HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding

Zhaorun Chen [*1] **Zhuokai Zhao** [*1] **Hongyin Luo** [2] **Huaxiu Yao** [3] **Bo Li** [1 4] **Jiawei Zhou** [5]

Abstract

While large vision-language models (LVLMs) have demonstrated impressive capabilities in interpreting multi-modal contexts, they invariably suffer from object hallucinations (OH). We introduce **HALC**, a novel decoding algorithm designed to mitigate OH in LVLMs. HALC leverages distinct fine-grained optimal visual information in vision-language tasks and operates on both local and global contexts simultaneously. Specifically, HALC integrates a robust auto-focal grounding mechanism (locally) to correct hallucinated tokens on the fly, and a specialized beam search algorithm (globally) to significantly reduce OH while preserving text generation quality. Additionally, HALC can be integrated into any LVLMs as a plug-and-play module without extra training. Extensive experimental studies demonstrate the effectiveness of HALC in reducing OH, outperforming state-of-the-arts across four benchmarks. Code is released at https://github.com/BillChan226/HALC.

1 Introduction

The confluence of natural language processing (NLP) and computer vision (CV) has undergone a transformative shift over the past years with the introduction of vision-language models (VLMs) (Zhu et al., 2023; Liu et al., 2023b; Zhang et al., 2024). Although VLMs have shown exceptional proficiency in integrating and interpreting intricate data across both textual and visual modalities, a significant challenge emerged as the phenomenon of *object hallucination (OH)*, where VLMs erroneously generate hallucinated objects and descriptions within their outputs (Rohrbach et al., 2018). Based on the different parts of the sentences that are being hallucinated, OH can be categorized into three types: object *existence*, *attribute*, and *relationship* hallucinations (Gunjal et al., 2023; Zhai et al., 2023).

OH has been a persistent challenge since the earlier stages of the VLM development (Rohrbach et al., 2018). And it has been gaining increased attention, especially when recent research indicates that even the much more sophisticated and capable large vision-language models (LVLMs) are not immune to it (Dai et al., 2022; Li et al., 2023; Guan et al., 2023). Numerous efforts have been devoted to mitigating OH in the context of LVLMs, including a post-hoc approach that corrects the LVLM output after completion (Zhou et al., 2023), a self-correction pipeline for OH mitigation (Yin et al., 2023), and various decoding strategies that are tailored towards reducing OH via better textual or visual priors utilization (Huang et al., 2023; Leng et al., 2023).

Despite the efforts, these approaches are not yet fully satisfying in terms of eliminating OH. More importantly, they mainly focus on mitigating object existence hallucination, while assuming the attribute- and relationship-level hallucinations can be consequently corrected through autoregressive decoding. Furthermore, their reliance on more powerful external LVLMs (Yin et al., 2023), repeated processing (Zhou et al., 2023) or additional data (Gunjal et al., 2023) complicates their adaptations to existing LVLMs and restricts their use cases. The importance of OH reduction combined with the limitations in existing methods underscore the urgent need for developing novel approaches.

To this end, we introduce Object **H**allucination Reduction through **A**daptive Foca**L**-**C**ontrast decoding (**HALC**), a novel decoding strategy designed to effectively counter OH and can be easily integrated into any open-source LVLMs such as MiniGPT-4 (Chen et al., 2023), LLaVA (Liu et al., 2023b) and mPLUG-Owl2 (Ye et al., 2023). HALC addresses all three types of OH (existence, attribute, and relationship) while preserving generation quality in both local and global levels; locally, it employs an *adaptive focal-contrast grounding* mechanism to locate the fine-grained optimal visual information to correct each generated token that might be hallucinating; and globally, it incorporates a *matching-based beam search* that utilizes a visual matching

[*]Equal contribution [1]University of Chicago, Chicago IL, USA [2]Massachusetts Institute of Technology, Boston MA, USA [3]UNC-Chapel Hill, Chapel Hill NC, USA [4]University of Illinois at Urbana-Champaign, Champaign IL, USA [5]Toyota Technological Institute at Chicago, Chicago IL, USA. Correspondence to: Zhaorun Chen, Zhuokai Zhao, Bo Li <{zhaorun, zhuokai, bol}@uchicago.edu>, Jiawei Zhou <jzhou@ttic.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

score to steer the generation of the final outputs to balance both OH mitigation and text generation quality.

The main contributions of this paper are: (1) HALC, a novel, plug-and-play decoding algorithm that significantly reduces OH in LVLMs while preserving outputs generation quality; (2) an open-sourced platform that unifies all major OH reduction baselines and state-of-the-arts (SOTAs) (Chuang et al., 2023; Zhou et al., 2023; Yin et al., 2023; Huang et al., 2023; Leng et al., 2023), including HALC, into one framework providing convenient evaluations supporting major LVLM backbones (Zhu et al., 2023; Chen et al., 2023; Liu et al., 2023b; Dai et al., 2023) and OH benchmarks and evaluation metrics (Rohrbach et al., 2018; Fu et al., 2023; Li et al., 2023; Liu et al., 2023a); and (3) comprehensive experimental studies that thoroughly evaluates HALC, demonstrating its superior capability in OH reduction over existing approaches.

2 Related Work

OH and its assessment. OH refers to the phenomenon where vision-language models (VLMs), including both earlier BERT-based models (Li et al., 2019; Radford et al., 2021) and more recent LVLMs (Liu et al., 2023b; Zhu et al., 2023; Tu et al., 2023; Cui et al., 2023; Wang et al., 2024; Zhou et al., 2024b), erroneously generate unfaithful contents. More specifically, Gunjal et al. (2023) and Zhai et al. (2023) proposed that OH could be categorized into three types: object *existence* hallucination for the creation of non-existent objects, object *attribute* hallucination for providing misleading descriptions, and object *relationship* hallucination for depicting incorrect inter-object relationships.

The most well-adopted metric specifically designed to evaluate OH is CHAIR (Rohrbach et al., 2018), which was motivated after Rohrbach et al. (2018) discovered that existing metrics that measure the output's text quality, such as CIDEr (Vedantam et al., 2015), is misleading at representing hallucinations (higher CIDEr score may correlate with higher OH). Another notable and more recent metric is POPE (Li et al., 2023), which transforms the assessment of OH into a binary classification problem where metrics such as precision, recall and accuracy are used to represent the level of OH. In our evaluations, we utilize CHAIR and propose a new metric based on POPE, named *OPOPE*, for thorough assessments of OH, while keeping the standard text generation quality metrics such as BLEU (Papineni et al., 2002), as an additional indicator to make sure little sacrifice in quality was made when mitigating OH.

Challenges and existing approaches. OH has been a persistent challenge over the past years (Rohrbach et al., 2018). Despite numerous advancements in LVLMs (Dai et al., 2022; Li et al., 2023; Zhou et al., 2024a), none of them can produce faithful outputs without suffering from some level of OH. Various strategies have been developed to this matter. For instance, Zhou et al. (2023) and Yin et al. (2023) proposed post-hoc and self-correction pipelines, respectively. Huang et al. (2023) and Leng et al. (2023) developed decoding strategies emphasizing better prior utilization. While effective, these approaches often require powerful external LVLMs or additional data, limiting their adaptability.

Distinct from these methods, HALC offers a novel decoding strategy that effectively reduces OH without necessitating extra LVLMs, training, or data. Integrating a novel adaptive focal-contrast grounding mechanism, HALC addresses both local and global contexts in OH reduction. Its compatibility with open-source LVLMs like MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023b) further enhances its applicability. And as previous approaches often study the problem under different settings and metrics (Zhou et al., 2023; Yin et al., 2023; Huang et al., 2023; Leng et al., 2023), to promote the development of OH reduction in general, we implement an open-source platform which hosts both the proposed HALC and other methods, supporting various LVLM backbones and evaluation metrics.

3 Background and Motivation

3.1 Problem Formulation

We consider an LVLM $\mathcal{M}_\theta^{\text{LVLM}}$ parameterized by θ, with a general architecture consisting of a vision encoder, a vision-text interface module, and a text decoder. For an image-grounded text generation task, given a textual query x and an input image v, v is first processed by the vision encoder into a visual embedding, then transformed by the interface module as the input to the text decoder together with the query x, and finally decoded into a textual response y autoregressively. Formally, we have

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

where y_t denotes the tth token, $y_{<t}$ is the token sequence generated up to time step t, and f_θ is the logit distribution (unnormalized log-probabilities) produced by $\mathcal{M}_\theta^{\text{LVLM}}$.

OH happens when some parts of the text generation y conflicts with the input image v. The goal of OH reduction is to minimize the occurrence of hallucination tokens and preserve the faithfulness to v when addressing the query x, while maintaining a high-quality generation of text y.

3.2 Why Does OH Occur?

OH in VLMs can be attributed to various factors, including but not limited to the inherent biases in the training data caused by co-occurrence (Biten et al., 2022; Zhou et al., 2023), visual uncertainty due to model's statistical bias and priors (Leng et al., 2023), as well as the limitations in current models' ability to discern context and fact accurately

during the entire output generation process (Daunhawer et al., 2021). Studies have also shown that OH is not random but exhibits certain patterns and dependencies, such as its co-existence with knowledge aggregation pattern (Huang et al., 2023), and the tendency to occur with objects positioned later in the generated descriptions (Zhou et al., 2023).

A closer examination of these analysis suggests that the autoregressive nature of the LVLMs may be a fundamental factor contributing to their hallucinatory behaviors. Specifically, autoregressive decoding makes LVLMs progressively rely more on textual information including both the query x and the increasing history generations $y_{<t}$, while unavoidably reducing reliance on the visual input. This imbalance results in a significant deviation from accurate representation of the visual input, ultimately culminating in OH with behaviors and patterns observed in the aforementioned studies (Zhou et al., 2023; Leng et al., 2023). This is especially obvious when longer responses are generated, which explains the correlation between higher OH and larger maximum token lengths, as seen in Huang et al. (2023).

3.3 Fine-grained Visual Knowledge Reduces OH

To mitigate the disproportionate reliance on the textual and visual information during the autoregressive text generation, the process can be enhanced by continuously incorporating targeted visual information. As faithful text generations should guarantee that object-related text tokens are well grounded in the visual input, we hypothesize that the generation can benefit from focusing more on the *fine-grained visual context* for different object-related tokens. For example, for an image showing *a man holding a clock on the beach* as in Fig. 2, the generation of the *clock* token can be well grounded in a smaller region of the image, which we call a specific *visual context*, ideally excluding the beach which is distracting. Therefore, our key insight in mitigating OH lies in identifying a token-wise optimal visual context to provide the most informative visual grounding while decoding a specific token.

We verify our hypothesis through an empirical pilot study. Fig. 1 shows the oracle performance of OH levels when we rely on optimal visual contexts for tokens through brute-force search, with greedy decoding on the MME benchmark (Fu et al., 2023) on three categories of OH.[1] We can see that for most cases, there are optimal visual contexts where decoding from them eliminates over 84.5% of the hallucinations. This motivates our approach of identifying different visual contexts for object-related token generations through *adaptive focal-contrast decoding*, which is introduced in detail in the next section.

[1]Details of this oracle analysis can be found in Appendix C.2.

