

From Uncertainty to Trust: Enhancing Reliability in Vision-Language Models with Uncertainty-Guided Dropout Decoding

Yixiong Fang^{*,1}, Ziran Yang^{*,1}, Zhaorun Chen², Zhuokai Zhao^{†,2}, Jiawei Zhou^{†,1}
¹Stony Brook University ²University of Chicago
^{*}Equal contribution [†]Joint last author

Abstract

Large vision-language models (LVLMs) demonstrate remarkable capabilities in multimodal tasks but are prone to misinterpreting visual inputs, often resulting in hallucinations and unreliable outputs. To address these challenges, we propose DROPOUT DECODING, a novel inference-time approach that quantifies the uncertainty of visual tokens and selectively masks uncertain tokens to improve decoding. Our method measures the uncertainty of each visual token by projecting it onto the text space and decomposing it into aleatoric and epistemic components. Specifically, we focus on epistemic uncertainty, which captures perception-related errors more effectively. Inspired by dropout regularization, we introduce uncertainty-guided token dropout, which applies the dropout principle to input visual tokens instead of model parameters, and during inference rather than training. By aggregating predictions from an ensemble of masked decoding contexts, DROPOUT DECODING robustly mitigates errors arising from visual token misinterpretations. Evaluations on benchmarks including CHAIR, THRONE, and MMBench demonstrate that DROPOUT DECODING significantly reduces object hallucinations (OH) and enhances both reliability and quality of LVLM outputs across diverse visual contexts. Code is released at <https://github.com/kigb/DropoutDecoding>.

1. Introduction

Recent advancements in large vision-language models (LVLMs) have demonstrated impressive capabilities [11, 15, 17, 54, 59, 62, 66], in tasks such as image captioning [58], visual question answering (VQA) [1, 22, 29, 61], multimodal reasoning [6, 37, 39, 40, 63] and so on. However, LVLMs still face challenges in accurately perceiving and interpreting visual inputs, leading to inaccurate

outputs and hallucinations [35]. These issues often stem from LVLMs misrepresenting key image elements or overlooking critical details, compromising the reliability of their outputs in tasks demanding precise visual understanding [4, 5, 16, 57].

In practice, LVLMs typically process visual inputs token by token, which we refer to as *visual tokens*.¹ This can fall short in effectively focusing on the most informative parts of the visual context. While attention mechanisms are designed to prioritize relevant information, they are not always perfect [46, 52], especially when the inputs are complex or ambiguous for the model, or in other words, of high *uncertainty*. Existing methods to address these challenges in the training stage often involve fine-tuning on specific tasks [33, 34, 49, 56], or using additional supervision signals especially at lower level to guide the model [7, 55]. However, these approaches are resource-intensive and not easily extensible to new tasks. Alternative inference-time strategies, such as attention-based or logits-based mechanisms on decoding correction [4, 21, 45, 51, 60], attempt to identify important regions in the input without additional training, but they typically rely on heuristic design choices and largely increase inference cost. Therefore, enhancing the trustworthiness of LVLMs and reducing hallucinations require more principled methods that can more effectively emphasize the most informative parts of the visual input.

To address this challenge, we propose a novel approach that quantifies uncertainty in visual token contexts and removes uncertain tokens, both directly at inference time to improve the reliability of LVLM outputs. Inspired by traditional dropout [47] techniques—typically applied to model parameters but difficult to implement directly in pretrained LVLMs [13, 24]—we introduce *token dropout*, which applies the dropout principle to input context tokens instead of model parameters. Furthermore, it is applied to regularize the inference process instead of training, by introducing randomness in decoding contexts to reduce overfitting to noisy visual tokens.

^{*}Work was done during Yixiong Fang and Ziran Yang’s remote internship at Stony Brook University. Correspondence to {kfangyixiong, ziranyang0}@gmail.com.

¹We specifically refer to the tokens that are already in the input prompt to the text decoder. Concrete definition is in §3.1.

Our method measures the uncertainty of each visual token by projecting it into the text token space *through the text decoder directly*, and decomposing this uncertainty into two components: *aleatoric* (data-related) and *epistemic* (model-related) [20, 43, 50]. By focusing on epistemic uncertainty, which reflects the model’s lack of knowledge, we identify visual tokens with high uncertainty and selectively target them for suppression. At inference time, we adjust the visual inputs by selectively suppressing tokens with high epistemic uncertainty. Specifically, we create an *ensemble of predictions* by generating multiple subsets of visual inputs, each with different combinations of high-uncertainty tokens dropped out. These subsets are processed independently, and their corresponding outputs are aggregated using majority voting to produce the final prediction.

This approach, which we term **DROPOUT DECODING**, enhances the reliability and accuracy of LVLM outputs without modifying the underlying model parameters or requiring additional training. By leveraging uncertainty quantification and token dropout, **DROPOUT DECODING** robustly mitigates errors arising from uncertain visual token interpretations. Experiments are conducted on LVLM decoding benchmarks including CHAIR [41], THRONE [23], and MMBench [36], demonstrating the effectiveness of our approach in both reducing object hallucinations (OH) and improving reliability of model outputs consistently across diverse visual contexts. With **DROPOUT DECODING**, we make the following contributions:

- We introduce a novel approach that quantifies and decomposes uncertainty on tokens in the visual inputs at inference time without additional supervision, by projecting visual input tokens onto text token interpretations.
- We propose a decoding strategy that uses epistemic uncertainty measurements to guide the selective dropout of high-uncertainty visual tokens in the context, analogous to performing dropout on the model but applied to the input tokens and during inference.
- Comprehensive experiments are conducted on various benchmarks, showing significant reductions in OH and improved fidelity in pre-trained LVLMs without additional fine-tuning.

2. Related Work

Reliable generation. Reliable generation in LLMs is often challenged by hallucinations, where the model generates irrelevant or factually incorrect information [18, 48, 64]. These hallucinations stem from issues in data, training, and inference stages [53], with attention mechanisms exacerbating the problem as sequence lengths grow [8]. To mitigate these, methods like factual-nucleus sampling have been proposed to balance output diversity and accuracy [26]. Besides, while Arias et al. [2] leverage quantified uncertainty to guide the decoding process for LLM, our method differs

significantly. We quantify uncertainty at the level of visual input context rather than of model ensemble which is heavy.

OH in LVLMs. OH is a common issue in LVLMs, where models generate descriptions containing objects, attributes, or relationships not present in the actual image. The CHAIR metric [41] is widely used to evaluate OH, measuring the hallucination rate on the MSCOCO dataset [31]. Another benchmark, POPE [30], treats object hallucination as a binary classification task. More recently, THRONE [23] takes a more holistic approach, using open-ended, object-based image descriptions for evaluation. In our work, we use CHAIR and THRONE to assess OH.

OH reduction. Recent methods addressing OH in LVLMs include internal signal guidance, contrastive decoding, and selective information focusing, all of which are inference-time strategies. OPERA [19] uses internal signals like attention patterns to refine outputs and improve alignment with visual content. Contrastive decoding methods, like VCD [27], enhance coherence by comparing image-specific outputs. Selective information focusing approaches, such as HALC [4], prioritize key image regions, while CDG [10] uses CLIP embeddings to align generation with visual input. In contrast, **DROPOUT DECODING** works with any LVLM by 1) selecting visual information from visual tokens during generation, unlike HALC which selects regions initially; 2) using uncertainty to guide visual information selection, requiring no external models, unlike HALC and CDG; 3) introducing a token-level majority voting strategy.

