Mitigating Hallucination of Large Vision-Language Models via Dynamic Logits Calibration

Jiahe Chen[1], Jiaying He[2], Qian Shao[1], Qiyuan Chen[1], Jiahe Ying[3], Hongxia Xu[1], Jintai Chen[4], Jianwei Zheng[2]*, Jian Wu[1]*
[1]Zhejiang University [2]Zhejiang University of Technology [3]Fudan University
[4]Hong Kong University of Science and Technology (Guangzhou)

Abstract

Large Vision-Language Models (LVLMs) have demonstrated significant advancements in multimodal understanding, yet they are frequently hampered by hallucination-the generation of text that contradicts visual input. Existing training-free decoding strategies exhibit critical limitations, including the use of static constraints that do not adapt to semantic drift during generation, inefficiency stemming from the need for multiple forward passes, and degradation of detail due to overly rigid intervention rules. To overcome these challenges, this paper introduces **D**ynamic **L**ogits **C**alibration (**DLC**), a novel training-free decoding framework designed to dynamically align text generation with visual evidence at inference time. At the decoding phase, DLC step-wise employs CLIP to assess the semantic alignment between the input image and the generated text sequence. Then, the Relative Visual Advantage (RVA) of candidate tokens is evaluated against a dynamically updated contextual baseline, adaptively adjusting output logits to favor tokens that are visually grounded. Furthermore, an adaptive weighting mechanism, informed by a real-time context alignment score, carefully balances the visual guidance while ensuring the overall quality of the textual output. Extensive experiments conducted across diverse benchmarks and various LVLM architectures (such as LLaVA, InstructBLIP, and MiniGPT-4) demonstrate that DLC significantly reduces hallucinations, outperforming current methods while maintaining high inference efficiency by avoiding multiple forward passes. Overall, we present an effective and efficient decoding-time solution to mitigate hallucinations, thereby enhancing the reliability of LVLMs for more practices. Code will be released on Github.

1 Introduction

The field of multimodal artificial intelligence has witnessed remarkable progress through the development of Large Vision-Language Models (LVLMs) [7, 12, 23, 26, 30, 42, 47, 57]. By integrating visual encoders [35, 52] with large language models (LLMs) [2, 10, 40, 41], contemporary LVLMs have successfully unified visual and textual information within a common semantic framework. These models demonstrate impressive capabilities across diverse visual tasks [3, 4, 21, 31], ranging from straightforward Visual Question Answering (VQA) to sophisticated scene reasoning.

Despite their remarkable advances in multimodal comprehension, LVLMs remain plagued by a critical limitation: hallucination, the generation of text contradicting visual input [4, 5, 17, 22, 58]. This phenomenon introduces deceptive inconsistencies, as the model produces plausible but factually unfaithful descriptions of the observed scene. This issue is especially critical in high-stakes domains like medical diagnosis [6, 24] and autonomous systems [8, 34], where such inconsistencies may incur serious real-world consequence, further hindering the practical deployment of LVLMs.

*Corresponding authors.

Preprint. Under review.

To address the hallucination issue, numerous training-free decoding strategies have emerged. These approaches have garnered significant attention due to their independence from high-quality annotated datasets and resource-intensive computational requirements. Based on their intervention mechanisms, existing methods can be systematically categorized into three main paradigms: 1) Contrastive Decoding [18, 22, 43]; 2) Guided Decoding Frameworks [9, 13, 54]; and 3) Latent Space Intervention Methods [17, 32, 39, 58]. While these strategies demonstrate some effectiveness in generation tasks, systematic analysis reveals three fundamental limitations. **First**, static constraint mechanisms fail to dynamically adapt to semantic drift during generation, leaving the model vulnerable to hallucinations in later stages as linguistic priors dominate over visual grounding. **Second**, most methods require multiple forward passes or regeneration, drastically reducing inference efficiency. **Third**, rigid intervention rules lack contextual sensitivity, either enforcing overly strict visual alignment or allowing unchecked deviations, thereby distorting the balance between fidelity and creativity.

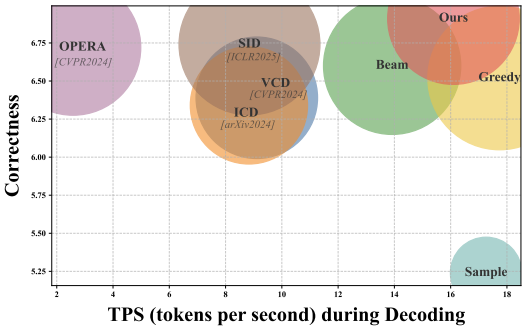

Figure 1: **Comparison of decoding methods across three dimensions**: Speed, Correctness, and Detailedness using LLaVA-1.5 [30] (max new tokens: 64), with all experiments conducted on a single NVIDIA A100 GPU. The vertical axis represents correctness, while the horizontal axis shows tokens per second (TPS). Bubble size corresponds to Detailedness of responses. Our method achieves superior accuracy and detailedness while maintaining decoding efficiency.

Addressing these limitations-**namely the lack of dynamic adaptation to context, the high inference cost, and the distortion of reality due to rigid rules**-demands a deeper investigation. Our work identifies that a core issue underlying hallucination is the increasing reliance on linguistic priors during decoding, causing progressive detachment from visual evidence despite initial alignment. Motivated by this observation and the need to overcome the shortcomings of existing methods, we introduce **D**ynamic **L**ogits **C**alibration (**DLC**) as an efficient and training-free decoding strategy. DLC employs the CLIP model as a dynamic probe to continuously monitor semantic correspondence between the image and generated text during token generation. It evaluates the semantic alignment of both the current text prefix and potential next tokens relative to the image, subsequently adjusting the logits to favor visually grounded options. This mechanism empowers tokens possessing a high "relative visual advantage" to supersede strong linguistic priors, thus diminishing the likelihood of visual hallucination while preserving generation coherence and maintaining high inference speed. As initially evidenced in Figure 1, DLC exhibits superior efficiency compared to other leading approaches. The concerned efficacy is validated through extensive experiments across multiple benchmarks, demonstrating its successful integration into diverse LVLM architectures such as LLaVA [30], InstructBLIP [12], and MiniGPT-4 [57]. In summary, our key contributions are threefold:

- We revisit the problem of semantic priors in LVLM decoding, highlighting the resulting detachment from visual evidence and onset of hallucination during later generation stages, and revealing a sharp drop in visual alignment for the high-probability tokens selected by the model during hallucination.
- To address the problem, we craft DLC, a novel training-free scheme, which dynamically adjusts token logits by leveraging CLIP to assess vision-language alignment in real-time, thereby mitigating hallucinations based on contextual visual feedback. The integration of semantic alignment into the logits space offers a new avenue for tackling dynamic vision-language alignment challenges.
- Comprehensive evaluations on diverse benchmarks confirm the efficacy of DLC, showing significant improvements in reducing hallucinations with appealing running efficiency.

2 Related Work

Hallucination issue in LVLMs. The root cause of LVLM hallucination lies in the inherent tension between parametric knowledge dominance and visual evidence grounding: while users expect models to generate responses strictly conditioned on visual inputs, LVLMs frequently default to priors learned from their textual training corpus, favoring statistically likely but visually ungrounded

outputs [15, 16, 22]. This over-reliance on linguistic patterns on top of pixel-accurate reasoning leads to hallucinations when the model's generative biases override fidelity to the input image (see Appendix A.1 for additional background on LVLMs).

The prior approaches broadly fall into three paradigms. **1) Data-driven optimization** [5, 19, 29, 45, 49–51, 55], including fine-grained contrastive learning, reinforcement learning, and curated instruction tuning with noisy or adversarial datasets. While striving for alignment enhancement, these methods demand expensively annotated data and often risk overfitting issue. **2) Post-hoc correction mechanisms** [37, 44, 48, 53, 56], such as auxiliary reviser networks, image regeneration feedback, or external knowledge grounding. Though effective, they introduce significant inference latency and reliance on auxiliaries. **3) Decoding strategy refinement** [9, 13, 17, 18, 22, 32, 39, 43, 54, 58], exemplified by VCD [22] that attenuates textual bias by contrasting original and distorted visual inputs, OPERA [17] that utilizes over-trust penalty and a retrospection allocation to alleviate aggregation patterns, and DoLa [11] that mitigates factual hallucinations via layer-wise logit calibration. Despite data dependency being avoided, these training-free methods exhibit critical limitations as discussed in the Introduction. DLC addresses these issues by employing real-time visual alignment assessment (Section 4.1) and adaptive logit modulation (Section 4.2), providing a more faithful decoding stage.

3 Preliminary and Motivation

3.1 Generative Process of LVLMs

An LVLM, parameterized by θ, processes jointly visual images v and textual queries x. Typically, a pre-trained network encodes v into visual features. An alignment module (*e.g.*, Q-Former [26] or linear projection [30]) then maps these features into the textual semantic space. These aligned visual representations are concatenated with the embedded text query x to form a joint input representation. Conditioned on this unified input, the LVLM auto-regressively predicts the next token y_t:

$$y_t \sim p_\theta\left(y_t \mid v, x, y_{<t}\right) \propto \exp\left(logit_\theta\left(y_t \mid v, x, y_{<t}\right)\right) \tag{1}$$

where $y_{<t}$ are previous tokens and $logit_\theta$ is the logits function. Decoding strategies then select the next token, which is appended to the input for subsequent iterations until generation is complete.

Our proposed DLC enhances this generative process as a training-free decoding strategy. Notably, it is **model-agnostic**, requiring no architectural changes or fine-tuning, allowing flexible application across diverse LVLMs. Furthermore, DLC is **compatible** with mainstream decoding strategies like nucleus and temperature sampling, seamlessly integrating by modulating logits before token selection. Critically, this approach maintains **high inference efficiency**, introducing only minimal computational overhead and latency, making it a practical enhancement for visually grounded generation.

3.2 CLIP Model and CLIP Score

The Contrastive Language-Image Pre-training (CLIP) model [35], trained on extensive datasets of image-text pairs using a contrastive learning objective, learns to embed visual and textual information into a shared, high-dimensional semantic space. Formally, for a given image v and text sequence s, we calculate their semantic similarity using cosine similarity, denoted as $\mathrm{CLIP}(v, s)$:

$$\mathrm{CLIP}(v, s) = \mathrm{Cosine\_Similarity}(E_I(v), E_T(s)) \tag{2}$$

where $E_I(v)$ and $E_T(s)$ are embeddings from CLIP's pre-trained image (E_I) and text (E_T) encoders.

A higher $\mathrm{CLIP}(v, s)$ value signifies greater semantic relevance. Section 4 details how our method leverages this score to guide text generation by assessing visual consistency during the decoding.

3.3 Motivation

To understand the origins of hallucinations in LVLMs and identify real-time signals for their mitigation, we empiri-

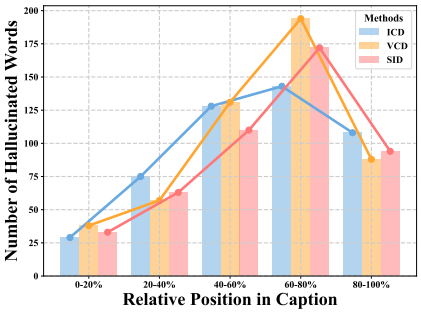

Figure 2: **Distribution of hallucinatory words by relative position** in generated captions, analyzed using the CHAIR benchmark, comparing various decoding methods (ICD, VCD, SID). All methods use sampling decoding strategy, with a Max Token setting of 512. Hallucinations predominantly emerge in the later stages (60%-80%) of generation, indicating progressive semantic drift.