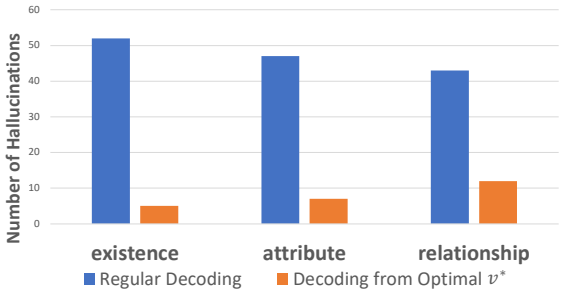


Figure 1: On average, over 84.5% of the observed existence, attribute, and relationship hallucinations are reduced by leveraging some optimal visual context v^*. Blue bar denotes number of hallucinated tokens on each corresponding MME sub-task, while orange bar denotes results when decoding from the oracle v^*.

4 Methodology

An overview of the proposed HALC method is shown in Fig. 2. It operates at the token level during generation, with reliance on fine-grained visual information represented by samples of different visual contexts. By recomputing the token distributions from different visual context inputs and contrasting them, object-related token probabilities are redistributed to reduce hallucinations dynamically within the generation steps. We describe the full procedures below.

4.1 Object-related Token Identification

To focus on the most-probable hallucination sources and optimize time efficiency, we first identify tokens that are related to objects to be processed by HALC. In particular, at each generation step t, we acquire the part-of-speech (POS) tag (Honnibal & Montani, 2017)[2] of the currently generated token from the model $\mathcal{M}_\theta^{\text{LVLM}}$. If the token belongs to noun, adjective/adverb/number/verb/pronoun, or preposition, which correspond to object existence, attribute, and relationship hallucinations, respectively, we redo the current token generation with HALC. For example, as seen in Fig. 2, the newly generated token *surfboard* is identified as it may contribute to the object existence hallucination. Notice that we do not make any assumptions on whether or not the current token is hallucinating, instead, we only determine if the token *can* be prune to hallucination solely based on its syntactic category.

4.2 Visual Context Retrieval

To identify the fine-grained visual information for the current token, we first retrieve a visual context window $v_d = (w_d, h_d, p_d)$ corresponding to the token, where w_d and h_d are the width and height of the visual window, and p_d is the center point. Specifically, we employ a zero-shot detector $\mathcal{G}_d$ such as Grounding DINO (Liu et al., 2023c) or OWLv2 (Minderer et al., 2023) to locate the token within the original image input v. Notably, despite the most com-

[2]We use the small-sized spaCy English pipeline (`https://spacy.io/models/en`) for tagging each complete word.

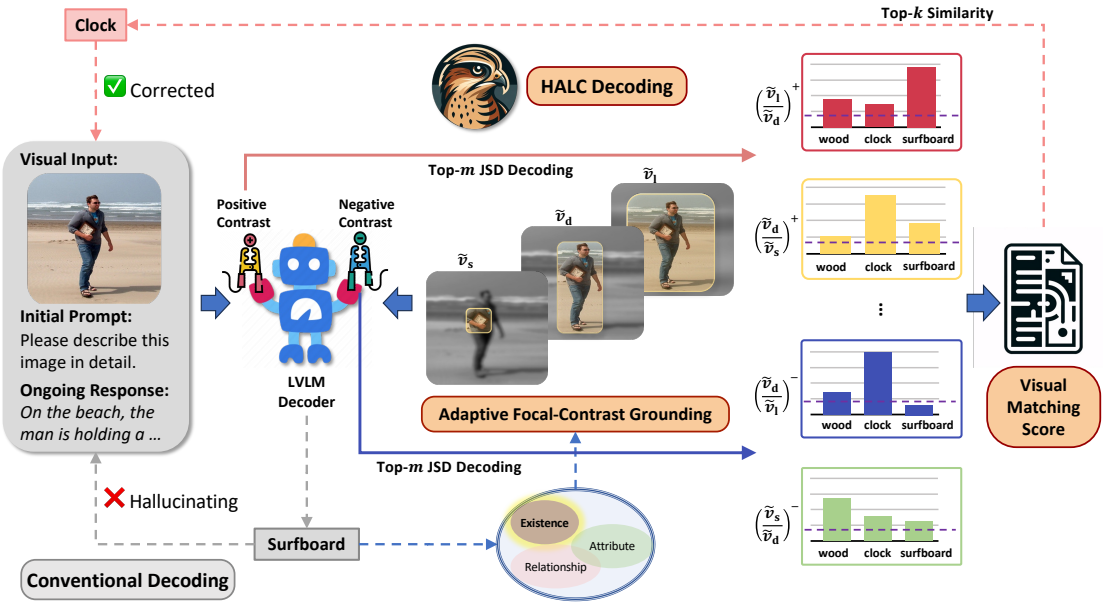


Figure 2: An overview of HALC. As LVLM autoregressively generates texts w.r.t. an image input (e.g. a man holding a clock on the beach), the conventional decoding method may hallucinate the *clock* as *surfboard*. However, HALC corrects this potential hallucination by first locating its visual grounding v_d, then sample n distinctive yet overlapping FOVs (e.g. $\tilde{v}_s$, $\tilde{v}_d$, $\tilde{v}_l$). Next, all FOVs are fed back into the LVLM, along with the current ongoing response, obtaining n logits distributions. Then we compute Jensen-Shannon Divergence (JSD) between each pair of the n distributions, and select the top m pairs, providing $2m$ next-token candidates by bi-directional contrasted logits distributions. Each of the $2m$ candidates are then appended to the k ongoing beams (beam search omitted in the figure for simplicity), resulting in $2mk$ response candidates. Finally, k best responses are selected according to the global visual matching score between current text and original image, completing the current decoding round with the hallucinating token *surfboard* successfully corrected to *clock*.

mon use case of these zero-shot detectors is to locate objects, they are trained to also provide good visual reference for adjective or prepositional phrase. This is because during pre-training, the objective of these detection models is to associate words in text descriptions with specific regions in images (Liu et al., 2023c), which naturally includes attributes and relationships besides names.

Interestingly, we find that although the current token may technically be non-existing when it represents a hallucination (e.g., *surfboard* in Fig. 2), it can still be accurately located by the detector in practice, especially when the detector confidence threshold is set to lower values.

4.3 Adaptive Focal-contrast Grounding

While off-the-shelf detectors establish a meaningful reference v_d within the original image input v, it is often not the optimal visual context for decoding. In Fig. 3, we show an example of how token probabilities representing different objects change with different visual context windows, or field of views (FOVs) input to the vision model in $\mathcal{M}_\theta^{\text{LVLM}}$. In this generation step, the ground-truth token "clock" (we call a *victim token*) is hallucinated to "surfboard". Although direct decoding from v_d does not correct the hallucination as the probability of "clock" is still low, we can see that there exists a better visual context window v_1 that can correct the hallucination, and the curve corresponding to the

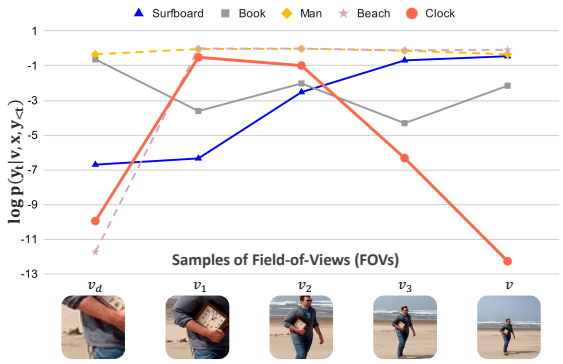


Figure 3: Log-likelihood of object tokens w.r.t. visual context samples in the FOV space, at the generation step in the example of Fig. 2. Exponentially expanding FOVs are adopted. While obvious objects (e.g. *beach*, *man*) are stable with high likelihood, hallucinating objects are either noisy (e.g. *book*) or shift gradually with the context (e.g. *surfboard*). The victim token (e.g. *clock*) usually display a drastically peaking pattern (local maximum).

faithful token "clock" displays a drastically peaking pattern. This is a sharp difference from the patterns of other tokens, which display smaller contrasts when the visual contexts vary. This observation motivates our approach of *focal-contrast grounding* to adaptively adjust the object-related token probabilities, by sampling and selecting a range of most contrasting FOVs based on their decoding probabilities to best approximate the optimal visual contexts.

FOV sampling. We first sample a sequence of n FOVs,

$v_1, v_2, \ldots, v_n$, based on the initial visual context v_d. There could be different approaches to come up with different FOVs conditioning on v_d. To attain a larger coverage of the input image quickly, one strategy to sample FOVs is through an exponential expanding function, by setting

$$v_i = (w_i, h_i, p_i) = \big((1+\lambda)^i w_d, (1+\lambda)^i h_d, p_d\big) \quad (2)$$

where w_i, h_i, p_i are the width, height, and center of FOV v_i.

Dynamic visual context selection. Based on the observation from Fig. 3, we now select a set of FOVs based on a contrastive criterion in the text decoding space to better approximate the optimal visual context for the current token. In particular, after obtaining n different FOVs, we feed these visual contexts back into the model[3] $\mathcal{M}_\theta^{\text{LVLM}}$, resulting in n different probability distributions $p_i = p_\theta(\cdot|v_i, x, y_{<t})$ with $i = 1, 2, \ldots, n$. Between any two candidate FOVs, we adopt the following distance measure for the discrepancy between their decoded token probability distributions

$$d(v_i, v_j) = \text{JSD}(p_\theta(\cdot|v_i, x, y_{<t}) \parallel p_\theta(\cdot|v_j, x, y_{<t})) \quad (3)$$

where JSD is the Jensen-Shannon divergence, a symmetric metric that measures the difference between two distributions. With the idea that more different FOV pairs are more likely to include the optimal visual context for the current victim token generation, we dynamically select the top m pairs with the largest distance according to Eq. (3).

Contrastive decoding. After obtaining top m visual context pairs with most discrepancies in influencing the token output, we contrast the decoding probability distributions (p_i, p_j) within each pair in order to amplify the information residing in one visual context over the other. This would potentially recover the victim token over the hallucinated token as the victim token enjoys a sharper contrast in the probability comparisons, especially when one of the visual contexts under comparison is near the optimal grounding. Specifically, we redistribute the probabilities based on the contrast in log space (Li et al., 2022b) for a given FOV pair (v_i, v_j), resulting in the following distribution

$$p_{v_i/v_j}(\cdot|v_i, v_j, x, y_{<t}) \propto \exp\Big[(1+\alpha) f_\theta(\cdot|v_i, x, y_{<t}) - \alpha f_\theta(\cdot|v_j, x, y_{<t})\Big] \quad (4)$$

where f_θ again is the logit distribution, α is the amplification factor where larger α indicates a stronger amplification of the differences between the distribution pair ($\alpha = 0$ simplifies Eq. (4) to regular decoding from v_i without contrast).

Unlike existing uni-modal contrastive decoding methods (Chuang et al., 2023; Gera et al., 2023; Shi et al., 2023) that assign an expert and an amateur distribution in the contrast by assuming the final or context-aware layer contains more factual knowledge, in our case defining an asymmetric expert distribution among a random pair of FOVs is non-trivial. For example, the optimal visual context usually resides midway among growing FOVs, making either overflowing or insufficient context result in hallucination, as seen in Fig. 3. Therefore, as we have no knowledge where the optimal visual context resides, for each pair of FOVs, we propose to contrast them bi-directionally, which contains both *positive* (larger over smaller-sized FOV) and *negative* (smaller over larger-sized FOV) contrast to preserve the completeness of FOV representations (as shown in Fig. 2). Essentially, this process results in $2m$ candidate tokens by individual greedy decodings which will be further selected by the matching-based beam search algorithm next.