3. Preliminaries

3.1. Vision-Language Model Decoding

Widely adopted LVLM architectures [28, 32, 33] typically include a vision encoder, a vision-text interface module, and a Transformer-based LLM decoder. As we mostly focus on the decoder side inference optimization, we assume the LLM decoder is with parameter θ .

The visual input, such as an image, is segmented into patches and processed by the vision encoder,² followed by the vision-text interface module, to produce a sequence of *visual tokens* $x^v = (x_1^v, x_2^v, \dots, x_N^v)$. Each token x_i^v is a contextualized embedding of an image patch, serving as the direct input to the text decoder. The text input such as a query or instruction is $x^t = (x_1^t, x_2^t, \dots, x_M^t)$. The input to the text decoder is denoted as $x = [x^v, x^t]$, which is the concatenation of visual and text tokens. At this point, the visual and text tokens are aligned and serve as a sequential input to the LLM decoder.

During autoregressive decoding, the decoder generates output text tokens $y = (y_1, y_2, \dots)$ as continuation from

²We assume a general Transformer architecture for the vision encoder as well. Our approach could also apply to other types of vision encoders.

prompt x , following the conditional probability distribution

$$\begin{aligned} h_j &= f_\theta(x^v, x^t, y_{<j}) \\ p_\theta(y_j | x^v, x^t, y_{<j}) &= \text{softmax}(W_{\mathcal{V}}h_j) \end{aligned} \quad (1)$$

where $y_{<j} = (y_1, \dots, y_{j-1})$ is the sequence of previously generated tokens, f_θ denotes the LLM forward pass to produce hidden states $h_j \in \mathbb{R}^d$ on top of the Transformer layers, $W_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the output projection matrix onto the text vocabulary \mathcal{V} , and $y_j \in \mathcal{V}$ the output token at j -th step.

3.2. Uncertainty Quantification

Our approach quantifies the information uncertainty of visual tokens used for decoding by adapting the concept of epistemic uncertainty for measurement, as detailed in §5, and drawing inspiration from classical uncertainty decomposition [20, 43, 44]. To provide the necessary background, we first introduce the concept of uncertainty decomposition.

Uncertainty decomposition separates the total uncertainty of a model’s prediction into two components: *aleatoric* uncertainty, which is inherent to the data, and *epistemic* uncertainty, which relates to the model’s lack of knowledge. The Bayesian framework offers a principled way to quantify uncertainty about some candidate model with weights w , through the posterior estimation over the hypothesis space for a given dataset \mathcal{D} . The Bayesian model average (BMA) predictive distribution is defined as³

$$p(y | x, \mathcal{D}) = \int_w p(y | x, w)p(w | \mathcal{D}) dw. \quad (2)$$

The total information uncertainty is measured by the entropy of BMA: $\mathbb{H}[p(y | x, \mathcal{D})]$, which equals the posterior expectation of the cross-entropy between the predictive distribution of the candidate model and the BMA distribution:

$$\begin{aligned} \underbrace{\mathbb{H}[p(y | x, \mathcal{D})]}_{\text{Total Uncertainty}} &= \mathbb{E}_{p(w|\mathcal{D})} [\text{CE}[p(y | x, w), p(y | x, \mathcal{D})]] \\ &= \underbrace{\mathbb{E}_{p(w|\mathcal{D})} [\mathbb{H}(p(y | x, w))]}_{\text{Aleatoric Uncertainty}} \\ &\quad + \underbrace{\mathbb{E}_{p(w|\mathcal{D})} [D_{\text{KL}}(p(y | x, w) \| p(y | x, \mathcal{D}))]}_{\text{Epistemic Uncertainty}} \end{aligned}$$

The epistemic uncertainty, expressed as the KL divergence between candidate models’ predictive distributions and the BMA, has proven effective in various applications [3, 13, 38, 65]. Our approach, adopts a similar formulation for uncertainty quantification, calculating the KL divergence between candidate prediction distributions on individual visual tokens and an aggregated average distribution.

³ $p(y | x, w, \mathcal{D}) = p(y | x, w)$ because of conditional independence.

4. Textual Interpretation of Visual Tokens

As discussed in §1, identifying the visual tokens that carry significant information and quantifying their uncertainty is critical for improving the reliability of LVLMs. To address this, we propose a supervision-free, scalable approach that maps visual tokens to the text token space, effectively translating visual content into an interpretable text-based representation. This mapping acts as a heuristic for understanding visual tokens, leveraging the LVLm’s inherent ability to align visual and textual contexts.

Text-space projection of visual tokens. While LVLMs are trained to generate text *only after* processing all visual tokens x^v and text instruction tokens x^t , the hidden representations h on top of the text decoder layers inherently capture textual semantics. This is due to their proximity to the text vocabulary projection, even at visual token positions where the model is *not explicitly trained to generate text*.

Building on this intuition, we adopt a heuristic approach to interpret visual tokens by projecting them onto the text vocabulary at the top Transformer layers. In particular, for each visual token x_i^v at position i ,⁴ we obtain its textual projected distribution over the vocabulary \mathcal{V} from the last layer of the LLM decoder in the LVLm as:

$$\begin{aligned} h_i^v &= f_\theta(x_{\leq i}^v) \\ q_i^{\text{proj}} &= p_\theta(\cdot | x_{\leq i}^v) = \text{softmax}(W_{\mathcal{V}}h_i^v) \end{aligned} \quad (3)$$

where h_i^v is the LLM decoder top-layer hidden representation aligned at the i -th visual token positions, $x_{\leq i}^v$ denotes the visual tokens up until index i .⁵

Here, q_i^{proj} , which we refer to as *visual-textual distribution*, represents the projection of the visual input onto the text space. It encapsulates the model’s interpretation of the i -th visual token. This projection offers a text-based summarization, akin to an unordered caption or a “bag-of-words” representation of the visual content. As we will demonstrate in §6, this heuristic method serves as an effective proxy for uncertainty estimation.

An illustrative example with projection uncertainty. To illustrate the effectiveness of this projection method and motivate our approach, consider the example shown in Fig. 1. The image is processed into patches, and for five selected patches, we compute their corresponding distribution over the text space. Then we obtain the top-5 predicted text tokens for each.

Some patches produce specific and informative text tokens, often corresponding to meaningful visual content like “Berlin,” “computer,” or “map.” These tokens are relatively

⁴Note that i indexes are only used over visual tokens x^v , not text tokens x^t or generations y .

⁵For the models we use, the visual tokens x^v are all placed before the text tokens x^t in the concatenated sequence x , so $x_{\leq i}^v$ are purely visual tokens. But our approach also applies to other cases.

