq \mathcal{U} \setminus \{i\}$, $f(\mathcal{S} \cup \{i\}) = f(\mathcal{S})$, $f(\{i\}) = \phi^\tau(i)$;

(4) Recursivity. If $\mathcal{S} = \{i, j\}$, $\phi(i \mid \mathcal{U} \setminus \mathcal{S} \cup \{[\mathcal{S}]\}) = \phi(i \mid \mathcal{U} \setminus \{i\}) + \phi(j \mid \mathcal{U} \setminus \{j\}) + \phi^\tau(i)$.

Proof. Our Tree-based Shapley value is the metric that satisfies the following axioms: Linearity, Symmetry, Dummy, and Recursivity:

Linearity Axiom. To prove this, we first apply the linearity property of Shapley values to the combined game w. We know from the properties of Shapley values that: $\phi_w^\tau(i) = \phi_u^\tau(i) + \phi_v^\tau(i)$ Apply this property directly to each term in the interaction formula:

$$\begin{aligned}\phi_w^\tau(i) = &\phi_w([\mathcal{P}(i)] \mid \mathcal{T}\backslash\mathcal{P}(i) \cup \{[\mathcal{P}(i)]\}) \\ &- \sum_{j \in \mathcal{P}(i)} \phi_w(j \mid \mathcal{T}\backslash\mathcal{P}(i) \cup \{j\}),\end{aligned} \tag{8}$$

Using the linearity of Shapley values, we can obtain:

$$\begin{aligned}\phi_w^\tau(i) = &\phi_u([\mathcal{P}(i)] \mid \mathcal{T}\backslash\mathcal{P}(i) \cup \{[\mathcal{P}(i)]\}) \\ &- \sum_{j \in \mathcal{P}(i)} \phi_u(j \mid \mathcal{T}\backslash\mathcal{P}(i) \cup \{j\}) \\ &+ \phi_v([\mathcal{P}(i)] \mid \mathcal{T}\backslash\mathcal{P}(i) \cup \{[\mathcal{P}(i)]\}) \\ &- \sum_{j \in \mathcal{P}(i)} \phi_v(j \mid \mathcal{T}\backslash\mathcal{P}(i) \cup \{j\}) \\ &= \phi_u^\tau(i) + \phi_v^\tau(i)\end{aligned} \tag{9}$$

Symmetry Axiom. Assume two players, i and j, and we need to show that if for all subsets $\mathcal{S} \subseteq$

[3] https://github.com/BillChan226/HALC

Algorithm 1 Depth-First Search (DFS) for Game Contribution Allocation on CFTree

Require: The node v_i in CFTree $\mathcal{T}$, the path of the current node $\mathcal{P}$, the set $\mathcal{O}$ used to store TSVs of all nodes, which is initialized to $\{\emptyset\}$.

Add the current node to the path $\mathcal{P} = \mathcal{P} \cup \{v_i\}$
Compute the TSV $\phi^{\tau}(v_i)$ of node v_i {§3.2, Eq. (5)}
Add current TSV to the results $\mathcal{O} = \mathcal{O} \cup \{\phi^{\tau}(v_i)\}$
for each child node v of the current node v_i **do**
  Recursive call DFS(v, $\mathcal{P}$)
end for
Remove the current node from the path $\mathcal{P} = \mathcal{P} \backslash \{v_i\}$

Algorithm 2 Complete GTHM Decoding Algorithm

Require: LVLM θ, question query x, image query v, the sequence of tokens generated prior to t-th step $y_{<t}$. Constructed CFTree $\mathcal{T}$ by v, beam size k, top-m candidate pool m.