4.4 Matching-based Beam Search

While our adaptive focal-contrast grounding in §4.3 focuses on local token corrections at a single generation step, we adopt a sequence-level beam search algorithm (Anderson et al., 2016) to globally maintain the text generation qualities. Specifically, with a beam size of k, at an HALC decoding step at time t, the k beam sequences would generate $2mk$ token candidates for y_t in total from top m focal-contrast pairs. Different from existing beam score designs (Borgeaud & Emerson, 2019) based only on textual information, we rely on a global visual matching score to select the top k beams from $2mk$ candidates, by comparing the similarity between the current text sequence $y_{\leq t}$ and the original image v. This maintains a diverse but faithful set of generations within the search. In practice, we employ the Bootstrapping Language-Image Pre-training (BLIP) model (Li et al., 2022a) for both text and image encoding and compute their similarity scores.

Combining all components, the full procedure of HALC is summarized in Algorithm 1. Notice that by utilizing the fine-grained visual information at different levels for a single generation step, we admittedly trade in some computation time for correcting token hallucinations. The detailed analysis on time complexity is in Appendix B. One way to increase the HALC decoding speed is through parallelization of decoding from different visual contexts, where we can hope to spend at worst roughly twice of the regular decoding time at HALC steps considering the whole sequence.[4]

5 Theoretical Analysis on FOV Sampling

Based on our observation (in Fig. 1 and Fig. 3) that there exists some underlying optimal visual context v^* within the original image v that can largely reduce the object hallucination at the token level, our method aims to recover this optimal visual context v^* based on a sampling process conditioned on v_d. To do so, we first select the visual contexts,

[3]We directly feed the cropped image to the FOV in the model.

[4]As HALC does not happen at every decoding step. There are also other overhead such as visual grounding affecting the runtime.

Algorithm 1 HALC Decoding

Require: LVLM $\mathcal{M}_\theta^{\text{LVLM}}$, text query x, image input v, grounding detector $\mathcal{G}_d$, FOV sample size n, beam size k, number of contrast FOV pairs m.
output Model response y_{new}.

repeat
  At every decoding step t:
  for $b = 1$ to beam size k **do**
    $\mathcal{M}_\theta^{\text{LVLM}}$ decoding, obtain current token y_t^b
    if $y_t^b \in \{\text{existence}, \text{attribute}, \text{relationship}\}$ **then** ▷ §4.1
      Retrieve visual context $v_d^b \leftarrow \mathcal{G}_d(y_t^b, v)$ ▷ §4.2
    end if
    if $v_d^b \neq \{\varnothing\}$ **then**
      Sample n FOVs $v_1, \ldots, v_n$ by expanding v_d^b
    else
      Randomly sample n FOVs $v_1, \ldots, v_n$ from v
    end if ▷ §4.3
    Compute pair-wise JSDs $d(v_i, v_j), \forall i \neq j$ ▷ §4.3, Eq. (3)
    Select top-m candidate pairs ▷ §4.3
    for $i = 1$ to m **do**
      Apply bi-directional contrast $(p_{v_i/v_j}, p_{v_j/v_i})$,
      get a pair of redistributed logits ▷ §4.3, Eq. (4)
    end for ▷ y_{new}^b with $2m$ candidates obtained
  end for
  Select top k candidates by visual matching ▷ §4.4
  if $v_d^b \neq \{\varnothing\}$ **and** $y_{\text{new}}^b = y_t^b$ **then**
    $y_{\text{new}}^b \leftarrow$ [IDK] ▷ y_t^b is hallucinating, but no correction token was found
  end if
  $y_t^b \leftarrow y_{\text{new}}^b$ ▷ Hallucinating token y_t^b corrected
until each beam has terminated

or FOVs, by taking a sequence of FOV samples starting from the initial v_d based on an off-the-shelf detector. While we cannot guarantee that the initial visual grounding v_d is sufficiently accurate to approximate v^* (and directly using v_d could result in unstable behaviors), we could effectively certify the robustness of our FOV sampling strategy in Theorem 5.1. To preserve generality, consider the sampled FOVs are taken from a distribution $\pi(\cdot|v_d)$, where π can either follow normal distribution sampling around v_d, or obey an exponential expansion sampling strategy starting from v_d.

Theorem 5.1. *Let $v^* = (w^*, h^*, p^*)$ be the optimal visual context. Assume there exists a tolerable neighborhood $\mathcal{B}(v^*, \epsilon) = \{\hat{v} : \|\hat{v} - v^*\| \leq \epsilon\}$ around v^*, such that decoding from visual contexts within the neighborhood is robust:*

$$D(p_\theta(\cdot|v^*), p_\theta(\cdot|\hat{v})) \leq \delta \ll 1, \forall \hat{v} \in \mathcal{B}(v^*, \epsilon) \tag{5}$$

where $D(\cdot, \cdot) \in [0, 1]$ is a symmetric discrepancy measure between two probability distributions, such as the Jensen-Shannon divergence, or the total variation distance.

Let $v_d = (w_d, h_d, p_d)$ be the initial detection and $v_d = v^* + \eta$ with perturbation η. The minimum deviation of token probabilities from the optimum with n samples $v_1, v_2, \ldots, v_n$ distributed according to $\pi(\cdot|v_d)$ is denoted as

$$h_\pi(v^*, n) = \min_{i=1,\ldots,n} D\left(p_\theta(\cdot|v^*), p_\theta(\cdot|v_i)\right) \tag{6}$$

(a) For normal distribution sampling $\pi_g(\cdot|v_d) \sim \mathcal{N}(v_d, \sigma^2 I)$, the minimum deviation above is bounded as

$$h_{\pi_g}(v^*, n) \leq \delta + (1 - C_g(\epsilon, \eta; \sigma))^n \tag{7}$$

where $C_g(\epsilon, \eta; \sigma) \in (0, 1)$ is a constant depending on ϵ, η, σ, and the upper bound goes to δ when $n \to \infty$.

(b) For exponential expansion sampling $\pi_e(\cdot|v_d) \sim \mathcal{U}(r \in [r_{\min}, r_{\max}])$ with samples $v_r = ((1+\lambda)^r w_d, (1+\lambda)^r h_d, p_d)$ uniformly from the r-space, under the conditions (i) $|p_d - p^*| < \epsilon$ and (ii) $w_d/h_d = w^*/h^*$, the minimum deviation in Eq. (6) is bounded below

$$h_{\pi_e}(v^*, n) \leq \delta + (1 - C_e(\epsilon, v^*, v_d; \lambda))^n \tag{8}$$

where $C_e(\epsilon, v^*, v_d; \lambda) \in (0, 1]$ is a constant depending on $\epsilon, v^*, v_d, \lambda$, and the upper bound goes to δ when $n \to \infty$.

The proof of Theorem 5.1 is detailed in Appendix A. The neighborhood radius ϵ around the optimal v^* can be roughly interpreted as a valid range of optimal visual context to yield the correct prediction (e.g., $[v_1, v_2]$ in Fig. 3). Typically the detection perturbation $\|\eta\| > \epsilon$, making v_d outside of the ϵ-neighborhood of v^*. Through FOV sampling according to some $\pi(\cdot|v_d)$, the above theorem establishes a formal guarantee that at least one of the n samples achieves good approximation of the optimal v^* in the decoding probability space, as the deviation is closer to δ when n grows. The normal sampling distribution, concentrated around v_d, is preferred when v_d has minimal perturbations from v^*. And an exponential expansion sampling distribution, with a more averaged coverage of the sampling space, is preferable when less prior of the task is available. In practice of our algorithm, we take discrete integer values of r under the exponential expansion distribution for deterministic sampling with $n = 4$, acquiring good efficiency and performance.

6 Experiments

Benchmarks. We evaluate HALC on three benchmarks including (1) quantitative metrics CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023) on MSCOCO (Lin et al., 2014) dataset; (2) general-purposed Multimodal Large Language Model Evaluation (MME) (Fu et al., 2023) benchmark; and (3) qualitative evaluation benchmark LLaVA-Bench (Liu et al., 2023a). These experiments comprehensively assess HALC's capability on reducing OH in image captioning, visual-question answering (VQA) and more challenging tasks that generalize to novel domains.

Table 1: CHAIR evaluation results on MSCOCO dataset of LVLMs with different decoding baselines and SOTAs designed for mitigating OH. Lower CHAIR_S and CHAIR_I indicate less OH. Higher BLEU generally represent higher captioning quality, although existing work has reported weak correlation between CHAIR and text overlapping quality metrics. Bold indicates the best results of all methods.

Method	MiniGPT-4			LLaVA-1.5			mPLUG-Owl2		
	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	BLEU↑	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	BLEU↑	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	BLEU↑
Greedy	$30.87_{\pm 5.45}$	$12.33_{\pm 2.07}$	$14.33_{\pm 0.00}$	$20.80_{\pm 0.08}$	$6.77_{\pm 0.07}$	$15.93_{\pm 0.00}$	$23.20_{\pm 0.35}$	$8.33_{\pm 0.28}$	$15.37_{\pm 0.00}$
Beam Search	$29.56_{\pm 6.09}$	$11.36_{\pm 0.99}$	$14.94_{\pm 0.00}$	$18.67_{\pm 0.38}$	$6.30_{\pm 0.05}$	$16.17_{\pm 0.00}$	$21.67_{\pm 1.61}$	$7.63_{\pm 0.40}$	$15.77_{\pm 0.00}$
DoLA	$30.87_{\pm 2.52}$	$11.70_{\pm 0.13}$	$14.93_{\pm 0.00}$	$21.00_{\pm 0.67}$	$6.70_{\pm 0.38}$	$15.93_{\pm 0.00}$	$24.60_{\pm 0.24}$	$8.73_{\pm 0.30}$	$15.40_{\pm 0.00}$
OPERA	$30.00_{\pm 0.43}$	$11.67_{\pm 0.22}$	$14.87_{\pm 0.00}$	$21.13_{\pm 0.12}$	$6.73_{\pm 0.18}$	$16.27_{\pm 0.01}$	$22.13_{\pm 0.86}$	$7.57_{\pm 0.16}$	$15.53_{\pm 0.00}$
VCD	$30.27_{\pm 0.44}$	$12.60_{\pm 0.45}$	$14.33_{\pm 0.00}$	$23.33_{\pm 5.66}$	$7.90_{\pm 0.53}$	$14.67_{\pm 0.01}$	$27.27_{\pm 7.32}$	$9.73_{\pm 1.22}$	$14.40_{\pm 0.00}$
Woodpecker	$28.87_{\pm 2.20}$	$10.20_{\pm 0.85}$	$\mathbf{15.30}_{\pm 0.01}$	$23.85_{\pm 4.62}$	$7.50_{\pm 0.01}$	$\mathbf{17.05}_{\pm 0.00}$	$26.33_{\pm 1.98}$	$8.43_{\pm 0.80}$	$\mathbf{16.43}_{\pm 0.00}$
LURE	$27.88_{\pm 2.25}$	$10.20_{\pm 0.85}$	$15.03_{\pm 0.11}$	$19.48_{\pm 2.35}$	$6.5_{\pm 0.38}$	$15.97_{\pm 0.01}$	$21.27_{\pm 0.06}$	$7.67_{\pm 0.16}$	$15.65_{\pm 0.05}$
HALC	$\mathbf{17.80}_{\pm 0.03}$	$\mathbf{8.10}_{\pm 0.14}$	$14.91_{\pm 0.00}$	$\mathbf{13.80}_{\pm 0.08}$	$\mathbf{5.50}_{\pm 0.14}$	$16.10_{\pm 0.01}$	$\mathbf{17.33}_{\pm 4.30}$	$\mathbf{7.43}_{\pm 0.11}$	$16.27_{\pm 0.00}$

Baselines. To effectively evaluate HALC, besides regular greedy decoding and beam search baselines, we further involve layer-wise contrastive decoding SOTA DoLa (Chuang et al., 2023), as well as SOTA methods specifically designed to mitigate OH, including OPERA (Huang et al., 2023), VCD (Leng et al., 2023), Woodpecker (Yin et al., 2023) and LURE (Zhou et al., 2023) in our analysis. All the results are acquired and benchmarked consistently with our unified implementation. Please refer to Appendix C.1 for the detailed setting of our HALC.