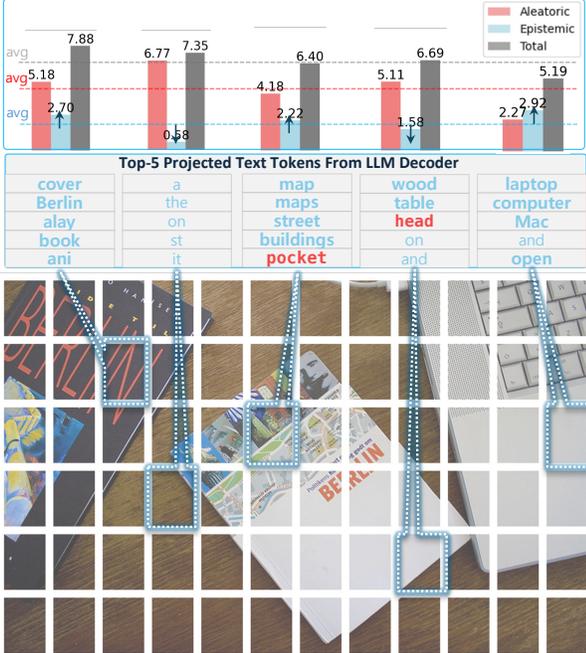


Figure 1. An illustrative example where visual tokens are projected into the text space. We show 5 image patches and their corresponding top-5 projected text tokens. Words in **bold** indicate high information content. Red text highlights projections that are evidently misaligned with the image content. This demonstrates that the visual tokens sometimes cannot perfectly capture the image’s information, which motivates the token dropout (§5.2). We also show the uncertainty values quantified on top (dotted lines are average among all visual tokens for reference). High epistemic uncertainty correlates well with high information in visual tokens, whereas aleatoric and total uncertainty do not.

closer to the long tail in vocabulary, indicating that the corresponding visual tokens capture unique and informative visual context. In contrast, patches resulting in common words carry less specific information, because high-frequency words (e.g. “a,” “the,” or “on”) contribute less to the uniqueness of the visual content. This observation suggests that the projected text tokens can serve as a proxy for the information content of the visual tokens.

Building on these insights, we introduce uncertainty measures derived from textual projective distributions q_i^{proj} , which we use in our method to quantify the uncertainty associated with each visual token, as shown at the top of Fig. 1. Specifically, inspired by §3.2, we decompose the total uncertainty into two components: *aleatoric* (data-related), calculated directly from q_i^{proj} , and *epistemic* (model-related), obtained by comparing q_i^{proj} to an average distribution. Detailed definitions are in §5.1.

As shown in Fig. 1, the epistemic uncertainty accurately reflects the information content of the visual tokens: *visual tokens with high epistemic uncertainty correspond to patches with significant information* (e.g., “Berlin”), while

those with low epistemic uncertainty correspond to less informative patches (e.g., “the”). In contrast, the aleatoric and the total uncertainty do not correlate well. This finding motivates our focus on epistemic uncertainty as a reliable indicator of the significance of visual information.

5. Method

We propose DROPOUT DECODING, which leverages visual uncertainty to selectively drop out visual tokens and guide decoding. As shown in Fig. 2 and Algorithm 1, our approach comprises two stages: uncertainty quantification (§5.1) before decoding and uncertainty-guided token generation (§5.2) for decoding.

5.1. Uncertainty Quantification Before Decoding

Average visual-textual distribution. We begin by defining the averaged distribution q^{proj} , which represents the overall projection of the entire visual input (e.g. an image) into the text space. Using the projected distribution defined in Eq. (3), we define the average projection distribution over all visual tokens as:

$$q^{\text{proj}} = \mathbb{E}_i[q_i^{\text{proj}}] = \frac{1}{N} \sum_i q_i^{\text{proj}} \quad (4)$$

where q_i^{proj} represents the text-space projection of the i -th visual token, and N is the total number of visual tokens. Note that the subscript i indicates different distributions rather than elements within a single distribution. This provides us with a “baseline” representation of the visual input, against which we can quantify the surprisal of a specific visual token. This idea is grounded in classical uncertainty decomposition where a Bayesian average distribution is needed to quantify epistemic uncertainty [20, 43].

Uncertainty measurement for visual tokens. We aim to quantify the uncertainty associated with each visual token at inference time. To distinguish from those uncertainty terms in classical settings as introduced in §3.2, we refer to ours as *perception uncertainty*. We start by quantifying the *perception total uncertainty* of the visual input as the entropy of the average visual-textual distribution $\mathbb{H}[q^{\text{proj}}]$. Then, to attribute this total uncertainty to individual visual tokens, we decompose it (more details in Appendix A) as follows:

$$U_{\text{total}} = \mathbb{H}[q^{\text{proj}}] = \mathbb{E}_i \left[\text{CE} \left(q_i^{\text{proj}}, q^{\text{proj}} \right) \right] \quad (5)$$

Further decomposing the cross-entropy (CE), the perception total uncertainty can be expressed as:

$$\begin{aligned} U_{\text{total}} &= \mathbb{E}_i \left[\mathbb{H} \left[q_i^{\text{proj}} \right] + D_{\text{KL}} \left(q_i^{\text{proj}} \parallel q^{\text{proj}} \right) \right] \\ &= \mathbb{E}_i [U_{\text{ale}}(i) + U_{\text{epi}}(i)] \end{aligned}$$

Here we have the *perception aleatoric uncertainty* of the

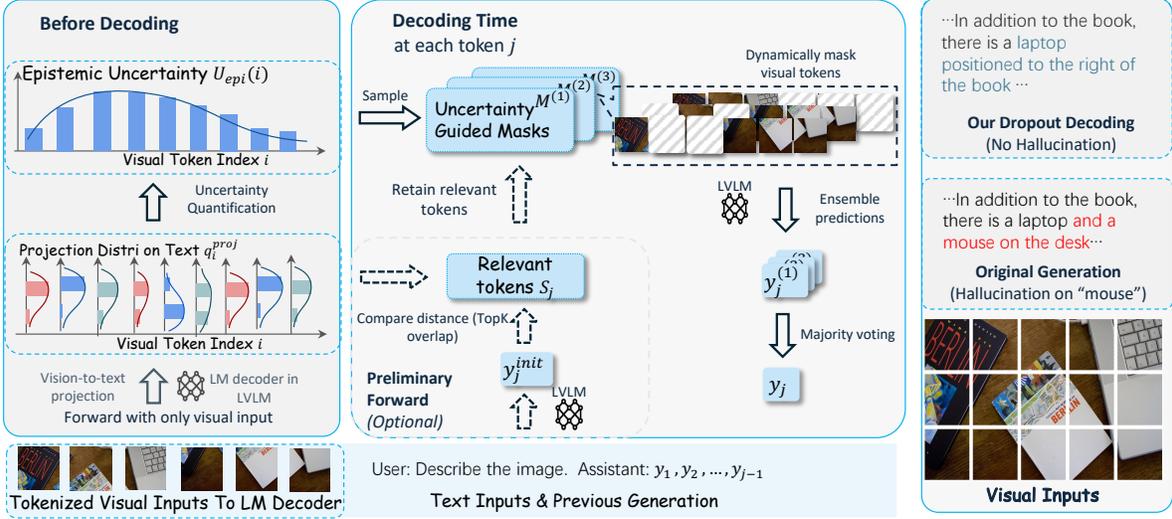


Figure 2. An overview of our DROPOUT DECODING. The method includes uncertainty measurement of visual tokens (under “Before Decoding”) and uncertainty-guided visual context dropout decoding algorithm (under “Decoding Time”). The pseudocode is in Algorithm 1.

i -th visual token $U_{\text{alc}}(i) = \mathbb{H} [q_i^{\text{proj}}]$, capturing the inherent noise or ambiguity of the i -th token, and the *perception epistemic uncertainty*—

$$U_{\text{epi}}(i) = D_{\text{KL}}(q_i^{\text{proj}} \parallel q^{\text{proj}}) \quad (6)$$

quantifying the divergence between the visual token’s textual projection and the overall projection. It indicates how much the model’s belief about this token differs from its belief about the entire visual input. A higher $U_{\text{epi}}(i)$ suggests that the i -th visual token conveys information that is surprising or not well-represented in the overall visual content, which can be critical for identifying tokens that might introduce uncertainty in the decoding process.