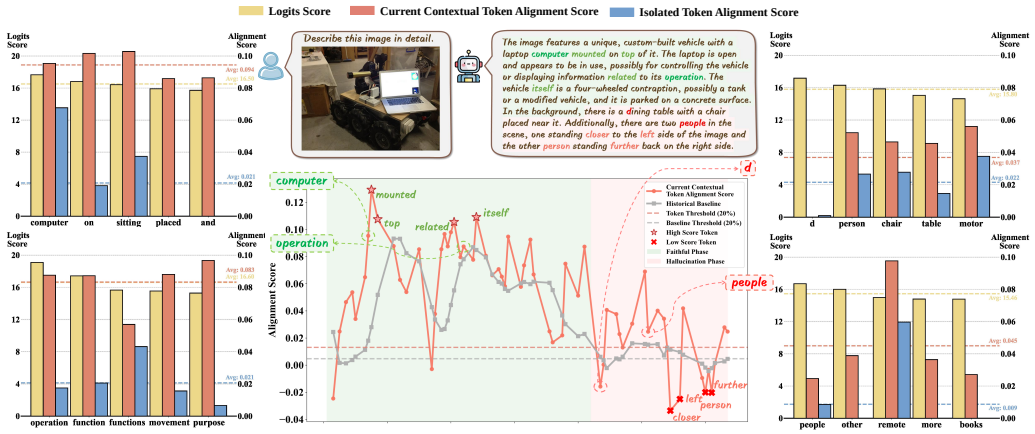

Figure 3: **Visualization of semantic drift and token selection in LVLMs.** The central plot tracks Current Contextual Token Alignment (CCTA) against a Historical Baseline $\bar{B}_t$, contrasting visually faithful (**green**) and hallucination (**red**) phases. Surrounding bar charts show Language Model Logits, Isolated Token Alignment (ITA), and potential Current Contextual Token Alignment (CCTA) if selected for candidate tokens at chosen steps, illustrating the divergence between high logits and low visual alignment during hallucination.

cally analyze the dynamic evolution of the autoregressive generation. As illustrated in Figure 2, a common empirical observation across state-of-the-art decoding methods is that while models often successfully anchor their initial generation in visual evidence, hallucinations tend to emerge progressively, exhibiting a pronounced concentration in the later stages of sequence generation (typically peaking within the 60%-80% range). This pattern of semantic drift-wherein the model increasingly relies on learned linguistic priors and patterns over continuous visual grounding-highlights a tendency to gradually deviate from the visual input as the sequence unfolds. Consequently, this necessitates a closer examination of how visual consistency evolves at the token level during decoding, which is crucial for identifying real-time signals to monitor fidelity and counteract drift.

Building upon this, to dissect the dynamics of semantic drift at a granular level, we leverage the CLIP score (Eq. 2) as a proxy for the semantic alignment between the input image v and the evolving text $y_{<t}$ during decoding. Our subsequent analysis, exemplified by the case study in Figure 3, focuses on the trajectory of visual consistency scores. This approach allows us to contrast the dynamics observed during phases of visually faithful description versus those containing hallucinations.

- **Analyzing the Trajectory of Visual Consistency:** The central plot in Figure 3 illustrates the evolving visual alignment during text generation. The **Current Contextual Token Alignment (CCTA)** trajectory quantifies the visual alignment at each step t. This score is derived from the CLIP similarity between the input image v and the sequence segment formed by the most recently chosen token y_t appended to its immediate preceding context $y_{<t}$ (detailed in Section 4). It effectively measures how well the current state of generation, including the latest token, aligns with the visual input. Complementing this, the **Historical Baseline** $\bar{B}_t$ provides a smoothed average of these CCTA scores from the recent generation history (detailed in Section 4). This baseline acts as a dynamic reference, reflecting the level of visual coherence that the model has recently established and against which the alignment of new tokens can be compared. During the visually-grounded stages of generation, the CCTA trajectory typically remains high, often rising with or closely tracking the $\bar{B}_t$, indicating successful selection of tokens that maintain or enhance semantic coherence. However, as generation progresses into hallucination phases, a characteristic downward trend emerges: the CCTA often drops significantly and exhibits increased volatility, frequently falling below the $\bar{B}_t$ and other indicative thresholds (*e.g.*, Baseline Threshold). This decline signals a weakening connection to visual evidence and the onset of semantic drift.
- **Examining Candidate Token Selection:** The bar charts surrounding the central graph in Figure 3 provide snapshots at different generation stages, illustrating the relationship between the language model's preference for a token (Logits Score) and that token's visual consistency. In visually faithful stages, alignment is typical: tokens with high logits scores also tend to exhibit strong visual consistency. Conversely, during hallucinatory stages, a critical divergence is often observed. The