LVLM Backbones. Three LVLMs including MiniGPT-4 V2 (Chen et al., 2023), LLaVA-1.5 (Liu et al., 2023b) and mPLUG-Owl2 (Ye et al., 2023) are used for both HALC and all aforementioned baselines except Woodpecker and LURE, where Woodpecker utilizes ChatGPT (Brown et al., 2020) during its self-correction process and LURE distills an extra reviser model from GPT-4 (Achiam et al., 2023).

6.1 CHAIR and POPE on MSCOCO

Following existing evaluation procedures (Huang et al., 2023; Yin et al., 2023; Liu et al., 2023b), we randomly sampled 500 images from the validation split of MSCOCO (Lin et al., 2014) and conduct evaluations with both CHAIR and POPE. For each metric, we repeat the experiments five times with different random seeds and report average and standard deviations of all the runs.

CHAIR. Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018) is a tailored tool created to evaluate the occurrence of OH in the task of image captioning. Specifically, CHAIR measures the extent of OH in an image description by determining the proportion of the mentioned objects that are absent in the actual label set. This metric includes two separate evaluation aspects: CHAIR_S, which performs assessments at the *sentence* level (proportion of the hallucinated sentences over all sentences), and CHAIR_I, which operates at the object *instance* level (proportion of the hallucinated objects over all generated objects). Lower scores indicate less OH.

We prompt all methods with "*Please describe this image in detail.*" and the results are illustrated in Table 1. Besides CHAIR_S and CHAIR_I, we also report BLEU (Papineni et al., 2002) as an assessment of the text generation quality. Table 1 demonstrats that our proposed HALC consistently outperforms all the existing methods by a large margin. Notably, a major advantage of HALC is its strong robustness, as can be observed by its much lower standard deviations, especially when compared to the non-OH specific baselines. While Woodpecker (Yin et al., 2023) has the highest generation quality BLEU scores, this can be largely attributed to the fact that Woodpecker adopts ChatGPT, a much more capable LLM, to organize the final outputs, which is not exactly a fair comparison to the other methods.

We also investigate how HALC performs with longer responses, as showed in Fig. 4, where we plot both the number of generated (dashed) and hallucinated (solid) objects with randomly sample 100 images. This experiment is important to further assess HACL's robustness, as it is commonly believed that OH happens more with objects positioned later in the responses (Zhou et al., 2023), as well as in longer responses (Huang et al., 2023). We observe that HALC is the only method that can keep even smaller number of hallucinations while the number of generated objects increases, demonstrating its superior performance and advantageous robustness in reducing OH.

POPE. Polling-based Object Probing Evaluation (POPE) (Li et al., 2023) evaluates OH via a streamlined approach, which incorporates a list of yes-or-no questions to prompt LVLMs for presence of positive and negative objects. When selecting negative (non-existing) objects for prompting, POPE provides three sampling options: random, popular, and adversarial. We refer detailed explanations of the different options to its original paper (Li et al., 2023).

One distinct difference between POPE and CHAIR is that POPE relies on interacting with the examined LVLM directly. While this requirement is not an issue when evaluating the decoding-based baselines, it limits its adaptation

Table 2: Proposed OPOPE evaluation results on MSCOCO dataset of LVLMs with different decoding baselines and SOTAs designed for mitigating OH. Higher accuracy, precision, and F score indicate better performance. Bold indicates the best results of all methods.

Method	MiniGPT-4			LLaVA-1.5			mPLUG-Owl2		
	Accuracy↑	Precision↑	$F_{\beta=0.2}$ ↑	Accuracy↑	Precision↑	$F_{\beta=0.2}$ ↑	Accuracy↑	Precision↑	$F_{\beta=0.2}$ ↑
Greedy	$66.78_{\pm 1.27}$	$90.43_{\pm 25.1}$	$85.79_{\pm 18.7}$	$70.56_{\pm 1.51}$	$91.08_{\pm 20.6}$	$87.72_{\pm 16.3}$	$69.77_{\pm 1.18}$	$91.07_{\pm 17.8}$	$87.45_{\pm 13.9}$
Beam Search	$67.22_{\pm 0.74}$	$91.20_{\pm 14.4}$	$86.57_{\pm 10.8}$	$69.87_{\pm 1.37}$	$91.72_{\pm 20.4}$	$88.01_{\pm 15.97}$	$69.20_{\pm 0.90}$	$91.90_{\pm 15.1}$	$87.91_{\pm 11.7}$
DoLA	$67.06_{\pm 1.19}$	$90.84_{\pm 23.1}$	$86.22_{\pm 17.3}$	**70.69**$_{\pm 1.50}$	$90.87_{\pm 19.8}$	$87.59_{\pm 15.74}$	**70.17**$_{\pm 1.69}$	$91.97_{\pm 24.5}$	$88.30_{\pm 19.26}$
OPERA	$67.26_{\pm 1.04}$	$90.76_{\pm 20.0}$	$86.25_{\pm 15.0}$	$69.73_{\pm 1.34}$	$91.10_{\pm 19.4}$	$87.46_{\pm 15.3}$	$69.26_{\pm 0.45}$	**93.06**$_{\pm 8.01}$	**88.83**$_{\pm 6.14}$
VCD	$65.78_{\pm 0.96}$	$90.02_{\pm 20.7}$	$85.00_{\pm 15.1}$	$70.67_{\pm 1.22}$	$91.62_{\pm 16.7}$	$88.19_{\pm 13.3}$	$69.81_{\pm 0.65}$	$92.70_{\pm 11.0}$	$88.76_{\pm 8.49}$
Woodpecker	$67.78_{\pm 0.88}$	$91.33_{\pm 16.66}$	$86.91_{\pm 12.6}$	$69.80_{\pm 0.54}$	$91.80_{\pm 8.41}$	$88.04_{\pm 6.56}$	$68.90_{\pm 1.02}$	$92.22_{\pm 17.98}$	$88.05_{\pm 13.77}$
LURE	**68.14**$_{\pm 0.99}$	$90.95_{\pm 17.34}$	$86.76_{\pm 13.23}$	$70.00_{\pm 1.53}$	$90.89_{\pm 21.9}$	$87.38_{\pm 17.3}$	$69.24_{\pm 1.60}$	$90.54_{\pm 23.37}$	$86.85_{\pm 18.28}$
HALC	$66.76_{\pm 0.68}$	**91.95**$_{\pm 15.0}$	**86.92**$_{\pm 11.1}$	$70.59_{\pm 0.82}$	**92.94**$_{\pm 12.18}$	**89.22**$_{\pm 9.55}$	$70.12_{\pm 0.98}$	$91.94_{\pm 15.1}$	$88.26_{\pm 11.85}$

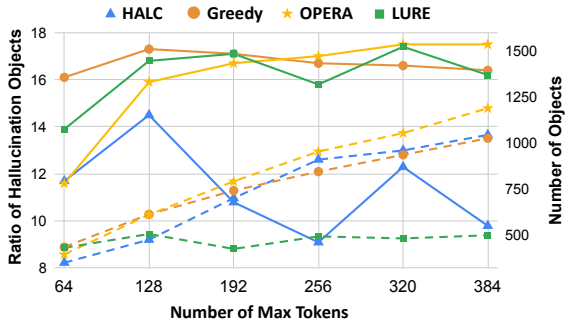


Figure 4: Comparing four mainstream methods on the ratio of hallucination objects (CHAIR_I) v.s. the number of max tokens. The right axis (dashed line) indicates the total number of generated objects. HALC outperforms all other methods by maintaining a low ratio of hallucination with the increasing of generated objects.

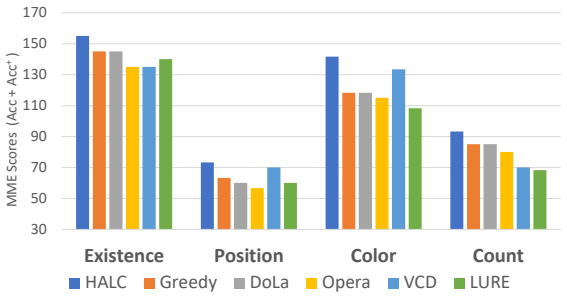


Figure 5: Comparison across OH baselines and SOTAs on four OH-critical MME subsets. All methods adopt MiniGPT-4 as LVLM backbone. HALC outperforms all other methods with a large margin: *existence*: +10.7%; *position*: +18.3%; *color*: +19.4% and *count*: +20.2% in average.

to post-hoc methods such as LURE (Zhou et al., 2023). It also creates larger instabilities when the examined LVLM incorporates smaller language backbones such as LLaMA-7B (Touvron et al., 2023), which has less robust chat capability. To these concerns, we propose *offline POPE (OPOPE)*, which keeps the object sampling and yes/no query strategy from POPE, but replaces the live interactions with offline checks. Specifically, instead of querying the model with "*Is there a {} in the image?*", where "{}" is the queried object, we first ask the examined LVLM to give its detailed descriptions of the image, and then manually check if the sampled positive/negative objects exist in the captions when computing the OPOPE scores.

We also adjust the main metrics for comparison. As it is more random for descriptions to include the exact sampled hallucinated objects, false-negative (FN) and the resulting recall become less trustable in the offline checks. Therefore, we propose to use F-beta, instead of F-1, as the main metric of OPOPE, so that the final score relies less on the FN. Specifically, we have $F_\beta = (1+\beta^2)\cdot(\text{precision}\cdot\text{recall})/(\beta^2\cdot\text{precision} + \text{recall})$, where we use $\beta = 0.2$ throughout our experiments. The evaluation results incorporating OPOPE is shown in Table 2. All the numbers are averaged results of the three sampling methods (random, popular and adversarial, as in the original POPE), while the complete version of the table is shown in Appendix F. We also include the original POPE evaluation results in Appendix E where HALC also outperforms other methods in most of the settings.

6.2 MME

The Multimodal Large Language Model Evaluation (MME) (Fu et al., 2023) benchmark is a comprehensive tool designed to quantitatively compare multimodal LLMs. Following Yin et al. (2023); Leng et al. (2023), we utilize the "existence" and "count" subsets to evaluate the object existence hallucinations and the "position" and "color" subsets for object attribute and relationship hallucination. Please refer to Appendix D for experiment details. The comprehensive results across six methods are reported in Fig. 5, where HALC significantly outperforms all the other methods on each sub-task, indicating an overall performance gain in reducing OH while preserving generation quality.

6.3 LLaVA-Bench Qualitative Study

LLaVA-Bench (Liu et al., 2023a) is a collection of 24 images, where each image is paired with a detailed, manually-crafted description and carefully selected questions. The questions are divided into three categories: simple QA (conversation), detailed descriptions, and complex reasoning. In this experiment, we leverage LLaVA-Bench as a case study

to qualitatively compare the decoding outputs of HALC with other methods. The results are shown in Appendix G.