5.2. Uncertainty-Guided Decoding

During the text decoding process, we leverage the computed uncertainty measures to guide the generation of each token. Our method involves two main steps for each generated *text* token: (1) identifying relevant visual tokens (optional), and (2) performing *token dropout* with uncertainty-guided masking. The first step is optional, designed to enhance decoding by retaining more relevant visual tokens.

Identifying relevant visual tokens (optional). We selectively retain only the most relevant visual tokens from the context, which are excluded for dropout. To do this, when generating each output text token, y_j , we first perform a preliminary forward pass to generate an initial prediction token y_j^{init} :

$$y_j^{\text{init}} \sim p_{\theta}(\cdot \mid x^v, x^t, y_{<j}) \quad (7)$$

Next, we determine the set of visual tokens that are relevant to this initial prediction. Specifically, a visual token x_i^v

is considered relevant if the initial prediction y_j^{init} appears among the top- k tokens of its visual-textual projection q_i^{proj} . Formally, the set of relevant visual tokens for the j -th generation is:

$$\mathcal{S}_j = \left\{ x_i^v \mid y_j^{\text{init}} \in \text{TopK}(q_i^{\text{proj}}) \right\} \quad (8)$$

where $\text{TopK}(\cdot)$ denotes the function returning the top- k entries of a given distribution.

To illustrate the intuition behind this step, consider an image depicting a cat. Suppose the model correctly predicts the token “cat” during the preliminary forward pass. In that case we retain the visual tokens associated with “cat” and drop out among the remaining visual content. Conversely, if the model incorrectly predicts “dog” or unrelated tokens irrelevant to an object, these predictions will not align with the top text projections of any q_i^{proj} if the visual interpretation is accurate. In such cases, no visual tokens are retained due to a lack of clear relevance, and dropout is applied across the entire visual context as the best alternative.

It is worth noting that this step is optional. Omitting it can improve efficiency by reducing the computational overhead of the preliminary forward pass. As shown by the ablation studies in §7, while skipping this step may lower performance on certain benchmarks like THRONE [23], it still achieves comparable results on others such as CHAIR [41].

Visual token dropout with uncertainty guidance. Using the epistemic uncertainty measurements $U_{\text{epi}}(i)$ from Eq. (6), we introduce dropout masks over visual tokens. As illustrated in Fig. 1, the projected visual-textual distributions sometimes misalign with the image content, and regions of high information can lead to substantial errors, resulting in hallucinations. Building on this intuition, we

selectively target visual tokens with high epistemic uncertainties for dropout.

Specifically, we formulate a controllable series of sample distributions for visual token dropout based on $U_{\text{epi}}(i)$, for each visual position i :

$$P_{\text{dropout}}^{(k)}(x_i^v) = \gamma^{(k)} \left(\frac{U_{\text{epi}}(i) - U_{\text{epi}}^{\min}}{U_{\text{epi}}^{\max} - U_{\text{epi}}^{\min}} \right) + \delta^{(k)} \quad (9)$$

where U_{epi}^{\min} , U_{epi}^{\max} are the minimum and maximum epistemic uncertainty values across all visual tokens, and $\gamma^{(k)}$ and $\delta^{(k)}$ are hyperparameters controlling the probability range of the dropout. By adjusting the values of $\gamma^{(k)}$ and $\delta^{(k)}$, we can modulate the intensity of visual token dropout.

With the dropout distributions, we can sample dropout masks for each visual token independently. Denote the binary mask as $M^{(k)} \in \{0, 1\}^N$, consisting of a binary indicator $M_i^{(k)}$ for each visual token x_i^v , where the corresponding visual token is retained if $M_i^{(k)} = 1$, and dropped if $M_i^{(k)} = 0$. The dropout mask sampling follows $P(M_i^{(k)} = 0) = P_{\text{dropout}}^{(k)}(x_i^v)$, and the sampling is done for each visual token position independently. A higher value of $P_{\text{dropout}}^{(k)}(x_i^v)$ indicates that x_i^v is more likely to be dropped out. If we performed the optional preliminary forward pass to identify relevant visual token set \mathcal{S}_j , these visual tokens are never dropped, *i.e.*, $\forall x_i^v \in \mathcal{S}_j$, set $M_i^{(k)} = 1$ directly.

Ensemble-based reliable generation. Our inference-time context dropout introduces stochasticity, so we employ an ensemble decoding approach by independently sampling K distinct dropout masks, $\{M^{(k)}\}_{k=1}^K$, to enhance generation quality. Since the masks are independent, the text generative distribution from K masks can be efficiently computed in a parallel forward pass

$$y_j^{(k)} \stackrel{\text{Decoding}}{\sim} p_{\theta}(\cdot \mid x_{/M^{(k)}}^v, x^t, y_{<j}) \quad (10)$$

where $x_{/M^{(k)}}^v$ denotes the visual tokens after applying dropout mask $M^{(k)}$, and $\stackrel{\text{Decoding}}{\sim}$ denotes invariance to the decoding algorithm used (e.g., greedy search in our implementation, though others are applicable).

Each $y_j^{(k)}$ serves as a *candidate prediction* for the next text token, with the final token y_j selected via majority voting among the K masked inputs. In case of a tie, we choose the prediction from the forward pass with the fewest dropped tokens, as it retains the most information and is deemed more reliable. By forming an ensemble of predictions derived from various subsets of the visual input, enabled through token dropout, we diversify the model’s perspective on the visual content. This diversity mitigates the impact of any single misinterpretation, ultimately leading to more reliable and robust generation, which is also observed in other ensemble-based methods [4, 12, 14, 25, 42].

Algorithm 1 Pseudocode of DROPOUT DECODING.

- 1: **Input:** visual tokens x^v , Text tokens x^t , Number of dropout masks K , Generation length L
 - 2: **Output:** Generated sequence y
 - 3:
 - 4: **Before Decoding:**
 - 5: Obtain visual text projecting distributions q_i^{proj} . \triangleright Eq. (3)
 - 6: Compute average distribution q^{proj} . \triangleright Eq. (4)
 - 7: Compute epistemic uncertainty $U_{\text{epi}}(i)$. \triangleright Eq. (6)
 - 8: **for** $j = 1$ to L **do**
 - 9: **Identifying relevant visual tokens (optional):**
 - 10: Generate preliminary token y_j^{init} . \triangleright Eq. (7)
 - 11: Get relevant tokens \mathcal{S}_j with y_j^{init} and q_i^{proj} . \triangleright Eq. (8)
 - 12:
 - 13: **Visual token dropout with uncertainty-guidance:**
 - 14: Get K dropout prob $P^{(k)}$ with $U_{\text{epi}}(i)$. \triangleright Eq. (9)
 - 15: Generate K dropout masks $M^{(k)}$ based on $P^{(k)}$ while retain relevant tokens \mathcal{S}_j .
 - 16: Forward candidates $y_j^{(k)}$ with masks $M^{(k)}$. \triangleright Eq. (10)
 - 17: Majority voting on $y_j^{(k)}$ and get y_j .
 - 18: **end for**
 - 19: **Return** Generated sequence y
-

6. Experiments

We evaluate the proposed DROPOUT DECODING from two aspects: OH reduction and overall generation quality. For OH, we use the CHAIR [41] and THRONE [23] metrics to assess the performance of different decoding methods on the MSCOCO dataset. Additionally, we employ MM-Bench [36] to evaluate the overall generation quality and general ability of these methods.