repeat
  for $b = 1$ to beam size k **do**
    Autoregressive decoding by θ, obtain current token $w_i^b \sim p_\theta(y_t|v, x, y_{<t})$
    if POS tags $w_i^b \in \{\text{entity}, \text{attribute}, \text{relationship}\}$ **then**
      Obtain game contributions of all view nodes $\mathcal{O} = \text{DFS}(v_0, \{\emptyset\})$ {§A, Alg. (1)}
    else
      $\mathcal{O} = \{\emptyset\}$
    end if
    Calculate pair-wise $d(v_i, v_j) = |\phi^{\tau}(v_i) - \phi^{\tau}(v_j)|$ {§3.3}
    Select top-m candidate pairs {§3.3}
    for $i = 1$ to m **do**
      Apply game-augmented contrast $p\left(\cdot \mid v_i, v_j, x, y_{<t}\right)$
      get a redistributed logits {§3.3, Eq. (6)}
    end for{y_{new}^b with m candidates obtained}
  end for
  Select top k candidate responses by BLIP and beam search {§3.3}
  if $\mathcal{O} \neq \{\emptyset\}$ **and** $y_{\text{new}}^b = w_i^b$ **then**
    $y_{\text{new}}^b \leftarrow$ [IDK] {w_i^b is hallucinating, but no correct token was found}
  end if
  $w_i^b \leftarrow y_{\text{new}}^b$ {Hallucinating token w_i^b corrected}
until each beam has terminated

$\mathcal{U} \setminus \{i, j\}$, we have $f(\mathcal{S} \cup \{i\}) = f(\mathcal{S} \cup \{j\})$, then their interaction scores should also be equal, i.e., $\phi_w^{\tau}(i) = \phi_w^{\tau}(j)$. Under the assumption of symmetry, since i and j contribute equally to all possible coalition subsets, it implies that:

For all $\mathcal{S} \subseteq \mathcal{U} \setminus \{i, j\}$, within the Shapley value formula for each combination, ϕ should be equal for i and j. Specifically, for subsets $\mathcal{S} = \{i\}$ and $\mathcal{S} = \{j\}$, we have:

$$\phi(\{i\} \cup (N \setminus \{i\})) = \phi(\{j\} \cup (N \setminus \{j\})) \quad (10)$$

For $\sum_{j \in \mathcal{P}(i)} \phi_w(j \mid \mathcal{T} \backslash \mathcal{P}(i) \cup \{j\})$, since $\mathcal{S}$ contains only one player, this term will also be equal, as each player individually has the same impact on all other coalition subsets.

Thus, the interaction scores: $f(\{i\}) = \phi^{\tau}(i)$. This satisfies the Symmetry Axiom.

Dummy Axiom. Let i be a dummy player, $\forall \mathcal{S} \subseteq \mathcal{U} \backslash \{i\}$, $f(\mathcal{S} \cup \{i\}) = f(\mathcal{S})$. This means that joining a player i does not change the value of any subset of collaborators. Since $\forall \mathcal{S}, f(\mathcal{S} \cup \{i\}) = f(\mathcal{S})$, $f(\mathcal{S} \cup \{i\}) - f(\mathcal{S}) = 0$. Thus, we have $\phi(i) = 0$.

When i is a dummy player and $\mathcal{S}$ includes i,

$$\phi(\{S\} \cup N \backslash \mathcal{S}) = \phi(\{\mathcal{S} \backslash \{i\}\} \cup \mathcal{U} \backslash (\mathcal{S} \backslash \{i\})) \quad (11)$$

because adding i does not change the value of the set. Similarly, for each $i \in \mathcal{S}$, we have

$$\phi(\{i\} \cup \mathcal{U} \setminus \{i\}) = 0 \quad (12)$$

Therefore,

$$\phi^{\tau}(\mathcal{S}) = \phi(\{\mathcal{S} \setminus \{i\}\} \cup \mathcal{U} \setminus (\mathcal{S} \setminus \{i\})) - \sum_{j \in \mathcal{S} \setminus \{i\}} \phi(\{j\} \cup \mathcal{U} \setminus \{j\}) = \phi^{\tau}(\mathcal{S}\{i\}). \tag{13}$$

When $\mathcal{S}$ contains only the dummy player i, since $\phi(i) = 0$, it follows that $\phi^{\tau}(i) = 0$