7 Analysis and Ablation Studies

7.1 Adaptive Focal-contrast Grounding

FOV Sampling initialization. The visual context retrieval process described in §4.2 utilizes detector output as a key component of the adaptive focal-contrast grounding algorithm introduced in §4.3. However, it is important to note that HALC primarily uses the detector output as a *initialization* for the field of view (FOV) sampling process, rather than depending heavily on it. In this section, we present empirical results to compare different methods of sampling initialization, which include random sampling (selecting a random FOV within the image), center initialization (selecting a fixed region in the center of the image), original image initialization (using the entire image) and detector initialization (using the detector output). More specifically, we include an extra detector model, OWLv2 (Minderer et al., 2024), in addition to the Grounding Dino (Liu et al., 2023c) illustrated in previous sections.

Table 3: HALC performance with different sampling initialization.

Init.	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	OPOPE ↑	POPE ↑	BLEU ↑
Random	25.6	11.8	83.33	67.67	15.10
Center	23.9	11.2	86.62	69.10	14.80
Original	27.8	12.2	85.20	68.33	15.50
G. Dino	**22.0**	**8.8**	**88.20**	**70.67**	**16.40**
OWLv2	23.4	10.8	84.47	67.50	15.70

As shown in Table 3, both random and center initialization perform better than using the original image as the visual input. This result confirms the robustness of the proposed FOV sampling process. Additionally, both detectors deliver better performance than the other initializations, further demonstrating that using a detector-grounded FOV provides an effective starting point for the subsequent conditional FOV sampling process.

Exponential Expanding ratio. Besides initialization, another important parameter used in adaptive focal-contrast grounding is the expanding ratio λ, which determines each sampling FOV as in Eq. (2). Thus we further analyze the performance of HALC with different expanding ratios.

Table 4: HALC performance with different expanding ratios.

λ	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	OPOPE ↑	POPE ↑	BLEU ↑
0.2	22.0	8.5	86.45	69.63	**16.60**
0.4	**18.0**	**7.6**	87.33	70.20	16.10
0.6	22.0	8.8	**88.20**	**70.67**	16.40
0.8	28.0	9.6	86.45	69.63	14.80
1.0	26.0	8.9	84.32	69.63	14.70

Table 4 demonstrates that an expanding ratio of 0.6 is optimal. We hypothesize that the poorer performance associated with smaller or larger expanding ratios is due to that smaller ratios increase the number of FOV samples, which presents greater challenges for the global beam search. On the other hand, larger ratios decrease the granularity of the FOV in the image, potentially leading to more severe hallucinations.

7.2 Global Beam Search

Beam sizes. As is common with all beam search algorithms, beam size k is a major hyperparameter. Thus here we examine the performance of HALC w.r.t. different values of k.

Table 5: HALC performance with different values of beam size k.

k	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	OPOPE ↑	POPE ↑	BLEU ↑
1	36.0	14.6	88.20	70.49	15.40
2	**22.0**	**8.8**	**88.74**	**70.67**	**16.40**
3	26.0	9.8	87.65	**70.67**	15.40
5	29.6	11.1	86.33	70.14	15.70
8	33.3	13.8	87.73	70.14	15.50

Table 5 shows improved performance as the beam size initially increases from one. However, when the beam size reaches or exceeds two, the number of FOV samples also increases, making it more challenging for the global beam search module to select the optimal visual context from all the samples, thus leading to a higher rate of hallucination. Furthermore, as the beam size continues to increase, the variance of HALC's performance also increases, indicating that it will be more difficult to select the top candidate as the global matching model also suffers from hallucination.

Scoring methods. Finally, we compare the BLIP and CLIP scoring models with random selection to rank the beams.

Table 6: HALC performance with different scoring methods.

	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	OPOPE ↑	POPE ↑	BLEU ↑
Random	26.6	12.8	85.45	68.45	15.20
BLIP	**22.0**	**8.8**	**88.20**	70.67	**16.40**
CLIP	23.4	10.0	87.67	**71.96**	15.60

As shown in Table 6, different scoring methods do not lead to large variations and they all outperform random selection.

8 Conclusion

We present HALC, a novel decoding algorithm designed to mitigate OH in LVLMs. HALC operates on both local and global levels, integrating a robust adaptive focal-contrast grounding mechanism to better utilize fine-grained visual information for correcting hallucinated tokens, and a specialized beam search algorithm that promotes further visually matched generations. Comprehensive experiments demonstrate that HALC effectively reduces OH, achieving SOTA performance while preserving sequence generation quality, and can be conveniently integrated into existing LVLMs without additional training or data. A benchmarking tool was also built to support convenient comparisons across all available OH reduction strategies comprehensively.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

We thank Lingyu Gao for initial discussion and constructive suggestions. This work was supported in part by the Research Computing Center at the University of Chicago, and Cisco Faculty Research Award. We also thank Center for AI Safety and Google Cloud Research Credits program for supporting our computing needs. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. Longman grammar of spoken and written english, 2000.

Biten, A. F., Gómez, L., and Karatzas, D. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1381–1390, 2022.

Borgeaud, S. and Emerson, G. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. *arXiv preprint arXiv:1908.06288*, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., and Elhoseiny, M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., and Yao, H. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Dai, W., Liu, Z., Ji, Z., Su, D., and Fung, P. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*, 2022.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B. A., Fung, P., and Hoi, S. C. H. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. URL https://api.semanticscholar.org/CorpusID:258615266.

Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., and Vogt, J. E. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*, 2021.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Gera, A., Friedman, R., Arviv, O., Gunasekara, C., Sznajder, B., Slonim, N., and Shnarch, E. The benefits of bad advice: Autocontrastive decoding across model layers. *arXiv preprint arXiv:2305.01628*, 2023.

Guan, T., Liu, F., Li, X. W. R. X. Z., Wang, X. L. X., Yacoob, L. C. F. H. Y., and Zhou, D. M. T. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.

Honnibal, M. and Montani, I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7 (1):411–420, 2017.

Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., and Yu, N. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022a.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., and Lewis, M. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022b.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023c.

Minderer, M., Gritsenko, A., and Houlsby, N. Scaling open-vocabulary object detection. *arXiv preprint arXiv:2306.09683*, 2023.

Minderer, M., Gritsenko, A., and Houlsby, N. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., and Yih, S. W.-t. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tu, H., Cui, C., Wang, Z., Zhou, Y., Zhao, B., Han, J., Zhou, W., Yao, H., and Xie, C. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Wang, X., Zhou, Y., Liu, X., Lu, H., Xu, Y., He, F., Yoon, J., Lu, T., Bertasius, G., Bansal, M., et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

Zhai, B., Yang, S., Xu, C., Shen, S., Keutzer, K., and Li, M. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

Zhang, Y., Zhao, Z., Chen, Z., Feng, Z., Ding, Z., and Sun, Y. Rankclip: Ranking-consistent language-image pretraining. *arXiv preprint arXiv:2404.09387*, 2024.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

Zhou, Y., Cui, C., Rafailov, R., Finn, C., and Yao, H. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024a.

Zhou, Y., Fan, Z., Cheng, D., Yang, S., Chen, Z., Cui, C., Wang, X., Li, Y., Zhang, L., and Yao, H. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024b.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Proof of Robust Certification of FOV Sampling in Theorem 5.1

This section proves theoretical analysis on the robustness of HALC in approximating the optimal visual context v^* via sampling in the FOV space (Theorem 5.1). With certain assumptions on v^* and v_d, we focus on demonstrating the certified robustness on the decoding token probability distribution compared with that from the optimal visual context v^*, when sampling different FOVs based on v_d which is initially determined by an detector $\mathcal{G}_d$.

The objective of HALC is to approximate the unknown optimal visual context for a decoding step, thereby mitigating hallucination and enhancing the truthfulness of the LVLM outputs. We approach the optimal proxy by sampling a series of n FOVs in the original image v, starting from v_d according to some sampling function $\pi(\cdot|v_d)$. We focus on bounding the minimum deviation of the decoding token probabilities from the optimum among the n FOV samples, with the hope that we can always find some sample that is close to the optimal v^* during this process. And as the sample size n becomes larger, the minimum deviation becomes smaller, indicating that we can better cover the optimal visual context v^* within the samples.[5]

Proof. Let $v^* = (w^*, h^*, p^*)$ be the optimal visual context, represented by a 3-tuple of its width, height, and center point. The corresponding optimal token decoding probability distribution is $p_\theta(\cdot|v^*)$, where θ denotes the parameters of the LVLM $\mathcal{M}_\theta^{\text{LVLM}}$, and we ignore the condition on the textual query x and previously generated tokens $y_{<t}$ for simplicity. We rely on a symmetric discrepancy measure $D(\cdot,\cdot) \in [0,1]$ to compare the disparity between two probability distributions, such as the Jensen-Shannon divergence, or the total variation distance. We assume that the model prediction is robust around v^* against small perturbations. In particular, we assume that there exists a tolerable small ϵ-neighborhood $\mathcal{B}(v^*, \epsilon) = \{\hat{v} : \|\hat{v} - v^*\| \leq \epsilon\}$ around v^*, such that

$$g(v^*, \hat{v}) = D(p_\theta(\cdot|v^*), p_\theta(\cdot|\hat{v})) \leq \delta \ll 1, \quad \forall \hat{v} \in \mathcal{B}(v^*, \epsilon) \tag{9}$$

Essentially, for any visual context window (or FOV) close enough to v^*, the output token probability disparity is tiny, which is likely to result no difference in greedy decoding.

From the FOV detector $\mathcal{G}_d$, the output visual context is denoted as $v_d = (w_d, h_d, p_d)$, which is in general not the optimal. We assume $v_d = v^* + \eta$ in the 3-tuple vector space, where η is the perturbation vector from the optimal. The detection perturbation is often large enough with $\|\eta\| > \epsilon$, making v_d outside of the ϵ-neighborhood of v^*.

$v_d \to v^*$: If we directly use the detector output v_d as an approximation of the optimal visual context v^*, the output distribution deviation from the optimum, measured by $g(v^*, v_d)$, is often unpredictable, when v_d does not fall in the hypothetical tolerable region $\mathcal{B}(v^*, \epsilon)$. An example can be seen as the inaccurate detection v_d in Fig. 3 results in the wrong token prediction *book*. This prompts the need for our proposed FOV sampling approach with the hope to find samples close to the optimal v^*.

$\pi(\cdot|v_d) \to v^*$**:** Thus we consider sampling conditioned on v_d in the FOV space to enhance the robustness of optimal visual context approximation, hoping to find some sample that is close to the optimal. To do this, we obtain an upper bound on the minimum deviation from the output distribution among a collection of FOV samples. Assume $\pi(\cdot|v_d) \in \Omega$ is an arbitrary sampling function conditional on the initial FOV detection v_d, where Ω denotes the sampling space over all potential visual contexts in the entire image v. π can either be a deterministic sampling function, or a stochastic sampling process with a probabilistic distribution over Ω. Suppose we acquire n samples $v_1, v_2, \ldots, v_n$ according to $\pi(\cdot|v_d)$, we denote the minimum deviation of the resulted token probability from that of the optimal visual context v^* as

$$h_\pi(v^*, n) = \min_{i=1,\ldots,n} g(v^*, v_i) = \min_{i=1,\ldots,n} D\left(p_\theta(\cdot|v^*), p_\theta(\cdot|v_i)\right) \tag{10}$$

where D is the aforementioned symmetric discrepancy measure between two probability distributions, which is within the range of $[0,1]$. Having a small value of $h_\pi(v^*, n)$ would indicate that we can find some visual context that is close to the optimal v^* through n samples.