6.1. Experimental Setup

Base LVLMS. We evaluate all methods on three representative LVLMS: LLaVA-1.5 [32], InstructBLIP [9] and LLaVA-NEXT [34]. LLaVA-1.5 employs linear projection layers to align image and text features, generating 576 visual tokens for detailed visual representation, while LLaVA-NEXT extends this approach by utilizing thousands of visual tokens. In contrast, InstructBLIP uses a Q-former with only 32 visual tokens to bridge the modalities. This diversity highlights the flexibility of our approach, validating its effectiveness across models with both high and low token counts, and confirming its robustness and adaptability.

Hallucination reduction baselines. In addition to the original LVLMS outputs, we compare our method with beam search as well as two state-of-the-art decoding methods: VCD [27], which contrasts original and distorted visuals to reduce hallucinations, and OPERA [19], which applies penalties and token adjustments for better grounding.

Model	Method	CHAIR		THRONE			
		CHAIR _S ↓	CHAIR _I ↓	F _{all} ¹ ↑	F _{all} ^{0.5} ↑	P _{all} ↑	R _{all} ↑
LLaVA-1.5	Greedy	42.20 _{±2.86}	12.83 _{±0.36}	0.795 _{±0.006}	0.784 _{±0.009}	0.772 _{±0.015}	0.847 _{±0.010}
	Beam Search	46.33 _{±1.10}	13.9 _{±0.60}	0.790 _{±0.007}	0.772 _{±0.004}	0.759 _{±0.003}	0.862 _{±0.009}
	OPERA	41.47 _{±0.92}	12.37 _{±0.72}	0.802 _{±0.003}	0.791 _{±0.004}	0.782 _{±0.009}	0.854 _{±0.011}
	VCD	49.20 _{±0.88}	14.87 _{±0.47}	0.786 _{±0.012}	0.771 _{±0.017}	0.759 _{±0.020}	0.854 _{±0.015}
	DROPOUT DECODING	39.80 _{±2.3}	11.73 _{±0.25}	0.804 _{±0.002}	0.796 _{±0.006}	0.790 _{±0.009}	0.851 _{±0.005}
	DROPOUT DECODING (w/o prelim)	39.73 _{±2.15}	12.20 _{±0.70}	0.799 _{±0.002}	0.794 _{±0.004}	0.791 _{±0.007}	0.843 _{±0.005}
InstructBLIP	Greedy	27.87 _{±1.32}	7.90 _{±0.63}	0.809 _{±0.001}	0.826 _{±0.003}	0.832 _{±0.006}	0.803 _{±0.007}
	Beam Search	25.87 _{±2.77}	6.93 _{±0.569}	0.809 _{±0.002}	0.827 _{±0.006}	0.836 _{±0.005}	0.807 _{±0.015}
	OPERA	28.07 _{±1.75}	8.23 _{±0.53}	0.805 _{±0.004}	0.824 _{±0.003}	0.830 _{±0.004}	0.798 _{±0.008}
	VCD	39.33 _{±2.70}	19.10 _{±0.30}	0.737 _{±0.008}	0.746 _{±0.012}	0.751 _{±0.020}	0.757 _{±0.007}
	DROPOUT DECODING	24.53 _{±1.26}	6.63 _{±0.65}	0.814 _{±0.008}	0.833 _{±0.004}	0.838 _{±0.002}	0.808 _{±0.016}
	DROPOUT DECODING (w/o prelim)	26.2 _{±2.40}	7.10 _{±0.854}	0.807 _{±0.008}	0.823 _{±0.006}	0.827 _{±0.010}	0.804 _{±0.010}
LLaVA-NEXT	Greedy	28.80 _{±2.12}	8.10 _{±0.92}	0.815 _{±0.012}	0.832 _{±0.009}	0.830 _{±0.007}	0.799 _{±0.008}
	Beam Search	28.06 _{±1.30}	7.10 _{±0.20}	0.816 _{±0.007}	0.834 _{±0.006}	0.834 _{±0.004}	0.801 _{±0.002}
	OPERA	29.06 _{±1.89}	8.06 _{±1.07}	0.814 _{±0.011}	0.832 _{±0.011}	0.831 _{±0.006}	0.799 _{±0.007}
	VCD	33.19 _{±0.52}	8.10 _{±0.91}	0.818 _{±0.004}	0.822 _{±0.003}	0.808 _{±0.005}	0.822 _{±0.003}
	DROPOUT DECODING	26.26 _{±2.4}	7.39 _{±0.69}	0.821 _{±0.010}	0.840 _{±0.009}	0.842 _{±0.002}	0.805 _{±0.010}
	DROPOUT DECODING (w/o prelim)	27.0 _{±1.80}	7.53 _{±0.643}	0.814 _{±0.009}	0.835 _{±0.007}	0.837 _{±0.003}	0.793 _{±0.008}

Table 1. Comparison of methods on CHAIR_S, CHAIR_I, F_{all}¹, F_{all}^{0.5}, P_{all}, and R_{all} metrics for LLaVA-1.5, InstructBLIP, and LLaVA-NEXT.

6.2. CHAIR

CHAIR [41] is a benchmark designed to evaluate OH in image captioning situations. CHAIR provides two primary metrics to measure hallucination at different granularities: sentence-level and object-level. The sentence-level metric, CHAIR_S, calculates the proportion of captions that contain any hallucinated objects, giving an overall measure of hallucination frequency in captions. And the object-level metric, CHAIR_I, calculates the proportion of hallucinated objects out of all mentioned objects across captions, reflecting the prevalence of hallucination among the objects described.

Results. As shown in Table 1, DROPOUT DECODING consistently outperforms baseline approaches across various models, demonstrating its reliability and effectiveness in image captioning. Especially, on InstructBLIP, CHAIR_I is improved by approximately 16% over the second-best method, and CHAIR_S sees a gain of around 12%. These substantial improvements underscore the effectiveness of our approach, which aligns well with intuitive expectations that token dropout will reduce generated objects. Furthermore, DROPOUT DECODING reduces the generation of hallucinated objects without compromising the inclusion of relevant objects. This reduction in hallucinated content, as opposed to accurate content, is further validated by the recall metric (R_{all}) in THRONE.

6.3. THRONE

THRONE [23] assesses hallucinations in LVLM-generated responses, covering both “Type I” (mentions of non-existent objects, like CHAIR) and “Type II” (accuracy of object ex-

istence, like POPE [30]). It uses P_{all} (Precision), R_{all} (Recall), F_{all}¹, and F_{all}^{0.5}. Additionally, it employs F_β, which combines P_{all} and R_{all}, with the parameter β controlling the weight of R_{all} relative to P_{all}: $F_{all}^{\beta} = (1 + \beta^2) \cdot \frac{P_{all} \times R_{all}}{(\beta^2 \times P_{all}) + R_{all}}$.