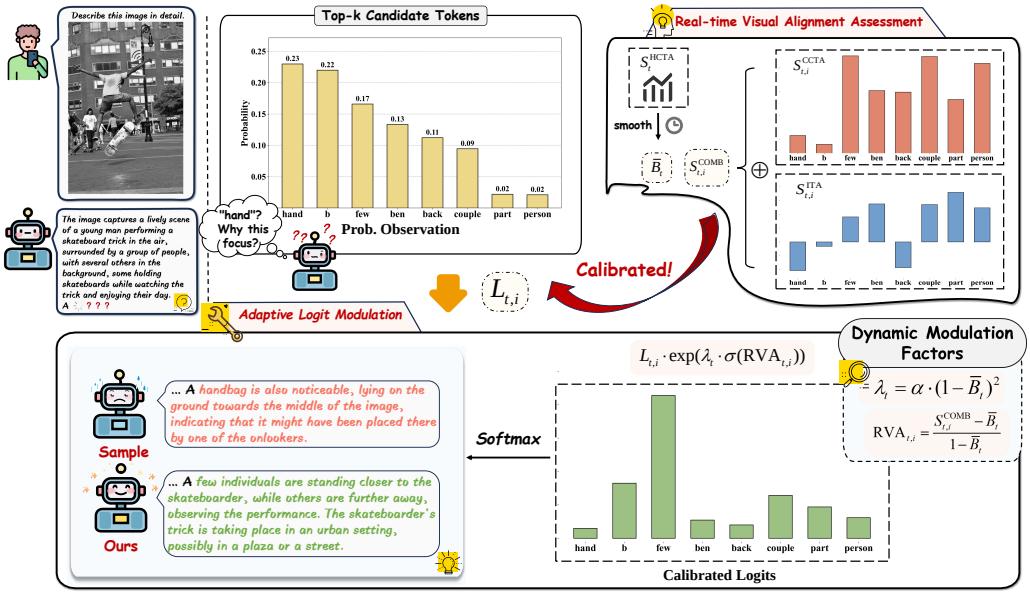

Figure 4: **Overview of Dynamic Logits Calibration (DLC).** Given an input image and prompt, DLC first performs Real-time Visual Alignment Assessment on top-k candidate tokens by calculating Current Contextual Token Alignment (CCTA) and Isolated Token Alignment (ITA) scores relative to a historical baseline ($\bar{B}_t$). These scores inform the Adaptive Logit Modulation step, which computes Relative Visual Advantage (RVA) and a dynamic modulation factor (λ_t) to adjust the original logits ($L_{t,i}$), favoring visually grounded tokens.

model may assign high logits scores to, and consequently select, tokens that possess comparatively low visual consistency scores. This indicates that the selected token is visually inappropriate, either when assessed for its intrinsic alignment with the image as an isolated entity (**Isolated Token Alignment (ITA)**, *i.e.*, $\text{CLIP}(v, c_i)$) or in terms of how it integrates with the recent sequence context (*i.e.*, its potential Current Contextual Token Alignment, $\text{CLIP}(v, y_{<t} + c_i)$). This detrimental selection often occurs even when other candidate tokens c_i offering better ITA or CCTA scores are available but were assigned lower initial probabilities by the language model.

Synthesizing observations from Figure 3, our key insight is that **hallucinations frequently arise when LVLMs, driven by strong linguistic priors, select tokens with high linguistic coherence but poor visual alignment (either intrinsic or contextual), even when better visual alternatives exist.** Standard decoding methods, focused on language model likelihood, are ill-equipped to prevent this. They lack the crucial real-time, fine-grained visual consistency checks needed to systematically assess a candidate token's true visual appropriateness against the unfolding narrative. This critical gap-the failure to dynamically detect and counteract incipient semantic drift by re-evaluating choices against visual evidence-motivates our *Dynamic Logits Calibration (DLC)*, which is designed to explicitly integrate these *real-time visual signals*, steering generation towards *higher visual fidelity* and tackling the shortcomings of existing approaches, as detailed in the Section 4 and Figure 4.

4 Methodology

Building upon the insights from our analysis in Section 3.3, we propose a **D**ynamic **L**ogits **C**alibration (**DLC**) mechanism integrated directly into the decoding process. As illustrated in Figure 4, DLC comprises two key components: Real-time Visual Alignment Assessment (Section 4.1) and Adaptive Logit Modulation (Section 4.2).

4.1 Real-time Visual Alignment Assessment

At each decoding step t, prior to selecting token y_t, DLC assesses the visual alignment of candidate subsequent tokens with the input image v. This assessment involves establishing a dynamic baseline for visual consistency based on recent generation history and then evaluating each candidate token's potential visual contribution.

First, to quantify the visual grounding of the preceding context, we introduce the **History Contextual Token Alignment (HCTA)**. This score measures the CLIP similarity between the input image v and a fixed size window of N tokens $y_{t-N:t-1}$ immediately preceding the current step, denoted as:

$$S_t^{\text{HCTA}} = \text{CLIP}(v, y_{t-N:t-1}) \tag{3}$$

S_t^{HCTA} reflects the degree to which this recent textual window aligns with the visual input.

To establish a more stable reference reflecting recent trends in visual consistency and mitigate transient fluctuations, we maintain a Historical Baseline $\bar{B}_t$. This baseline is calculated as the smoothed average of the N most recent HCTA scores (S_k^{HCTA} for $k < t$) following a brief warm-up period. $\bar{B}_t$ thus represents the established level of visual fidelity against which the visual contribution of subsequent candidate tokens will be compared.

Next, for each of the top-k candidate tokens $\{c_1, ..., c_k\}$ proposed by the LVLM for the current decoding step t, we evaluate their potential impact on visual alignment using two distinct scores:

- Current Contextual Token Alignment (CCTA). It evaluates the visual alignment resulting from appending a candidate token c_i to the preceding contextual window $y_{t-N:t-1}$, as follows:

  $$S_{t,i}^{\text{CCTA}} = \text{CLIP}(v, y_{t-N:t-1} \oplus c_i) \tag{4}$$

  where $\oplus$ denotes an operation of concatenation.
- Isolated Token Alignment (ITA). This score measures the intrinsic visual alignment of a candidate token c_i with the input image v, independent of the preceding textual context. It is calculated as:

  $$S_{t,i}^{\text{ITA}} = \text{CLIP}(v, c_i) \tag{5}$$

  Note that $S_{t,i}^{\text{ITA}}$ assesses the candidate's standalone visual relevance.

Evaluating candidate tokens based solely on their ITA or CCTA scores can be insufficient. For instance, the preceding context may have drifted from strong visual alignment, potentially skewing the contextual score. Conversely, a token's true visual contribution might only become evident when considered within the sequence, rather than in isolation.

Therefore, to obtain a more robust and nuanced measure of a candidate token's visual suitability, we compute a combined score for each candidate by integrating ITA and CCTA:

$$S_{t,i}^{\text{COMB}} = \frac{S_{t,i}^{\text{CCTA}} + S_{t,i}^{\text{ITA}}}{2} \tag{6}$$

This $S_{t,i}^{\text{COMB}}$ provides a more comprehensive assessment of each candidate's visual appropriateness, considering both its individual properties and its potential integration into the sequence.