We proceed to estimate the minimum deviation $h_\pi(v^*, n)$ from the optimal visual context v^* with n samples. We introduce a partition based on the occurrence of two probabilistic events: the event A where at least one of the samples falls into the ϵ-neighborhood $\mathcal{B}(v^*, \epsilon)$ close to v^*, and its complement. Let us denote the probability of at least one sample falling within

[5]The subsequent selection of the best sample is another question, which is not concerned in this proof. We theoretically justify the existence of an "optimal" sample in the proof here, and HALC selects such a sample by contrasting FOV pairs based on the observation illustrated in Fig. 3.

$\mathcal{B}(v^*, \epsilon)$ as $P(A)$, and the complementary event's probability as $P(\neg A) = 1 - P(A)$. Hence, we can express the minimum divergence $h_\pi(v^*, n)$ as a marginalization over these events:

$$h_\pi(v^*, n) = P(A) \cdot [h_\pi(v^*, n)|A] + P(\neg A) \cdot [h_\pi(v^*, n)|\neg A] \tag{11}$$

Recognizing that for the one sample in the vicinity of v^* in the event of A, its decoding token probability deviation from the optimal is bounded by $\delta \ll 1$ based on our assumption. Hence we have

$$h_\pi(v^*, n) \leq P(A) \cdot \delta + P(\neg A) \cdot 1 \leq \delta + P(\neg A) \tag{12}$$

Next, we consider two instances of the sampling function $\pi(\cdot|v_d)$ that yield an upper bound for $h_\pi(v^*, n)$.

Normal Distribution Sampling. Suppose sampling from π follows a stochastic process following a normal distribution around v_d. We denote this sampling process as $\pi_g(\cdot|v_d) \sim \mathcal{N}(v_d, \sigma^2 I)$, where we assume a variance of σ^2 for each element of the visual context representation (width, height, center) independently. For $\tilde{v} \in \Omega$, the probability of sampling $\tilde{v}$ following the multivariate normal distribution is

$$q(\tilde{v}; v_d, \sigma^2 I) = \frac{1}{\sqrt{(2\pi\sigma^2)^s}} \exp\left(-\frac{1}{2\sigma^2}(\tilde{v} - v_d)^\top (\tilde{v} - v_d)\right)$$

where $s = 3$ is the dimension of the FOV representation vector. The probability of event $\neg A$ happening, which is none of n FOV samples falling within the ϵ-neighborhood of v^*, is

$$\begin{aligned} P(\neg A) &= P(\|v_1 - v^*\| > \epsilon) \wedge P(\|v_2 - v^*\| > \epsilon) \wedge \cdots P(\|v_n - v^*\| > \epsilon) && (13)\\ &= P(\|\tilde{v} - v^*\| > \epsilon)^n && (14)\\ &= P(\|\tilde{v} - (v_d - \eta)\| > \epsilon)^n && (15) \end{aligned}$$

From the normal distribution assumption of $\tilde{v}$, we know that $\tilde{v} - (v_d - \eta)$ also follows a normal distribution $\mathcal{N}(\eta, \sigma^2 I)$. Therefore,

$$\begin{aligned} P(\neg A) &= (1 - P(\|\tilde{v} - (v_d - \eta)\| \leq \epsilon))^n && (16)\\ &= \left(1 - \int_{\nu:\|\nu\|\leq\epsilon} \frac{1}{\sqrt{(2\pi\sigma^2)^s}} \exp\left(-\frac{1}{2\sigma^2}(\nu - \eta)^\top(\nu - \eta)\right) d^s\nu\right)^n && (17)\\ &= (1 - C_g(\epsilon, \eta; \sigma))^n && (18) \end{aligned}$$

where we use $C_g(\epsilon, \eta; \sigma) \in (0, 1)$ to denote the constant value given ϵ, η, and σ. Following Eq. (12), we now have

$$h_{\pi_g}(v^*, n) \leq \delta + (1 - C_g(\epsilon, \eta; \sigma))^n \tag{19}$$

where the second term goes to 0 as n is increasing to larger values.

Exponential Expansion Sampling. Now suppose sampling from π follows an exponential expanding process, where a sample can be expressed as $v_r = (w_r, h_r, p_r) = ((1+\lambda)^r w_d, (1+\lambda)^r h_d, p_d)$ with an expanding factor λ (assuming $\lambda > 0$ without loss of generality) and some r.[6] Essentially, the sample space comprises all fields of view (FOVs) that maintain the same aspect ratio (i.e. w_d/h_d) and the same center p_d with v_d. Assume the sampling is uniform among all possible FOVs in the sample space, which we denote as $\pi_e(\cdot|v_d) \sim \mathcal{U}(r \in [r_{\min}, r_{\max}])$, where $r_{\min}$ and $r_{\max}$ correspond to the smallest FOV allowed (such as a few pixels) and the largest FOV possible (i.e. the entire original image v), respectively.

For this sampling distribution, we introduce two moderate assumptions regarding the initial detection v_d. First, the center of the detection is relatively close to the optimum, such that $|p_d - p^*| < \epsilon$. Second, The detection v_d and the optimum v^* share the same aspect ratio, meaning $w_d/h_d = w^*/h^*$. This assumption is reasonable since the optimum is unknown, and we can assume it adheres to the aspect ratio used by a standard detector.

[6] Besides expansion, this could also be an exponential shrinking process when r is negative. We abuse the use of "expansion" for both.

We begin by deriving the range of r such that v_r falls into the small neighborhood $\mathcal{B}(v^*, \epsilon)$ around v^*. We need

$$\|v_r - v^*\| \leq \epsilon \tag{20}$$

$$\implies \quad (w_r - w^*)^2 + (h_r - h^*)^2 + (p_r - p^*)^2 \leq \epsilon^2 \tag{21}$$

$$\implies \quad [(1+\lambda)^r w_d - w^*]^2 + [(1+\lambda)^r h_d - h^*]^2 + (p_d - p^*)^2 \leq \epsilon^2 \tag{22}$$

$$\vdots$$

$$\implies \quad (w_d^2 + h_d^2)\left((1+\lambda)^r - \frac{w_d w^* + h_d h^*}{(w_d^2 + h_d^2)}\right)^2 \leq \epsilon^2 - (p_d - p^*)^2 - \frac{h_d^2 h^{*2}}{(w_d^2 + h_d^2)}(\frac{w_d}{h_d} - \frac{w^*}{h^*})^2 \tag{23}$$

$$= \epsilon^2 - (p_d - p^*)^2 > 0 \tag{24}$$

Denoting constants $C_a = \frac{\epsilon^2 - (p_d - p^*)^2}{(w_d^2 + h_d^2)}$ and $C_b = \frac{w_d w^* + h_d h^*}{(w_d^2 + h_d^2)}$, we get the range of r such that $v_r \in \mathcal{B}(v^*, \epsilon)$ as

$$\max\left(r_{\min}, \frac{\log(C_b - \sqrt{C_a})}{\log(1+\lambda)}\right) \leq r \leq \min\left(r_{\max}, \frac{\log(C_b + \sqrt{C_a})}{\log(1+\lambda)}\right) \qquad \text{if} \quad C_b > \sqrt{C_a} \tag{25}$$

$$\text{Or} \qquad r_{\min} \leq r \leq \min\left(r_{\max}, \frac{\log(C_b + \sqrt{C_a})}{\log(1+\lambda)}\right) \qquad \text{if} \quad C_b \leq \sqrt{C_a} \tag{26}$$

We further denote this range as $r \in [C_{\min}(\epsilon, v^*, v_d; \lambda), C_{\max}(\epsilon, v^*, v_d; \lambda)]$, with $r_{\min} \leq C_{\min}(\epsilon, v^*, v_d; \lambda) < C_{\max}(\epsilon, v^*, v_d; \lambda) \leq r_{\max}$. Based on the independent uniform sampling assumption, the probability of the event $\neg A$ that none of the n samples fall into the ϵ-neighborhood around the optimum $\mathcal{B}(v^*, \epsilon)$ is

$$P(\neg A) = \left(1 - \frac{C_{\max}(\epsilon, v^*, v_d; \lambda) - C_{\min}(\epsilon, v^*, v_d; \lambda)}{r_{\max} - r_{\min}}\right)^n = (1 - C_e(\epsilon, v^*, v_d; \lambda))^n \tag{27}$$

where we use $C_e(\epsilon, v^*, v_d; \lambda) \in (0, 1]$ to denote the constant value depending on $\epsilon, v^*, v_d, \lambda$. Following Eq. (12), we then have

$$h_{\pi_e}(v^*, n) \leq \delta + (1 - C_e(\epsilon, v^*, v_d; \lambda)))^n \tag{28}$$

where the second term goes to 0 as n is increasing to larger values.

Discussion. In the above, we demonstrated that beginning with the initial detected visual context v_d, under certain mild conditions, acquiring n samples according to a distribution $\pi(\cdot|v_d)$ is an efficient method for identifying a sample that leads to a small bounded deviation in the token decoding probabilities from those derived from the optimal visual context v^*. The more samples acquired, the tighter the bound is. This provides a simple and robust way of approximating the optimum.

Different sampling distributions have distinct characteristics. For normal distribution sampling $\pi_g(\cdot|v_d) \sim \mathcal{N}(v_d, \sigma^2 I)$, the variance parameter σ^2 determines the spread of the samples and thus the likelihood of approximating the optimal v^* within $\mathcal{B}(v^*, \epsilon)$. For exponential expansion sampling $\pi_e(\cdot|v_d) \sim \mathcal{U}(r \in [r_{\min}, r_{\max}])$ with samples $v_r = ((1+\lambda)^r w_d, (1+\lambda)^r h_d, p_d)$, the parameter λ controls the rate of growth for the sampled visual contexts. In practice, we apply discrete integer values of r to acquire different samples efficiently, thus λ affects the sample coverage of the visual information around v^*.

The choice of the sampling distribution π is contingent upon factors such as the quality of the detector $\mathcal{G}_d$, the LVLM backbone $\mathcal{M}_\theta^{\text{LVLM}}$, the textual query x, and the visual input v. Specifically, the continuous normal distribution is advantageous for concentrated sampling around v_d, which is particularly effective when the detection perturbation η is small (meaning v_d is near v^*). In contrast, exponential expansion sampling covers an extended range of visual contexts quickly, which is preferable when limited context information is obtained. In scenarios where significant underestimation or overestimation in G_d detection is present, the exponential expanding strategy can discover the optimal visual context more effectively. □

B Time Cost Analysis

Since time complexity is a critical aspect for VLM decoding algorithms, in this section we analyze the additional runtime overhead and time cost of HALC. According to Biber et al. (2000), nouns, adjectives, adverbs, numbers, verbs, and pronouns, which are tokens that will actually pass through HALC decoding, comprise approximately 35% of the total words in modern English (we observe similar sparse patterns in our experiments). POS tagging is observably fast in practice (we used the spaCy package, which is highly optimized on CPU with the smallest tagger model, which is only 12 MB in size[7]). Thus we

[7] https://spacy.io/models/en#en_core_web_sm

will mainly discuss the time cost w.r.t. other modules in HALC.

For each individual token, after its original decoding, HALC will utilize the detection module to initialize the FOV sampling, for which we use T_d to represent the detector time cost. Next, each one of the n FOVs (in our experiments, $n = 4$, as shown in Table 8) is fed back into the LVLM for decoding, resulting in $n * T_{LVLM}$ time cost, where T_{LVLM} represents the LVLM decoding time for a single step (although this may increase slightly as the sequence grows longer). Other computations on top of the multiple decodings such as contrasting the distributions can be ignored in comparison. Therefore, in summary, without any parallelization, for a sequence of L tokens, HALC will cost approximately:

$$L * T_{LVLM} + L * 0.35 * (T_d + n * T_{LVLM}) = L * ((1 + 0.35n) * T_{LVLM} + 0.35T_d) \tag{29}$$

In practice, when $n = 4$ and T_d is relatively much smaller than T_{LVLM} (the detection model Grounding DINO we used was based on the Swin-Tranformer[8] with 341M parameters), we expect HALC to cost around 2.4x of the normal greedy decoding time expense.