Results. The test results in Table 1 illustrate that DROPOUT DECODING surpasses nearly all baseline methods across various metrics, highlighting its effectiveness in reducing both Type I and Type II hallucinations. Specifically, DROPOUT DECODING demonstrates notable strengths in InstructBLIP, excelling in the P_{all} metric and achieving the highest performance in R_{all}. Across models, P_{all} metric achieves larger improvement while the R_{all} score also exceeds that of the Greedy method, confirming that retaining overlap tokens effectively preserves relevant objects. The large increase in F_{all}^{0.5} further shows its comprehensiveness.

6.4. MMBench

MMBench [36] is a comprehensive benchmark designed to evaluate the multimodal capabilities of LVLMs across various tasks and data types, including image captioning, question answering, and object recognition. It provides a holistic view of a model’s strengths and weaknesses in multimodal understanding. Since the prompt length limits in MMBench exceed InstructBLIP’s token allowance, we report results only on LLaVA-1.5 and LLaVA-NEXT.

Results. As shown in Table 2, DROPOUT DECODING outperforms all the other baselines on LLaVA-1.5, which demonstrates not only its effectiveness in hallucination mitigation but also its robustness and adaptability across a broader range of multimodal tasks.

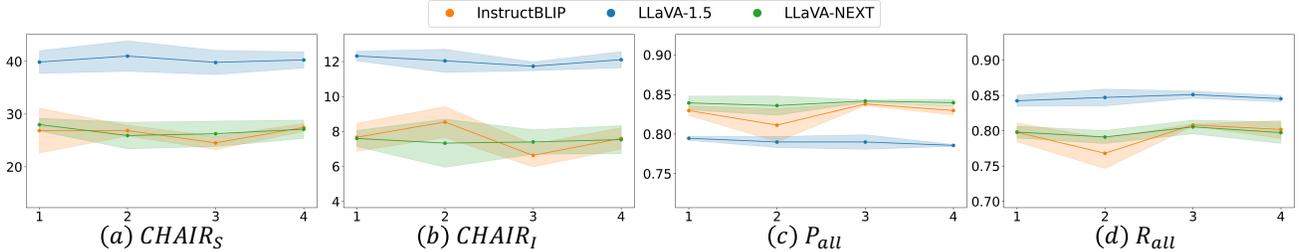


Figure 3. Comparison of $CHAIR_S$, $CHAIR_I$, P_{all} and R_{all} scores with standard deviations across different candidate numbers.

Method	Original	VCD	OPERA	DROPOUT	DECODING
LLaVA-1.5	71.86	72.35	73.86		74.01
LLaVA-NEXT	74.57	69.65	74.54		74.31

Table 2. Results of different methods on MMBench.

7. Analysis and Ablation Studies

7.1. Number of Parallel Dropouts

As described in §5.2, we generate K candidate predictions by applying token dropout with different dropout masks $M^{(k)}$. In this section, we investigate how varying K from 1 to 4 affects generation quality.

We fix $\delta^{(k)} = 0.1$ and adjust $\gamma^{(k)}$ based on a predefined order: $\gamma^{(1)} = 0.3$, $\gamma^{(2)} = 0.5$, and $\gamma^{(3)} = 0.7$. However, setting $\gamma^{(k)}$ to 0.9 leads to excessive dropout of visual tokens and degrades InstructBLIP’s performance, so we set $\gamma^{(4)} = 0.1$. Moreover, our majority voting algorithm favors candidates with fewer dropped tokens in the event of a tie, meaning that when comparing only two candidates, both will yield identical outputs. To address this, we remove Candidate 1 in the second round, leaving only Candidate 2.

As shown in Fig. 3 (a) and (b), both $CHAIR_S$ and $CHAIR_I$ scores peak at $K = 3$ for LLaVA-1.5 and InstructBLIP. Increasing K to 4 introduces a less-masked candidate that slightly negatively impact our method’s effectiveness in reducing hallucinations. Conversely, using fewer candidates (*e.g.*, only Candidate 1 and Candidate 2) lacks the balance needed for stable voting outcomes, resulting in increased randomness. Similarly, Fig. 3 (c) and (d) shows that THRONE’s R_{all} and P_{all} metrics also perform best at $K = 3$. Overall, we find that selecting three candidates strikes the optimal balance between increased certainty from additional votes and the controlled uncertainty introduced by candidate dropout probability, allowing DROPOUT DECODING to achieve more trustworthy and stable generation results.

7.2. Initial Identification of Relevant Visual Tokens

As discussed in §5.2, DROPOUT DECODING may employ a preliminary forward pass to retain most relevant objects during generation, which helps reduce hallucinated objects

while maintaining high-quality outputs. In contrast, bypassing this step risks masking relevant visual tokens during the token dropout phase, potentially degrading overall performance. However, incorporating a preliminary forward pass roughly doubles the computational cost per generation. To strike a balance between accuracy and efficiency, we evaluate the approach both with and without this step. Specifically, our goals are: 1) to confirm the effectiveness of the preliminary forward pass, and 2) to explore a more efficient alternative when computational resources are limited.

As shown in Table 1, including the preliminary forward pass consistently improves most metrics, with particular notable gains in the F_{all} score on THRONE. We also observe further improvements in CHAIR metrics, especially when using InstructBLIP. Interestingly, for LLaVA-1.5, the variant without the preliminary pass performs slightly better on CHAIR, though the THRONE performance remains stable (*i.e.*, the version with the preliminary pass still performs better). We hypothesize that this discrepancy arises from differences in the abundance of visual tokens. LLaVA-1.5 uses 576 visual tokens, each carrying relatively less weight, whereas InstructBLIP relies on just 32, making each token’s contribution more critical. Consequently, omitting the preliminary forward pass in InstructBLIP risks losing critical information, lowering performance. These findings suggest that while a preliminary forward pass is highly beneficial for models with abundant visual tokens, models with fewer tokens may achieve better computational efficiency and performance by skipping this additional step.

8. Conclusion

We introduce DROPOUT DECODING, a novel uncertainty-guided context selective decoding approach aimed at enhancing the reliability of LVLMs. After quantifying the uncertainty in visual inputs, DROPOUT DECODING accordingly drops out visual tokens to regularize effect of information uncertainty, and employs an ensemble-based decoding approach to stabilize generation. Extensive experiments on benchmarks including CHAIR, THRONE, and MMBench validate the effectiveness, demonstrating consistent performance improvements over existing methods in both hallucination reduction and general multimodal capability tasks.