4.2 Adaptive Logit Modulation

Having established the historical visual alignment $\bar{B}_t$ and computed the combined scores $S_{t,i}^{\text{COMB}}$ for each candidate token c_i, the core technique lies in using these signals to adaptively modulate the original logits $L_{t,i}$ produced by LVLM. This modulation aims to increase the likelihood of selecting visually grounded tokens, especially when semantic drift is detected relative to the historical baseline.

The modulation process involves three key steps:

1. **Quantifying Relative Visual Contribution:** First, to determine how much a candidate c_i improves or degrades visual alignment relative to the established context, we calculate its **Relative Visual Alignment (RVA)** as follows:

   $$\text{RVA}_{t,i} = \frac{S_{t,i}^{\text{COMB}} - \bar{B}_t}{1 - \bar{B}_t} \tag{7}$$

   A positive RVA means that the candidate is more visually aligned than the recent baseline $\bar{B}_t$, and vice versa. This metric provides a normalized signal of the candidate's relative visual merit.
2. **Adaptive Intervention Strength:** The strength of visual guidance applied shall depend on current state of visual grounding. If the generation is already highly aligned with the image, strong intervention is unnecessary and potentially harmful to fluency. Conversely, if alignment is poor, stronger guidance is needed. We determine this dynamic using a **Dynamic Guidance Strength** λ_t:

   $$\lambda_t = \alpha \cdot (1 - \bar{B}_t)^2 \tag{8}$$

where α controls the maximum modulation strength. The factor $(1 - \bar{B}_t)$ ensures that intervention is applied significantly only when the baseline alignment $\bar{B}_t$ drops, thereby minimizing modulation when visual grounding is already sufficient.

3. **Applying the Logit Adjustment:** Finally, we combine RVA and λ_t to modify the original logit $L_{t,i}$ for each valid candidate c_i. The Adjusted Logit $L'_{t,i}$ is computed multiplicatively:

$$L'_{t,i} = L_{t,i} \cdot \exp(\lambda_t \cdot \sigma(\text{RVA}_{t,i})) \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

This adaptive logit modulation mechanism directly integrates visual feedback into token selection. By dynamically amplifying the probability of visually consistent candidates based on both their relative merit (RVA) and the overall context alignment λ_t, it effectively counteracts misleading semantic priors and promotes the generation of text faithful to the visual input. The complete algorithm and detailed implementation considerations are provided in Appendix A.2 and Appendix A.3, respectively.

5 Experiments

5.1 Experimental Settings

Evaluated LVLMs. We evaluate the effectiveness of our proposed DLC on three state-of-the-art LVLMs: LLaVA-1.5 [30], InstructBLIP [12] and MiniGPT-4 [57]. All benchmarked models adopt a dual-modality architecture, incorporating pre-trained vision encoders like CLIP [35] vision encoder and language models (*e.g.*, Vicuna [10], LLaMA [41]) as fundamental components. Notably, all of models used in our paper are equipped with a 7B LLMs.

Benchmarks. To comprehensively evaluate our proposed method, we employed a suite of established benchmarks. These include:**(1) CHAIR** [36], for quantifying object hallucination in image captions;**(2) GPT-4 Assisted Evaluations** [55], to detect more fine-grained hallucination types;(3) **GPT-4o assisted evaluations** [48], for a comprehensive analysis of hallucinations and text quality; (4) **LLaVA-Bench-in-the-Wild** [30], on which we performed targeted case studies to demonstrate the efficacy of our proposed DLC in addressing complex tasks and adapting to novel domains. Detailed descriptions of each benchmark, are provided in Appendix A.4.

Baselines. To evaluate our training-free LVLM decoding strategies, we have compared them with standard decoding methods, including Nucleus Sampling, Greedy Sampling. We also assessed the performance of our method against several state-of-the-art decoding approaches: VCD [22], ICD [43], SID [18], and OPERA [17]. For comprehensive comparisons, we implemented VCD, ICD, SID, and DLC using both nucleus sampling (Top-p=1). The reported performance of these baselines is based on our re-implementation using their publicly available code.

5.2 Experimental Results

CHAIR Evaluation. Following [17, 18], we randomly sample 500 images from COCO [28] validation and prompt the evaluated models with "Please describe this image in detail" to generate descriptions under two maximum new token limits. Moreover, to address CHAIR's [36] image sensitivity, three distinct image sets were used for robust evaluation. Table 1 shows our method generally surpasses others, validating its efficacy. Notably, our method also exhibits superior generation speed over other elaborately designed approaches. Moreover, as our approach mitigates semantic drift, its superiority is particularly evident with a 512 Max Token setting.

Table 1: **CHAIR hallucination evaluation results.** This table evaluates model hallucination using the CHAIR_S and CHAIR_I metrics, where lower values indicate reduced hallucination and improved performance. TPS indicates tokens per second during generation. All methods use nucleus sampling with Top-p=1. **Bold** and underlined values indicate the best and second-best results, respectively.