However, the decoding passes for the extra n FOVs can essentially run **in parallel** as they do not depend on each other. With parallelization, the time cost with n FOV decoding is equal to the time cost for 1 FOV decoding, so the expected time cost will be only approximately 1.35x of the greedy decoding. When the detection model time can not be ignored and in the worst case it is the same as the decoding step time (which is unlikely as the LVLMs we experimented with are 7B), the expected time cost would be 1.7x of the normal greedy decoding.

C Experimentation Details

C.1 Experimental Setups

The overall experiment settings is reported in Table 7. While the regular greedy decoding follows this setting, the beam search variant in our experiment essentially applies a token-wise beam search based on accumulated probability scores of the previous tokens $y_{<t}$. We use the default code for implementation of these two baselines in HuggingFace Transformers Repository (Wolf et al., 2020).[9]

Table 7: Overall Experiment Settings

Parameters	**Value**
Maximum New Tokens (CHAIR)	64
Maximum New Tokens (POPE)	64
Maximum New Tokens (MME)	128
Top-k	False
Top-p	1
Temperature τ	1

The complete hyper-parameters for HALC in our experiments in §6 is reported in Table 8. Specifically, there are four major hyper-parameters that can actively adjust the effectiveness of HALC to adapt to different task settings:

1. *FOV Sampling Distribution*: Typically, a normal distribution, which concentrated around v_d, provides a tighter bound under minimal perturbations, while an exponential expansion sampling distribution, with a more averaged coverage of the sampling space, is preferable when less contexts of the task is available. Thus to preserve generality in our experiment, we have employed the exponential expansion sampling with exponential growth factor $\lambda = 0.6$.

2. *Number of Sampled FOVs* n: n determines the number of sampled FOVs in the sample space. According to Theorem 5.1, while increasing n and adjusting the distribution parameters can efficiently reduce minimum token probability deviations and enhance the robustness against perturbed initial detection, it's notable that the runtime costs also raise with n. Consequently, we set $n = 4$ across all our experiments.

3. *JSD Buffer Size* m: For each beam in the overall beam search process (beam size k), our bi-adaptive visual grounding module samples n visual contexts, which through interpolated JSD calculation would produce $\frac{n \cdot (n-1)}{2}$ JSD values

[8] https://huggingface.co/docs/transformers/model_doc/swin
[9] https://huggingface.co/docs/transformers

in total. Then we select the top m FOV pairs with relatively large discrepancy to produce contrastive candidate distributions.

4. *Beam Size* k: The beam size k is set to adjust the diversity and range for HALC to search for the best candidate captions. Essentially, the global visual matching score module selects the top k diverse captions from $2m \cdot k$ text sequence candidates passed from the local adaptive visual grounding module. While a larger k involves a larger search space and hopefully a better generation, the runtime cost also raises linearly w.r.t. k. HALC adopts Bootstrapping Language-Image Pre-training (BLIP) (Li et al., 2022a) for both text and image encoding when computing their cosine similarity scores. Notably given the global search capability of our visual matching score module, HALC seeks to preserve a more diverse set of captions within the beam buffer.

5. *Other Hyperparameters*: Our implementation inherits an additional hyperparameter, adaptive plausibility threshold, originally from DoLA (Chuang et al., 2023).

Table 8: HALC Hyperparameter Settings

Parameters	**Value**
Amplification Factor α	0.05
JSD Buffer Size m	6
Beam Size	1
FOV Sampling	Exponential Expansion
Number of Sampled FOVs n	4
Exponential Growth Factor λ	0.6
Adaptive Plausibility Threshold	0.1

Regarding the comparison of HALC with SOTAs that are specifically designed for OH mitigation, we adopt the code, hyper-parameters, and pre-trained models of each method outlined in their public repositories and papers respectively. Specifically, the hyper-paratermers for DoLa (Chuang et al., 2023)[10] is reported in Table 9; OPERA (Huang et al., 2023)[11] is reported in Table 10; and the hyperparatermers for VCD (Leng et al., 2023)[12] is reported in Table 11. For each of these baselines, we strictly follow their implementations and default hyper-parameters as reported in the paper to reproduce their results.

Table 9: DoLa Hyperparameter Settings

Parameters	**Value**
Repetition Penalty θ	1.2
Adaptive Plausibility Threshold β	0.1
Pre-mature Layers	$[0, 2 \cdots, 32]$

Table 10: OPERA Hyperparameter Settings

Parameters	**Value**
Self-attention Weights Scale Factor θ	50
Attending Retrospection Threshold	15
Beam Size	3
Penalty Weights	1

[10] https://github.com/voidism/DoLa
[11] https://github.com/shikiw/OPERA
[12] https://github.com/DAMO-NLP-SG/VCD

Table 11: VCD Hyperparameter Settings

Parameters	Value
Amplification Factor α	1
Adaptive Plausibility Threshold	0.1
Diffusion Noise Step	500

Regarding post-hoc correction method woodpecker (Yin et al., 2023)[13] and LURE (Zhou et al., 2023)[14] , we also strictly follow their implementations and hyper-parameters as reported in the paper to reproduce their results. For woodpecker, we adopt their original code and use OpenAI API to access GPT-3.5 Turbo. In average, per 500 images would result in approximately $4.5 cost. For LURE, we also directly adopt their pre-trained projection layer model (based on Minigpt4) to reproduce the results reported in this paper. All the hyper-parameters are default.

Notably, to construct a standardized evaluation platform, we reorganize these repositories and form a unified object hallucination evaluation benchmark released at `https://github.com/BillChan226/HALC`. This benchmark repository provides at ease a unified access to most of the announced LVLMs for various VQA tasks, evaluated by CHAIR (Rohrbach et al., 2018) , POPE (Li et al., 2023), offline POPE (OPOPE), linguistic quality metrics and MME scores (Fu et al., 2023) in a standardized pipeline.

C.2 Empirical Studies on Optimal Visual Contexts

We verify our insight that optimal visual context is important in correcting object hallucination through an empirical pilot study. Fig. 1 shows the oracle performance of OH levels when we rely on optimal visual contexts for tokens through brute-force search, with greedy decoding on the MME benchmark (Fu et al., 2023) on three categories of OH sources. Specifically, each MME sub-task contains 30 images, and we have followed (Leng et al., 2023) and selected four sub-tasks (including *existence*, *count*, *color*, *position*) to evaluate the hallucination in our analysis, in total 110 distinct images. Based on these images, we manually constructed multiple challenging questions (2-4 per image) that are likely to induce the LVLM to hallucinate (e.g. queries based on co-occurrence statistics illustrated in (Li et al., 2023) on some plausible but unfaithful objects that are likely to co-occur, some minor objects in the distance). Then we take each question as a count unit and calculate the number of hallucinations on word level (instead of token level) which could be attributed for each of the three sources. Then for each question with a hallucination occurring, we search across the original image input using a brutal-force breadth-first algorithms until the hallucinating token is corrected to be consistent with the ground truth. This process effectively succeeds to retrieve the optimal visual context for 54.0% of the questions. For those questions that fail this brutal-force search, we further manually select the visual context candidates based on human priors. In total, 84.5% of the questions that contain these three sources of hallucinations can be eliminated with an explicit optimal visual context v^*.

D MME Experiment Details

The experiment details mostly follow Appendix C.2, where we adopt each sub-task of 30 images from the MME benchmark dataset[15], and reconstruct the question prompt following offline POPE. Specifically, instead of simply asking a question with a binary yes/no answer, we first ask the decoder to generate a detailed caption of the provided image and then check whether the target positive/negative word existes in the caption. The detailed results are reported in Table 12. The corresponding figure result is shown in Fig. 5.

[13] `https://github.com/BradyFU/Woodpecker`
[14] `https://github.com/YiyangZhou/LURE`
[15] `https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation`

Table 12: Comparison of Decoder Performances on 4 MME sub-tasks

Decoder	Existence	Position	Color	Count	Max Tokens	Num of Samples
HALC	155	73.33	141.67	93.33	128	110
Greedy	145	63.33	118.33	85	128	110
DoLa	145	60	118.33	85	128	110
Opera	135	56.67	115	80	128	110
VCD	135	70	133.33	70	128	110
LURE	140	60	108.33	68.33	128	110

E POPE Results

Although we argue that POPE is not suitable for post-correction decoding methods and as a result we propose OPOPE. We also conduct the POPE evaluation and demonstrate the result in Table 13. To adapt HALC for the original POPE benchmark, we use the **entire query** together with the initial answer (yes/no) as the text prompt for the detection model to provide a grounding for the focal area of the query.

Table 13: Detailed POPE results with random, popular and adversarial samplings.

Setting	**Model**	**Decoding**	**Accuracy**	**Precision**	**Recall**	**F_1 Score**
Random	MiniGPT-4	Greedy	61.00	56.32	98.00	71.53
		Beam Search	58.00	54.47	97.33	69.86
		OPERA	57.66	54.21	98.67	69.97
		VCD	60.33	57.87	76.00	65.71
		HALC	61.33	56.54	98.00	**71.70**
Popular	MiniGPT-4	Greedy	55.33	52.87	98.00	68.69
		Beam Search	50.33	50.17	97.33	66.21
		OPERA	51.00	50.51	98.67	66.82
		VCD	57.33	55.05	80.00	65.21
		HALC	55.67	53.07	98.00	**68.85**
Adversarial	MiniGPT-4	Greedy	54.00	52.15	96.7	67.76
		Beam Search	52.00	51.05	97.33	66.97
		OPERA	52.67	51.39	98.67	67.58
		VCD	53.67	52.53	76.00	62.13
		HALC	56.00	53.26	98.00	**69.02**

According to Table 13, HALC has also outperformed the other four methods by a large margin in terms of accuracy, precision, recall and F1 Score on all three types of POPE VQA tasks (random, popular, adversarial).

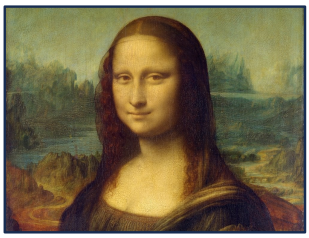
F Comprehensive OPOPE Results

Table 14: Detailed OPOPE results with random, popular and adversarial samplings.