Recursivity Axiom. By performing a simple numerical transformation on our TSV, we can obtain the form of the recursivity axiom:

$$\phi(i \mid \mathcal{U} \setminus \mathcal{S} \cup \{[\mathcal{S}]\}) = \phi(i \mid \mathcal{U} \setminus \{i\}) + \phi(j \mid \mathcal{U} \setminus \{j\}) + \phi^{\tau}(i). \tag{14}$$

□

A.3 Proofs of Theorem 2

Proof of Theorem 2 *Consider the CFTree, which consists of coarse-to-fine three hierarchical levels: event X_e, relationship X_r, and entity X_a. The level X_r depends on X_e, and the level X_a depends on X_r. By using our GTHM, additional information $\mathcal{I}(X_e, X_r) + \mathcal{I}(X_r, X_a)$ is introduced compared to vanilla decodings. This also reduces the overall decoding prediction error P_ϕ under the constraints of the given tree hierarchy:*

$$P_\phi \approx \frac{\mathcal{H}(X_r, X_a \mid X_e)}{\log(|\mathcal{X}| - 1) + \log e}, \tag{15}$$

where $\mathcal{I}(\cdot)$ denotes mutual information, $\mathcal{H}(\cdot)$ represents the calculation of information entropy, and $\mathcal{X}$ is the value space of all random variables.

Proof. Based on the hierarchical nature of CFTree, we can determine that each layer represents a random variable that conforms to Markov property, i.e $X_e \to X_r \to X_a$. For Markov chains $X_e \to X_r \to X_a$, we have $X_a \perp X_e \mid X_r$, i.e., given X_r, X_a and X_e are conditionally independent. $\mathcal{H}(X_r, X_a \mid X_e)$ represents the joint uncertainty of X_r and X_a given the known X_e. According to the definition of information entropy, we can deduce that the reduction in uncertainty of the relationship layer and entity layer after adding structured constraints is:

$$\begin{aligned} \mathcal{H}(X_r) - \mathcal{H}(X_r \mid X_e) &= \mathcal{I}(X_r; X_e). \\ \mathcal{H}(X_a) - \mathcal{H}(X_a \mid X_r) &= \mathcal{I}(X_r; X_a). \end{aligned} \tag{16}$$

Therefore, when using our GTHM for decoding on CFTree, the additional amount of information introduced is:

$$\begin{aligned} &\mathcal{H}(X_r) - \mathcal{H}(X_r \mid X_e) + \mathcal{H}(X_a) - \mathcal{H}(X_a \mid X_r) \\ &= \mathcal{I}(X_r; X_e) + \mathcal{I}(X_r; X_a). \end{aligned} \tag{17}$$

Using the definition of conditional entropy and the chain rule, we can obtain:

$$\mathcal{H}(X_r, X_a \mid X_e) = \mathcal{H}(X_r \mid X_e) + \mathcal{H}(X_a \mid X_r, X_e). \tag{18}$$

Since $X_a \perp X_e \mid X_r$, we have $\mathcal{H}(X_a \mid X_r, X_e) = \mathcal{H}(X_a \mid X_r)$. Thus, we can finally obtain:

$$\mathcal{H}(X_r, X_a \mid X_e) = \mathcal{H}(X_r \mid X_e) + \mathcal{H}(X_a \mid X_r). \tag{19}$$

Following the Fano's inequality, we can obtain:

$$\begin{aligned} \mathcal{H}(X_r \mid X_e) &\le H(P_r) + P_r \log(|\mathcal{X}| - 1), \\ \mathcal{H}(X_a \mid X_r) &\le H(P_a) + P_a \log(|\mathcal{X}| - 1) \end{aligned} \tag{20}$$

Following Eq. 19 and 20, we can obtain the overall decoding prediction error P_ϕ as:

$$P_\phi = P_r + P_a \approx \frac{\mathcal{H}(X_r, X_a \mid X_e)}{\log(|\mathcal{X}| - 1) + \log e}. \tag{21}$$

Due to the reduction of uncertainty, we essentially reduced the decoding error P_ϕ of the relationship layer and instance layer. □

A.4 More experimental results on POPE

We present the complete comparison results of our method and baselines on the POPE benchmark in Tables 6 and 7, and it can be seen that our GTHM outperforms the baselines in most of the metrics.