Max Token	Methods	LLaVA-1.5			InstuctBLIP			MiniGPT-4		
		$C_S \downarrow$	$C_I \downarrow$	TPS $\uparrow$	$C_S \downarrow$	$C_I \downarrow$	TPS $\uparrow$	$C_S \downarrow$	$C_I \downarrow$	TPS $\uparrow$
64	Sample	25.2	9.11	17.4	31.3	11.3	21.9	25.4	9.74	20.0
	VCD	24.9	8.15	8.92	27.8	9.56	10.7	22.5	8.27	10.2
	ICD	24.0	8.12	8.83	29.3	9.63	9.92	22.6	8.17	9.78
	SID	21.9	6.84	8.82	27.7	9.14	9.86	22.5	8.17	9.14
	DLC(Ours)	**20.9**	**6.22**	16.1	**25.4**	**8.38**	13.08	**21.6**	**7.62**	8.32
512	Sample	52.3	16.4	17.4	56.2	17.6	21.1	34.8	12.3	19.4
	VCD	56.5	17.0	8.74	62.0	18.1	10.4	30.7	10.2	10.0
	ICD	52.1	15.4	8.21	65.4	19.7	10.1	32.5	10.5	9.56
	SID	51.1	14.4	8.61	58.9	16.8	9.88	30.7	10.2	8.92
	DLC(Ours)	**38.4**	**10.8**	14.7	**51.8**	**15.2**	12.7	**29.4**	**9.62**	10.8

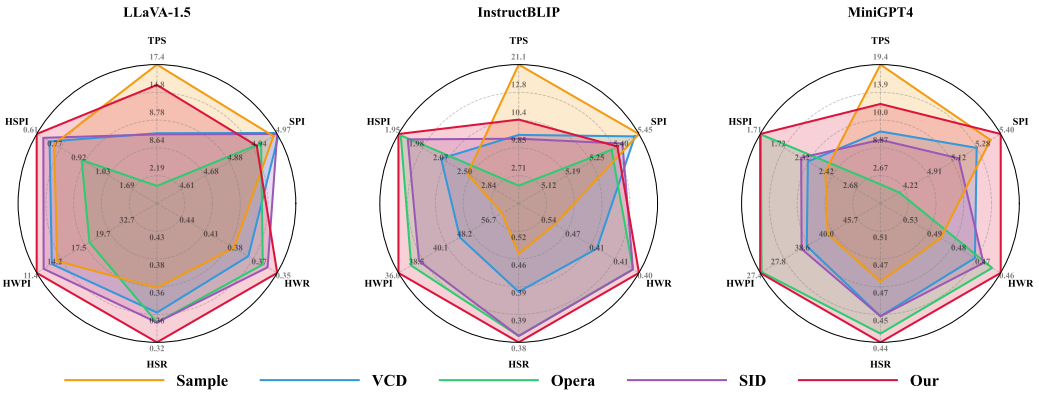

Figure 5: **GPT-4 assisted benchmark [55].** Six aspects of values are analyzed, including the number of sentences per image (SPI), the tokens per second (TPS), the number of hallucinated sentences per image (HSPI), the number of hallucinated words per image (HWPI), the ratio of hallucinated sentences (HSR), and the ratio of hallucinated words (HWR). Note that larger SPI and TPS, smaller HSPI, HWPI, HSR and HWR are better. Larger radar indicates better performance.

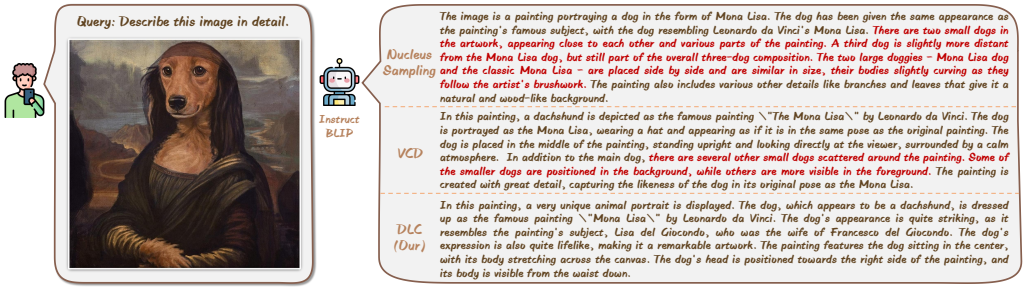

Figure 6: **Qualitative comparison on LLaVA-Bench-in-the-Wild [30].** Under identical sampling decoding strategy, our DLC approach is shown to significantly reduce hallucinatory content and enhance descriptive detail compared to baseline methods like Nucleus Sampling and VCD.

GPT-4 Assisted Evaluation. Following [17, 18, 55], we evaluate on 200 images from the VG dataset using the prompt "Please describe this image in detail," with a maximum of 512 new tokens generated. Additionally, we evaluated the model's generation speed using tokens per second (TPS). As illustrated in Figure 5, we observe that our DLC achieves state-of-the-art results on metrics related to both hallucinated sentences and words. This indicates that DLC can effectively assist models in mitigating hallucination issues stemming from strong semantic priors. Consistent with observations for OPERA [17], we also find that for certain models, DLC may reduce the length of the generated text, which can be attributed to the reduction of superfluous hallucinatory content.

GPT-4o Assisted Evaluation. To further analyze the hallucinations and text quality for open-end generation tasks, we employ GPT-4o as an advanced evaluation proxy. For evaluation, models generate detailed descriptions for 500 randomly selected COCO [28] images, prompted with: "Please describe this image in detail.". Comparative analyses are performed, contrasting our DLC with several established baselines, including Nucleus Sampling, VCD [22], OPERA [17], and SID [18]. The empirical results, summarized in Table 2, indicate that DLC consistently outperforms these baselines on most metrics. In particular, DLC demonstrates a significant improvement in correctness over conventional sampling-based decoding, while maintaining a comparable degree of detail. The close alignment of GPT-4o's perceptual and reasoning faculties with human judgment lends substantial credibility to these findings, suggesting that the observed enhancements in hallucination reduction and text quality are indicative of genuine improvements from a human evaluative perspective (full GPT-4/GPT-4o prompt is provided in the Appendix A.12).

Case Study on LLaVA-Bench. To complement our use of dedicated metrics like CHAIR [36] and evaluations facilitated by GPT-4/GPT-4o, we present a qualitative analysis of DLC's hallucination mitigation capabilities through case studies from the LLaVA-Bench-in-the-Wild dataset [30]. As

Table 2: **Assessing hallucination on COCO [28] with GPT-4o.** Performance is assessed on correctness (C) and detailedness (D).