Setting	Model	Decoding	Accuracy	Precision	Recall	$F_{0.2}$ Score
Random	MiniGPT-4	Greedy	68.30	97.24	37.67	91.67
		Beam Search	68.37	96.30	38.20	90.98
		DoLa	68.50	97.27	38.07	91.78
		OPERA	68.67	96.98	38.53	91.63
		VCD	67.10	96.22	35.60	90.30
		Woodpecker	69.07	96.99	39.366	91.83
		LURE	69.50	96.65	40.4	86.76
		HALC	67.90	97.36	40.4	91.74
	LLaVA-1.5	Greedy	72.20	97.17	45.73	93.14
		Beam Search	71.33	97.48	43.80	93.09
		DoLa	72.30	96.78	46.13	92.86
		OPERA	71.20	96.76	43.87	92.47
		VCD	72.07	96.89	45.60	92.87
		Woodpecker	70.83	95.89	43.53	91.65
		LURE	71.67	97.24	44.6	93.02
		HALC	71.87	97.86	44.73	93.58
	mPLUG-Owl2	Greedy	71.27	96.91	43.93	92.62
		Beam Search	70.50	97.26	42.20	92.61
		DoLa	71.47	96.92	44.33	92.69
		OPERA	70.17	96.92	41.67	92.22
		VCD	70.93	97.31	43.07	92.81
		Woodpecker	70.27	97.99	41.38	93.09
		LURE	70.83	96.71	43.13	92.30
		HALC	71.50	97.38	44.20	93.07
Popular	MiniGPT-4	Greedy	66.43	88.70	37.67	84.30
		Beam Search	67.00	90.09	38.20	85.62
		DoLa	66.8	89.50	38.07	85.08
		OPERA	66.80	88.65	38.53	84.43
		VCD	65.47	65.47	35.60	83.64
		Woodpecker	67.37	89.47	39.37	85.29
		LURE	67.8	89.38	40.4	85.40
		HALC	66.37	90.02	36.80	85.27
	LLaVA-1.5	Greedy	70.27	89.79	45.73	86.58
		Beam Search	69.80	91.25	43.8	87.6
		DoLa	70.43	89.75	46.13	86.60
		OPERA	69.63	90.51	43.87	86.95
		VCD	70.57	91.08	45.60	87.71
		Woodpecker	69.37	90.07	43.53	86.51
		LURE	69.63	89.32	44.6	86.00
		HALC	70.03	90.74	44.67	87.28
	mPLUG-Owl2	Greedy	69.30	89.13	43.93	85.74
		Beam Search	68.83	90.27	42.20	86.48
		DoLa	69.53	89.35	44.33	85.99
		OPERA	69.03	92.02	41.67	87.94
		VCD	69.43	91.10	43.07	87.35
		Woodpecker	68.58	90.73	41.38	86.75
		LURE	69.17	89.99	43.13	86.38
		HALC	69.63	89.95	44.20	86.50
Adversarial	MiniGPT-4	Greedy	65.60	85.35	37.67	81.38
		Beam Search	66.3	87.21	38.20	83.11
		DoLa	65.87	85.74	38.07	81.80
		OPERA	66.3	86.66	38.53	82.68
		VCD	64.77	85.44	35.60	81.08
		Woodpecker	66.88	87.53	39.37	83.60
		LURE	67.13	86.82	40.4	83.14
		HALC	66.00	88.47	36.80	83.94
	LLaVA-1.5	Greedy	69.23	86.30	45.73	83.44
		Beam Search	68.47	86.45	43.8	83.33
		DoLa	69.33	86.07	46.13	83.30
		OPERA	68.37	86.01	43.87	82.95
		VCD	69.37	86.91	45.60	83.99
		Woodpecker	69.20	89.45	43.53	85.96
		LURE	68.7	86.1	44.6	83.13
		HALC	69.87	90.21	44.67	86.80
	mPLUG-Owl2	Greedy	68.73	87.16	43.93	83.98
		Beam Search	68.27	88.17	42.20	84.63
		DoLa	68.87	87.02	44.33	83.91
		OPERA	68.57	90.22	41.67	86.35
		VCD	69.07	89.69	43.07	86.10
		Woodpecker	67.85	87.94	41.38	84.29
		LURE	67.73	84.91	43.13	81.86
		HALC	69.23	88.50	44.20	85.21

G Experiment Results on LLaVA-Bench

As discussed in §6.3, we leverage LLaVA-Bench (Liu et al., 2023a) as a case study to qualitatively compare the decoding outputs of HALC with other methods. Results generated by HALC and other OH reduction baselines incorporating mPLUG-Owl2 (Ye et al., 2023), MiniGPT-4 (Zhu et al., 2023; Chen et al., 2023), and LLaVA (Liu et al., 2023b) LVLM backbones are shown in Fig. 6, 7 and 8 respectively. In all the plots, red fonts indicate OH, including any of the object existence, attribute or relationship hallucinations.



HALC: The image is a painting of a face, resembling the famous Mona Lisa's portrait. The painting is done in a modern style, with a blend of colors and textures. The woman's face is the main focus of the painting, and her expression appears to be calm and serene. The background is slightly blurred, adding depth to the artwork.

Greedy: The image is a painting of a woman, possibly a Mona Lisa-style portrait, with a smile on her face. She is wearing a necklace and appears to be looking directly at the viewer. The painting is set against a backdrop of a mountain landscape, adding a sense of depth and beauty to the composition.

OPERA: The image features a painting of the famous Mona Lisa, a portrait of a woman attributed to the Renaissance artist Leonardo da Vinci. The Mona Lisa is known for her enigmatic smile and captivating gaze. The painting is displayed on a wall, showcasing the artwork's beauty and historical significance.

VCD: The image is a painting of the famous Mona Lisa, capturing the enigmatic smile of the subject. The artwork is displayed on a wall, and it features a blue background that accentuates the painting's colors and the subject's eyes, making the painting stand out. The Mona Lisa is the central focus of the artwork, showcasing her famous smile and captivating gaze.

HALC: The image showcases a beautiful view of a mountain range with a large, sandy beach in the foreground. The beach is bustling with activity, with numerous cars and trucks parked along the shore. In the background, the city can be seen, stretching across the landscape. The sky above the scene is filled with clouds, adding to the picturesque atmosphere. The combination of the mountain range, the beach, and the city creates a captivating and dynamic scene.

Greedy: The image showcases a beautiful view of a mountain range with a large, sandy beach in the foreground. The beach is bustling with activity, with numerous people enjoying the sunny day. Some of the people are standing closer to the water, while others are scattered along the beach, engaging in various activities. In the background, the mountain range stretches across the horizon, creating a picturesque scene. The combination of the sandy beach, the lush green mountains, and the blue ocean creates a serene and inviting atmosphere.

OPERA: The image showcases a breathtaking view of a mountainous area overlooked by water. A sandy beach stretches along the coast, surrounded on three sides by mountains. In addition to the beach, there are several boats visible in the water, adding to the picturesque scene. A few people can be seen scattered around the beach, enjoying the beautiful surroundings. Some of them are closer to the water, possibly swimming or sunbathing, while others are further back on the beach.

VCD: The image depicts a picturesque view of a mountainous landscape with a beautiful blue ocean in the background. The ocean water is visible near the bottom of the image, with a sandy beach extending along the shoreline. The water appears to be calm and inviting, creating a serene atmosphere. The mountainous area is characterized by a series of hills and mountains, with some of them rising steeply and others gently. A few people can be seen scattered throughout the scene, likely enjoying the natural beauty of the area.

Figure 6: LLaVA-Bench results comparing HALC and other methods with mPLUG-Owl2 (Ye et al., 2023) backbone.

HALC: The image shows a dining table with a variety of items on it, including a bowl of noodles, a cup of green tea, and some other dishes. The bowl is placed in the center of the table, surrounded by several bowls containing different foods. A cup of green tea is near the top right corner of the table. A spoon can be seen resting on the table, close to the bowl of noodles.

Greedy: The image shows a bowl of noodles with chopsticks and a spoon on the side. The bowl of noodles is filled with a variety of ingredients, including meat, vegetables, and sauce. The background of the image is a wooden table with a red and white checkered tablecloth. There are also two glasses of beer on the table.

OPERA: The image shows a bowl of ramen noodles with chicken and vegetables on the side. The bowl of ramen noodles has a clear broth with chicken and vegetables on the side. The noodles are cooked and have a chewy texture. The chicken is cooked and has a crispy texture. The vegetables are cooked and have a crunchy texture. The table is made of wood and has a red and white checkered tablecloth on it.

VCD: The image is a table with a bowl of noodles and some vegetables in it. There are also some chopsticks on the table. There are two chairs on either side of the table, and a vase of flowers on the table. The background of the image is a wall with a large window with a view of a city street.

HALC: The image is a toy of a dinosaur with glasses on its face. The dinosaur is made of plastic and has a red and orange flame on its back. The flame is made of plastic and has a red and orange color. The dinosaur is standing on its feet and has a pair of glasses on its face. The dinosaur has a smiling expression on his face.

Greedy: The image is a toy of a dinosaur with glasses on its face. The dinosaur is made of plastic and has a red and orange color scheme. The glasses are made of plastic and have a black frame and lenses. The dinosaur is standing on its hind legs and has a long tail. The image is a toy of a dinosaur with glasses on its face. The dinosaur is made of plastic and has a red and orange color scheme.

OPERA: The image is of a toy dinosaur with glasses on its face. The dinosaur is made of plastic and has a red and orange flame on its back. The dinosaur is standing on its hind legs and has a smile on its face. The image is of a toy dinosaur with glasses on its face. The dinosaur is made of plastic and has a red and orange flame on its back. The dinosaur is standing on its hind legs and has a smile on its face.

VCD: The image is a toy flamingo with a red body and orange wings. It has a pair of glasses on its head and is standing on its hind legs. The toy is made of plastic and has a smooth surface. The image is a toy flamingo with a red body and orange wings. It has a pair of glasses on its head and is standing on its hind legs. The toy is made of plastic and has a smooth surface. The flamingo has a cartoonish appearance, with large eyes and a smiling expression.

Figure 7: LLaVA-Bench results comparing HALC and other methods with MiniGPT-4 (Zhu et al., 2023; Chen et al., 2023) backbone.

HALC: The image features a man standing on the back of a yellow taxi cab, holding a blue white towel. He appears to be ironing clothes while riding in the taxi. The taxi is driving down a busy with other such as a and a car visible in the scene. The man manages to balance himself on the back of the taxi, making the scene quite unique and interesting.

Greedy: The image features a man standing on the back of a yellow taxi cab, holding a blue shirt. He appears to be ironing the shirt while the taxi is in motion. The taxi is driving down a busy street, with other vehicles such as a truck and a car visible in the scene. Additionally, a traffic light can be seen in the background, indicating that the taxi is driving through an intersection.

OPERA: In the image, a man is standing on the back of a yellow taxi cab, ironing clothes. He is wearing a yellow shirt and appears to be focused on his task. The taxi is surrounded by other vehicles, including a truck on the left side and a car on the right side of the scene. Additionally, there is a handbag placed near the center of the scene, possibly belonging to the man ironing clothes or someone else in the vicinity.

VCD: The image depicts a busy street scene where a man is standing on top of a car, ironing clothes. He is positioned towards the center of the scene, and there are multiple other people visible in the area. Various vehicles are present on the street, including a truck on the left side, a bus on the right side. There are also a few more cars on the right side of the image, near the background.

HALC: The image features a young boy wearing a suit and tie, with a red bow tie. He is wearing glasses and has a smiling expression on his face. The boy appears to be the main focus of the scene, with the rest of the image being blurry. The suit and tie give him a formal and polished appearance.

Greedy: The image features a young man wearing a suit and tie, with a red bow tie. He is wearing glasses and has a smiling expression on his face. The man appears to be the main focus of the scene, with the rest of the image being blurry. The suit and tie combination, along with the red bow tie, gives the impression that the man is dressed up for a special occasion or event.

OPERA: The image features a young boy wearing glasses and a red bow tie, standing in front of a building. He appears to be the main focus of the scene. The boy is wearing a suit and tie, giving him a formal and polished appearance. The building in the background adds to the overall setting of the image.

VCD: The image depicts a young boy wearing a blue suit with a red bow tie. He is dressed in a business suit, and his face is adorned with a pair of glasses. The boy has a charming smile and appears to be posing for a picture. In the background, there is a person partially visible, possibly taking the photo.

Figure 8: LLaVA-Bench results comparing HALC and other methods with LLaVA (Liu et al., 2023b) backbone.