A.5 Statistical results on CHAIR

We perform five experimental runs with different random seeds and reported the statistical mean and standard deviation. The mean values are reported in Table 1 and the standard deviations are reported in Table 8.

A.6 LLaVA-Bench Qualitative Study

The LLaVA-Bench (Liu et al., 2023b) is a benchmark comprising 24 images, each associated with a detailed, manually crafted description and a set of carefully selected questions. Following (Yin et al., 2023; Leng et al., 2023; Chen et al., 2024), we utilize LLaVA-Bench as a benchmark to evaluate the intuitive effect of our method on specific VH mitigation. We use 'Please describe this image in detail.' as the prompt to query LVLM for captions, as the results are shown in Figure 4, 5, 6 and 7.

Settings	Model	Decoding	Accuracy	Precision	Recall	F_0.2Score
Random	LLaVA-1.5	Greedy	74.24	97.48	48.91	92.45
		Beam Search	72.35	97.79	46.52	91.80
		DoLA	74.47	97.01	48.90	93.17
		OPERA	72.42	96.82	46.33	91.14
		VCD	73.75	96.92	47.69	92.78
		Woodpecker	71.10	94.82	45.15	91.06
		LURE	71.98	97.34	45.89	92.16
		HALC	72.19	97.44	48.73	93.01
		GTHM (Ours)	74.01	98.30	49.03	93.78
	MiniGPT-4	Greedy	69.43	96.96	36.35	91.09
		Beam Search	69.32	97.09	38.10	90.39
		DoLA	70.50	97.18	38.04	91.20
		OPERA	69.88	96.60	38.10	91.43
		VCD	69.02	96.66	37.04	90.79
		Woodpecker	70.17	97.40	40.20	91.67
		LURE	70.68	97.15	40.69	91.21
		HALC	69.83	97.88	40.80	91.76
		GTHM (Ours)	70.30	97.90	41.03	92.15
	mPLUG-Owl2	Greedy	73.06	96.50	44.29	92.18
		Beam Search	72.35	97.10	43.87	92.34
		DoLA	73.26	96.77	44.54	92.26
		OPERA	71.56	97.27	42.83	92.10
		VCD	71.60	97.53	43.79	93.10
		Woodpecker	72.69	97.71	42.16	93.25
		LURE	72.41	96.85	43.06	93.05
		HALC	72.65	97.13	44.18	93.20
		GTHM (Ours)	73.48	97.80	44.27	93.49
Popular	LLaVA-1.5	Greedy	72.13	90.97	46.33	88.68
		Beam Search	71.49	91.62	44.95	87.39
		DoLA	72.56	90.55	46.21	88.94
		OPERA	71.53	90.47	45.84	87.46
		VCD	72.70	91.06	46.72	88.63
		Woodpecker	71.38	91.46	45.21	87.20
		LURE	70.53	90.90	45.63	87.08
		HALC	72.05	91.34	45.80	87.78
		GTHM (Ours)	72.68	91.81	46.20	88.36
	MiniGPT-4	Greedy	67.89	89.16	39.27	84.92
		Beam Search	68.58	90.59	39.84	85.77
		DoLA	68.53	90.92	39.41	85.28
		OPERA	68.04	89.48	39.66	84.81
		VCD	67.60	89.41	38.48	84.07
		Woodpecker	68.70	90.01	40.22	85.92
		LURE	68.62	90.24	40.59	86.18
		HALC	68.04	90.31	39.48	86.00
		GTHM (Ours)	68.93	90.98	40.19	86.26
	mPLUG-Owl2	Greedy	70.34	89.63	43.87	86.10
		Beam Search	69.26	89.79	43.12	86.55
		DoLA	70.57	89.41	44.02	85.84
		OPERA	69.84	91.43	42.35	86.99
		VCD	70.05	90.49	43.41	87.12
		Woodpecker	69.48	90.10	42.59	86.44
		LURE	70.40	90.44	43.74	86.46
		HALC	70.18	90.20	43.97	87.11
		GTHM (Ours)	70.33	91.85	44.27	87.73

Table 6: Comparison of the mean of five OPOPE results on MSCOCO dataset with different decoding baselines under the 'random' and 'popular' settings. Higher accuracy (Acc.), precision (Prec.), and F-score ($F_{\beta=0.2}$) indicate better performance.