Settings	LLaVA-1.5		InstructBLIP		MiniGPT-4	
	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$
Nucleus	5.34	6.18	4.76	5.61	5.15	5.73
Our	**7.73**	**6.87**	**7.38**	**6.75**	**6.64**	**6.56**
VCD	5.55	6.38	5.32	6.09	5.55	5.97
Our	**7.45**	**6.67**	**6.94**	**6.54**	**6.39**	**6.44**
OPERA	6.16	6.59	6.02	6.17	5.98	5.93
Our	**7.10**	**6.67**	**6.51**	**6.55**	**6.16**	**6.45**
SID	5.69	6.54	5.34	6.07	5.55	6.05
Our	**7.17**	**6.55**	**6.89**	**6.62**	**6.54**	**6.45**

Table 3: **Assessing hallucination on LLaVA-Bench [30] with GPT-4o.** Performance is assessed on correctness (C) and detailedness (D).

Settings	LLaVA-1.5		InstructBLIP		MiniGPT-4	
	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$
Nucleus	4.71	5.79	4.21	5.00	5.17	5.33
Our	**8.42**	**7.04**	**7.25**	**6.42**	**6.00**	**5.75**
VCD	5.25	5.96	4.88	5.54	5.00	5.08
Our	**7.46**	**6.58**	**6.58**	**6.17**	**5.67**	**5.46**
OPERA	5.75	6.75	5.88	5.62	5.75	**6.08**
Our	**7.54**	**6.88**	**5.96**	**6.21**	**5.79**	5.62
SID	5.58	**6.79**	4.42	5.25	5.38	5.42
Our	**6.71**	6.08	**6.38**	**6.29**	**5.88**	**5.79**

illustrated in Figure 6 (with additional cases in Appendix A.13), the DLC approach markedly reduces hallucinatory content and enhances descriptive detail under identical image and prompt conditions. These qualitative observations are further supported by a comprehensive GPT-4o evaluation of all LLaVA-Bench generation outputs, detailed in Table 3, which substantiates DLC's superior performance over competing methods.

5.3 Ablation Studies

We primarily validate the effectiveness of the adaptive factor λ_t. The Appendix provides a large number of additional experiments, including comprehensive ablation studies on key components and hyperparameters. These experiments also demonstrate DLC's robustness across different CLIP models, its broader applicability to VQA tasks, and its scalability to larger 13B architectures.

Effectiveness of the adaptive factor in λ_t. To demonstrate the effectiveness of the adaptive factor $(1-\bar{B}_t)^2$ in dynamic guidance strength λ_t, as formulated in Eq. 8, we configure models with and without this factor to generate responses on the COCO [28] dataset. The correctness (C) and detailedness (D) of these responses are evaluated using GPT-4o. As illustrated in Figure 7, models including this factor achieve higher scores in both accuracy and details compared to their counterparts without the factor, thereby validating its efficacy.

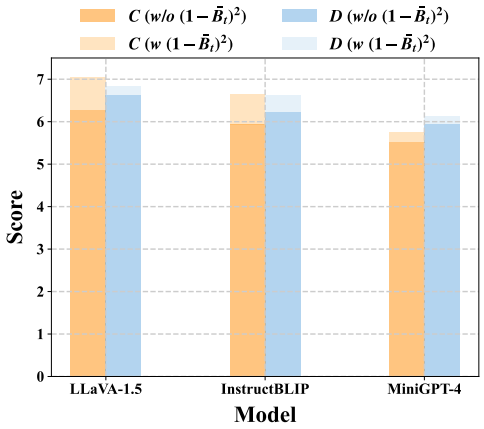

Figure 7: **Effectiveness of the adaptive factor $(1-\bar{B}_t)^2$ in λ_t.**

6 Conclusion

This paper introduces Dynamic Logits Calibration (DLC), an innovative training-free decoding framework designed to mitigate hallucinations in LVLMs by dynamically aligning text generation with visual evidence during inference. Observing that hallucinations often arise from the models' increasing reliance on linguistic priors and progressive detachment from visual input, DLC employs CLIP to step-wise assess image-text semantic alignment and evaluates the Relative Visual Advantage (RVA) of candidate tokens against a dynamically updated contextual baseline. This adaptively adjusts output logits to favor visually grounded tokens, further refined by an adaptive weighting mechanism that balances visual guidance with textual quality. Extensive experiments on various LVLMs demonstrate that DLC significantly reduces hallucinations and outperforms existing methods while maintaining high inference efficiency. In Appendix A.11, we discuss the limitations of our method and its broader implications.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[5] Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*, 2025.

[6] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024.

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[8] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14093–14100. IEEE, 2024.

[9] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

[11] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Th6NyL07na`.

[12] 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. URL `https://arxiv.org/abs/2305.06500`.

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

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics*, pages 4171–4186, 2019.

[15] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5078–5088, 2022.

[16] Yudong Han, Liqiang Nie, Jianhua Yin, Jianlong Wu, and Yan Yan. Visual perturbation-aware collaborative learning for overcoming the language prior problem. *arXiv preprint arXiv:2207.11850*, 2022.

[17] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[18] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rsZwwjYHuD.

[19] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 123(1):32–73, May 2017. doi: 10.1007/s11263-016-0981-7.

[21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[22] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[23] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild. May 2024. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

[24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=xozJw0kZXF.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014*, pages 740–755, 2014.

[29] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=J44HfH4JCg.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[32] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LBl7Hez0fF.

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[34] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[36] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. doi: 10.18653/v1/D18-1437.

[37] Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-LLMs via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8228–8248. Association for Computational Linguistics, November 2024.

[38] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[39] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=zGb4WgCW5i`.

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[42] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[43] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853, 2024. URL `https://aclanthology.org/2024.findings-acl.937`.

[44] Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024.

[45] Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*, 2024.

[46] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. In *Forty-second International Conference on Machine Learning*, 2025.