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1
- [2] Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. *arXiv preprint arXiv:2407.18698*, 2024. 2
- [3] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation, 2018. 3
- [4] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*. 1, 2, 6
- [5] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024. 1
- [6] Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprpm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1346–1362, 2024. 1
- [7] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 1
- [8] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7654–7664, Dublin, Ireland, 2022. Association for Computational Linguistics. 2
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 6
- [10] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding, 2024. 2
- [11] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 1
- [12] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. 6
- [13] Yarín Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059, New York, New York, USA, 2016. PMLR. 1, 3
- [14] Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 6
- [15] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024. 1
- [16] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023. 1
- [17] Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. Cliptrans: transferring visual knowledge with pre-trained models for multimodal machine translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2875–2886, 2023. 1
- [18] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. 2
- [19] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation, 2024. 2, 6
- [20] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021. 2, 3, 4
- [21] Lai Jiang, Hongqiu Wu, Hai Zhao, and Min Zhang. Chinese spelling corrector is just a language learner. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6933–6943, 2024. 1
- [22] Ehsan Kamaloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*, 2023. 1
- [23] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models, 2024. 2, 5, 6, 7
- [24] Alex Kendall and Yarín Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 1
- [25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 6
- [26] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation, 2023. 2

- [27] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. 2, 6
- [28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2
- [29] Wanhua Li, Zibin Meng, Jiawei Zhou, Donglai Wei, Chuang Gan, and Hanspeter Pfister. Socialgpt: Prompting llms for social relation reasoning via greedy segment optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1
- [30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 2, 7
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 6
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 1, 2
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 6
- [35] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1
- [36] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 2, 6, 7
- [37] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1
- [38] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn, 2016. 3
- [39] Zhenting Qi, Hongyin Luo, Xuliang Huang, Zhuokai Zhao, Yibo Jiang, Xiangjun Fan, Himabindu Lakkaraju, and James Glass. Quantifying generalization complexity for large language models. *arXiv preprint arXiv:2410.01769*, 2024. 1
- [40] Denisa Roberts and Lucas Roberts. Smart vision-language reasoners. *arXiv preprint arXiv:2407.04212*, 2024. 1
- [41] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. 2, 5, 6, 7
- [42] Lior Rokach. Ensemble-based classifiers. *Artificial intelligence review*, 33:1–39, 2010. 6
- [43] Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. Introducing an improved information-theoretic measure of predictive uncertainty. *arXiv preprint arXiv:2311.08309*, 2023. 2, 3, 4
- [44] Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of uncertainty with adversarial models. *Advances in Neural Information Processing Systems*, 36:19446–19484, 2023. 3
- [45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1
- [46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020. 1
- [47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 1
- [48] Chaoqi Wang, Zhuokai Zhao, Chen Zhu, Karthik Abinav Sankararaman, Michal Valko, Xuefei Cao, Zhaorun Chen, Madian Khabza, Yuxin Chen, Hao Ma, et al. Preference optimization with multi-sample comparisons. *arXiv preprint arXiv:2410.12138*, 2024. 2
- [49] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 1
- [50] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020. 2
- [51] Kelvin Xu. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 1
- [52] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016. 1
- [53] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2024. 2
- [54] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1
- [55] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1
- [56] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. 1

- [57] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pages arXiv–2310, 2023. 1
- [58] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 1
- [59] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [60] Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: Self logits evolution decoding for improving factuality in large language models. *arXiv preprint arXiv:2411.02433*, 2024. 1
- [61] Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. Beyond training: Dynamic token merging for zero-shot video understanding. *arXiv preprint arXiv:2411.14401*, 2024. 1
- [62] Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining. *arXiv preprint arXiv:2404.09387*, 2024. 1
- [63] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1
- [64] Zhuokai Zhao. Enhanced data utilization for efficient and trustworthy deep learning. 2024. 2
- [65] Zhuokai Zhao, Yibo Jiang, and Yuxin Chen. Direct acquisition optimization for low-budget active learning. *arXiv preprint arXiv:2402.06045*, 2024. 3
- [66] Zhuokai Zhao, Harish Palani, Tianyi Liu, Lena Evans, and Ruth Toner. Multimodal guidance network for missing-modality inference in content moderation. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–4. IEEE, 2024. 1

A. Details of Uncertainty Decomposition

A detailed derivation of Eq. (5):

$$\begin{aligned}
 U_{\text{total}} &= \mathbb{H} [q^{\text{proj}}] \\
 &= - \sum_{y \in \mathcal{V}} q^{\text{proj}}(y) \log q^{\text{proj}}(y) \\
 &= - \sum_{y \in \mathcal{V}} \left(\mathbb{E}_i \left[q_i^{\text{proj}}(y) \right] \right) \log q^{\text{proj}}(y) \\
 &= \mathbb{E}_i \left[- \sum_{y \in \mathcal{V}} q_i^{\text{proj}}(y) \log q^{\text{proj}}(y) \right] \\
 &= \mathbb{E}_i \left[\text{CE} \left(q_i^{\text{proj}}, q^{\text{proj}} \right) \right] \\
 &= \mathbb{E}_i \left[\mathbb{H} \left[q_i^{\text{proj}} \right] + D_{\text{KL}} \left(q_i^{\text{proj}} \parallel q^{\text{proj}} \right) \right] \\
 &= \mathbb{E}_i [U_{\text{ale}}(i) + U_{\text{epi}}(i)]
 \end{aligned} \tag{11}$$

B. Implementation Details

The experimental setup of DROPOUT DECODING is shown in Table 3. We set the maximum new tokens to 512 to ensure the complete generation of models, therefore achieving more reliable results from CHAIR and THRONE. In MMBench, as all questions are single-choice questions, we set the maximum new tokens to 1 for a more precise evaluation. We set other parameters in generation to greedy for more stable and repeatable results.

Parameters	CHAIR	THRONE	MMBench
	512	512	1
Top-k		False	
Top-p		1	
Temperature τ		1	
Number Beams		1	

Table 3. Parameter settings used in our experiments.

In addition to general generation settings, DROPOUT DECODING includes hyperparameters specified in §5.2. The details of these hyperparameter settings are provided below:

Top- k in identifying relevant visual tokens. Before the decoding process, we first obtain q^{proj} , which is then used in the decoding process for generating the relevant visual tokens. The higher the top- k is, the more visual tokens are expected to be kept during the decoding process. In LLaVA-1.5, we set $k = 5$, and in InstructBLIP, we set $k = 10$. The difference of k between LLaVA-1.5 and InstructBLIP derives from the informative level of each visual token, where in LLaVA-1.5, each visual token carries less information than in InstructBLIP, which only contains 32 visual tokens.

Number of mask K . K refers to the number of predictions that will join the majority vote progress. We set $K = 3$ in our experiment settings.

$\gamma^{(k)}$ and $\delta^{(k)}$ in uncertainty-guided masking We set $\delta^{(k)} = 0.1, \gamma^{(k)} = 0.2 * k + 1; k = 1, 2, \dots, K; K = 3$ in our experiment settings.

Moreover, we provide the hyperparameter settings of our baselines. OPERA’s hyperparameters can be referred to Table 4; VCD’s hyperparameters can be referred to Table 5.

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

Table 4. OPERA hyperparameter settings.

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

Table 5. VCD hyperparameter settings.

C. Further Discussion on Ablation Studies

To further validate our uncertainty guidance’s effectiveness, we select random masking strategy as an additional baseline to compare with DROPOUT DECODING’s uncertainty-guided masking. The experimental setup remains identical, except that tokens are masked randomly, that is, candidate k masks each vision token at $\gamma^{(k)}$ instead of using uncertainty guidance. The generated results using the random masking strategy often suffer from issues, with models producing repeated tokens until reaching the maximum token limit. For instance, the model might repeatedly generate “skiers” hundreds of times (generation: “The image shows shows a a snowy snowy slope with a skiers skiers skiers skiers ...”); this occurred in approximately 20–25 out of 500 cases, an issue nearly never encountered with our method. This behavior likely stems from random masking disrupting essential context information within LVLMs. In contrast, our uncertainty-guided masking applies a lower masking rate to tokens that the LVLMs are less “surprised” by and a higher rate to tokens that elicit greater surprise. This allows the model to generate content in a rather “expected” manner, even though many informative vision tokens are masked. By preserving base context information, our approach effectively maintains the LVLMs’ consistency and coherence.