Settings	Model	Decoding	Accuracy	Precision	Recall	F_0.2Score
Adversarial	LLaVA-1.5	Greedy	68.72	87.48	45.66	82.68
		Beam Search	67.35	86.98	44.01	82.23
		DoLA	68.43	86.13	45.27	82.43
		OPERA	67.61	85.88	44.30	83.18
		VCD	69.05	86.43	45.83	83.60
		Woodpecker	68.60	88.42	44.34	84.18
		LURE	68.27	87.40	44.69	83.42
		HALC	69.33	88.81	43.24	84.96
		GTHM (Ours)	69.86	89.47	44.88	85.30
	MiniGPT-4	Greedy	64.35	84.84	38.38	80.94
		Beam Search	65.49	86.22	39.42	82.60
		DoLA	66.37	85.24	38.65	81.11
		OPERA	65.10	85.83	39.59	82.04
		VCD	63.70	85.82	36.49	81.27
		Woodpecker	66.10	86.75	39.92	83.24
		LURE	66.82	86.33	40.37	83.38
		HALC	66.53	87.00	37.35	82.38
		GTHM (Ours)	66.34	87.66	39.99	83.46
	mPLUG-Owl2	Greedy	67.89	86.16	43.50	82.70
		Beam Search	67.31	87.24	43.53	83.19
		DoLA	67.68	87.60	43.38	82.95
		OPERA	68.06	88.98	42.06	85.30
		VCD	68.90	88.32	43.99	85.29
		Woodpecker	66.73	88.84	42.73	84.09
		LURE	67.83	86.38	43.00	82.48
		HALC	68.97	88.43	43.55	84.60
		GTHM (Ours)	69.64	88.83	43.52	85.03

Table 7: Comparison of the mean of five OPOPE results on MSCOCO dataset with different decoding baselines under the 'adversarial' setting. Higher accuracy (Acc.), precision (Prec.), and F-score ($F_{\beta=0.2}$) indicate better performance.

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	CHAIR_s	CHAIR_I	BLEU	CHAIR_s	CHAIR_I	BLEU	CHAIR_s	CHAIR_I	BLEU
Greedy	0.51	0.38	0.00	3.97	3.95	0.00	1.11	1.47	0.00
Beam Search	1.24	0.66	0.00	3.84	2.39	0.00	0.74	0.17	0.00
DoLA	0.95	0.96	0.00	1.58	0.46	0.00	1.50	0.75	0.00
OPERA	2.42	1.19	0.00	0.78	0.93	0.00	1.87	1.73	0.00
VCD	1.71	1.19	0.00	3.49	1.58	0.00	3.52	2.64	0.00
Woodpecker	3.24	2.60	0.03	1.14	0.81	0.01	2.65	1.49	0.00
LURE	1.66	1.31	0.04	1.23	1.60	0.01	1.80	1.24	0.06
HALC	0.15	0.64	0.00	0.16	0.50	0.00	0.92	0.32	0.00
GTHM	0.19	0.15	0.00	0.30	0.87	0.00	1.33	0.61	0.00

Table 8: Comparison of the standard deviations of five CHAIR evaluation results with different SOTA decoding baselines and our GTHM on MSCOCO datasets.

GTHM (Ours): The painting features a brown dog with a bandana over its eyes and ears, dressed in Renaissance garb. The dog is sitting on a bed, giving off an amusing yet artistic appearance.The background showcases mountains, creating a picturesque scene.

Greedy: The image features a dog wearing a large, flowing dress and a hat. The dog appears to be looking at something beyond the frame, possibly even focusing on the viewer. The dress is blue and envelops the dog, giving it a majestic appearance.

HALC: The image is a painting featuring a dog wearing a t-shirt and a veil. The dog's eyes are curiously inspecting the viewer. The painting is an interesting combination of an artwork and a portrait of a dog.

VCD: The image features a painting of a dog, dressed in Renaissance garb. In the foreground, a person can be seen sitting on a chair, with a potted plant located nearby, providing a sense of realism to the scene.

Figure 4: LLaVA-Bench results comparing our GTHM and other methods with LLaVA-1.5 backbone.

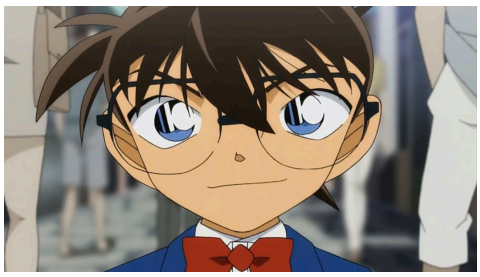
GTHM (Ours): The image features a group of three mugs, each one displaying a image of Mario, the iconic video game character. The mugs are placed side by side, varying in position, giving it an appealing and fun appearance.

Greedy: The image showcases a collection of coffee cups featuring the popular Mario character. There are four cups in total, each displaying Mario in different sizes, colors, and orientation. The cups are arranged in a row, showcasing their unique designs.

HALC: The image features a shelf with three white coffee mugs sitting side by side. The mugs are adorned with images of the same popular video game character, Mario. The mugs display close-ups of Mario, showcasing his iconic features.

VCD: The image features a display of four coffee cups designed with the popular Nintendo character, Mario. The cups are placed side by side and appear to be stacked. The Mario images are unique to each mug, creating a visually appealing display of the popular gaming character.

Figure 5: LLaVA-Bench results comparing our GTHM and other methods with LLaVA-1.5 backbone.



GTHM (Ours): The image features a man wearing a blue suit with a red bow tie. He appears to be posing and smiling for the camera. The man seems to be an anime character. He is wearing glasses, which emphasizes his distinct appearance.

Greedy: The image features a character with long hair, resembling a young boy or possibly a shaggy-haired man, wearing glasses and a bow tie. The character appears to be the protagonist of a cartoon or anime series, possibly "Ponch." The character's glasses are prominent, enhancing their unique style.

HALC: The image portrays a cartoon character, likely from the anime series "One Piece," dressed in an outfit featuring a red bow tie and eyeglasses. The character's appearance can be described as similar to a young boy. The character seems to be looking at the viewer in a smiling manner.

VCD: The scene features a cartoon character, likely Dracula or Johnny Sokko, with a red bow tie and glasses. The character appears to be smiling. In the background, there is a clock visible at the top-left corner, and a suitcase is placed in the lower-right corner of the image.

Figure 6: LLaVA-Bench results comparing our GTHM and other methods with MiniGPT-4 backbone.

GTHM (Ours): The scene features an artistic interpretation of a group of animals gathered around a dining table, enjoying a meal together. The animals include a rabbit, a cat, and a bear, all sitting on the grass by the table. A variety of food items can be seen on the table.

Greedy: The image features a painting of three animals, likely rabbits, sitting at a dining table together. The table is adorned with a cup, a spoon, and a bowl, giving the impression of a shared meal. In addition to the animals, there are two teddy bears seated at the table.

HALC: The image depicts an adorable scene of a bear, a cat, and a rabbit sitting around a small wooden dining table. There is a cup placed on the table, possibly containing a beverage for the group.

VCD: The image features a scene involving three animals sitting around a dining table. A large brown bear is sitting on one side, while a small kitten is sitting on the other side. A white rabbit is also present in the scene. The animals are enjoying a meal together, with a bowl and a spoon placed on the table.

Figure 7: LLaVA-Bench results comparing our GTHM and other methods with mPLUG-Owl2 backbone.