D. Generation Examples and Analysis

D.1. LLaVA-1.5 Examples



Generated Caption:

Greedy Generated caption:

The image features a black and white photograph of a small boat with three people rowing it down a river. The boat is being pulled by the people, **and there are two swans swimming alongside the boat.** The swans are swimming in the water near the boat, creating a peaceful and serene scene. The people in the boat are enjoying the ride, possibly on a leisurely outing.

DROPOUT DECODING Generated caption:

The image features a group of people in a small boat, paddling down a river. There are at least four people visible in the boat, with some of them sitting closer to the front and others towards the back. The boat is filled with people, and they seem to be enjoying their time together on the water.



Generated Caption:

Greedy Generated caption:

The image features a kitchen countertop with a microwave oven and a large, wrapped object sitting on top of it. The object is wrapped in aluminum foil, which is covering a large rectangular box or container. The microwave oven is positioned on the left side of the countertop, and the wrapped object is placed on the right side. **The countertop is also adorned with a few utensils, such as a spoon and a fork.**

DROPOUT DECODING Generated caption:

The image features a kitchen counter with a metal pan or tray sitting on top of it. The pan is covered in aluminum foil, which is commonly used to keep food warm or to prevent it from sticking to the pan. The kitchen counter appears to be a part of a larger countertop, and the scene suggests that the pan is being prepared for cooking or baking.



Generated Caption:

Greedy Generated caption:

The image features a dining table set with a white plate containing a variety of food items, including meat, vegetables, and bread. The plate is placed on a wooden table, and there is a glass of wine next to it. The table is set for a meal, with a fork and **knife placed on the table.** The food items on the plate are arranged in a visually appealing way, making the dish look appetizing.

DROPOUT DECODING Generated caption:

The image features a dining table with a white plate containing a delicious meal. The plate is topped with a variety of food items, including a piece of meat, a roll, and a side of vegetables. There are also a couple of carrots on the plate. In addition to the plate, there is a wine glass placed on the table, suggesting that the meal is accompanied by a glass of wine. A chair is positioned near the table, ready for someone to sit and enjoy the meal.

D.2. InstructBLIP Examples



Generated Caption:

Greedy Generated caption:

The image features a table with a laptop and a plate of food, including a bowl of salad and a plate of food. The food is arranged on the table, **with a fork** and a knife nearby. The laptop is placed on the table, likely used for work or entertainment. The table is set up for a comfortable dining experience, with a variety of food options available.

DROPOUT DECODING Generated caption:

The image features a plate of food on the table is aesthetically pleasing and well-balanced. There is a plate of food, including a salad, and a bowl of chicken, and a fork. The plate is placed on the table, which is also adorned with a laptop and a computer. The laptop is placed on the table, and the computer is positioned to the left of the plate.



Generated Caption:

Greedy Generated caption:

The image features a red fire hydrant placed in front of a stone wall. The fire hydrant is positioned in the middle of the scene, with a window above it. **There are two benches placed near the fire hydrant, one on the left and another on the right side of the image. The benches are positioned close to the fire hydrant, with one closer to the window and the other further away.**

DROPOUT DECODING Generated caption:

The image features a red fire hydrant placed in front of a stone building, with a window above it. The window is located on the second floor of the building, and the fire hydrant is positioned in front of the window. There are two metal bars on either side of the fire hydrant, which are likely used to secure the hydrant in place. The overall scene is quite unique, with the fire hydrant and the window creating a distinctive composition.



Generated Caption:

Greedy Generated caption:

The image features a dining table with a white plate topped with a slice of pie, a spoon, and a cherry on top. The pie is surrounded by various utensils, including a **fork, knife,** and spoon. There are also two cups on the table, one on the left and one on the right, each containing a beverage. The dining table is set with a variety of dishes, including a bowl, a cup, and a spoon. The scene is set for a delicious meal, with the pie and beverages adding to the overall ambiance.

DROPOUT DECODING Generated caption:

The image features a dining table is set with a variety of dishes, including a plate with a slice of pie, a cherry, and a spoon. There are also several cups of coffee and a bowl of soup. The table is set for a meal, and the dishes are arranged in a way that suggests a person is about to enjoy their food.

D.3. LLaVA-NEXT Examples



Generated Caption:

Greedy Generated caption:

In the heart of a bustling pizza shop, a woman stands, her attention captivated by the slice of pizza she holds in her hand. The pizza, with its golden crust and vibrant toppings, is a tantalizing sight. She's dressed casually, her green jacket contrasting with the warm tones of the shop. The shop itself is a feast for the eyes. **A large refrigerator stands in the background, its surface adorned with various signs and posters.** The words "We guarantee the best pizza" are prominently displayed, a testament to the shop's commitment to quality. ... The scene is a snapshot of everyday life, capturing the simple joy of enjoying a delicious slice of pizza.

DROPOUT DECODING Generated caption:

In the heart of a bustling pizza shop, a woman stands, her attention captivated by the slice of pizza she holds in her hand. The pizza, with its golden crust and vibrant toppings, is a tantalizing sight. She's dressed casually, her green jacket contrasting with the warm tones of the shop. The shop itself is a feast for the eyes. Boxes of pizza are neatly stacked on shelves, their white and red colors standing out against the darker tones of the shop. A sign on the wall proudly declares "We guarantee the best pizza in town", a testament to the shop's confidence in its offerings. ... Every object in this image tells a story - of delicious food, of a lively shop, and of a woman about to enjoy a slice of pizza.

Generated Caption:

Greedy Generated caption:



The image depicts a cozy indoor setting with a focus on a wooden table. The table is set with a few items, including a plate of food, a glass of water, and a small bowl. There are also a couple of utensils, a fork and **a knife**, placed on the table. The table is positioned in front of a large window that offers a view of a patio area with a bamboo fence and a small table. The room has a warm and inviting atmosphere, with a wooden floor and a beige cushion visible in the foreground. The lighting suggests it might be daytime. There are no visible texts or distinctive brands in the image. The style of the image is a candid, real-life photograph with a focus on the table setting and the view outside the window.

DROPOUT DECODING Generated caption:

The image depicts a cozy indoor setting with a focus on a wooden table. The table is set with a few items, including a plate with a small portion of food, a glass of water, and a pair of chopsticks. There are also a couple of small bowls and a spoon on the table. The table is positioned in front of a large window that offers a view of a wooden deck outside. The deck is furnished with a bench and a small table, and there are bamboo plants and a tree visible in the background. The room has a warm and inviting atmosphere, with a wooden floor and a cushion on the floor to the left of the table. The lighting suggests it might be daytime, and the overall style of the image is natural and serene.

D.4. Analysis

In our examples, we use red to denote hallucinations. Specifically, DROPOUT DECODING addresses hallucinations in two key ways: eliminating non-existent objects and correcting the generation of misidentified objects. For instance, in the third example of LLaVA-1.5, the Greedy method hallucinates a non-existent object, a knife, whereas DROPOUT DECODING effectively removes this hallucination. Similarly, in the second example of InstructBLIP, the Greedy method misidentifies metal bars as benches, but DROPOUT DECODING successfully corrects this, accurately recognizing the metal bars.