[47] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[48] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[49] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[50] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[51] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11766–11781. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.633.

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[53] Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P. Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=tTBXePRKSx`.

[54] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024.

[55] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

[56] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=oZDJKTlOUe`.

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[58] Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation for large vision-language model via visual-aware sparsification. *arXiv preprint arXiv:2501.06553*, 2025.

A Appendix

This appendix provides supplementary materials to support the main findings of our paper. We begin with an overview of Large Vision-Language Models (LVLMs) in Section A.1, offering additional context to help readers understand the foundation of our research. Section A.2 presents a detailed pseudo-code of our proposed method, providing an in-depth algorithmic understanding. Section A.3 elaborates on the implementation details, including parameter settings and computational workflows. For a comprehensive evaluation of our approach, Section A.4 provides further details on the benchmarks used in our experiments. Section A.5 analyzes the impact of two key hyperparameters on performance: the modulation strength α and the historical context window size N. Section A.6 analyzes the robustness of DLC across different CLIP model variants. Section A.7 presents the extension of DLC to Visual Question Answering tasks, demonstrating its applicability beyond open-ended generation. Section A.8 provides comprehensive ablation studies validating the importance of each component in our framework. Section A.9 demonstrates DLC's effectiveness on larger 13B-parameter models, confirming its scalability beyond 7B architectures. Section A.10 demonstrates the integration of DLC with various decoding strategies, including beam search, greedy search, top-k sampling, and temperature scaling. Section A.11 discusses the inherent limitations of our method, while Section A.13 presents additional comparative results of our approach against other methods on the LLaVA-Bench benchmark. Finally, Section A.12 showcases the prompts used for GPT-4 and GPT-4o evaluations.

A.1 More backgrounds about LVLMs

The development of LVLMs has evolved from BERT-based [14, 33, 38] cross-modal alignment to LLM-powered [1, 7, 12, 26, 30, 57] architectures. Building upon foundational text-image alignment techniques like CLIP [35] and BLIP [25], contemporary LVLMs integrate visual encoders with LLM decoders through distinct optimization strategies. For instance, LLaVA [30] employs lightweight adapters combined with instruction tuning, while InstructBLIP [12] and MiniGPT-4 [57] utilize Q-Former modules for visual token compression. Despite these innovations, significant challenges persist, particularly severe hallucination issues arising from misaligned multimodal representations. To systematically analyze mitigation strategies, we have examined with three representative architectures, including LLaVA, InstructBLIP, and MiniGPT-4, whose results demonstrate that various technical approaches commonly share critical vulnerabilities to the hallucination issue.

A.2 Pseudo-code of Our Method

In this section, we provide a detailed pseudo-code of our Dynamic Logits Calibration (DLC) method, as shown in Algorithm 1. The pseudo-code encompasses both key modules: Real-time Visual Alignment Assessment and Adaptive Logit Modulation.

As shown in Algorithm 1, DLC dynamically adjusts the output logits of LVLMs during decoding to mitigate hallucinations. First, the model undergoes a brief warm-up period (line 6), allowing generation of sufficient context for CLIP assessment. Subsequently, at each decoding step, the current History Contextual Token Alignment (HCTA) score is calculated (lines 7-8), and a historical baseline $\bar{B}_t$ is maintained (lines 9-13).

For each candidate token, DLC computes three key scores:

- Current Contextual Token Alignment (CCTA): measuring the visual alignment of the candidate token with the preceding context (line 16)
- Isolated Token Alignment (ITA): measuring the inherent alignment of the individual candidate token with the input image (line 17)
- Combined score: integrating CCTA and ITA by averaging (line 18)

Based on these scores, DLC calculates the Relative Visual Advantage (RVA) for each candidate token (line 19), which represents the improvement in visual alignment relative to the historical baseline. The dynamic guidance strength λ_t (line 20) adaptively adjusts according to the current state of visual alignment, providing stronger intervention when alignment decreases.

Finally, DLC modifies the original logits through RVA and λ_t (line 21), thereby enhancing the probability of visually consistent tokens. The entire process integrates into the normal decoding flow

Algorithm 1 Dynamic Logits Calibration (DLC)

Require: Input image v, text prompt x, vision-language model M_θ, CLIP model, window size N, modulation strength α
Ensure: Generated text sequence y

$y \leftarrow x$ ▷ Initialize output sequence with input prompt
$\bar{B} \leftarrow 0$ ▷ Initialize historical baseline
HCTA_history $\leftarrow$ empty queue ▷ Store historical HCTA scores
for $t = |x| + 1$ **to** $|x| +$ max_new_tokens **do**
  $L_t \leftarrow M_\theta(y_{<t}, v)$ ▷ Get original logits for current step
  if $t > |x| + 3$ **then** ▷ Start applying DLC after warm-up period
    $y_{t-N:t-1} \leftarrow$ get previous N tokens ▷ Get preceding context window
    $S_t^{\text{HCTA}} \leftarrow \text{CLIP}(v, y_{t-N:t-1})$ ▷ Calculate History Contextual Token Alignment
    Add S_t^{HCTA} to HCTA_history
    if HCTA_history length $> N$ **then**
      Remove oldest element from HCTA_history
    end if
    $\bar{B}_t \leftarrow$ average(HCTA_history) ▷ Calculate historical baseline
    $C \leftarrow$ get top-k candidate tokens(L_t) ▷ Get top-k candidates
    for each candidate token $c_i \in C$ **do**
      $S_{t,i}^{\text{CCTA}} \leftarrow \text{CLIP}(v, y_{t-N:t-1} \oplus c_i)$ ▷ Current Contextual Token Alignment
      $S_{t,i}^{\text{ITA}} \leftarrow \text{CLIP}(v, c_i)$ ▷ Isolated Token Alignment
      $S_{t,i}^{\text{COMB}} \leftarrow \frac{1}{2}(S_{t,i}^{\text{CCTA}} + S_{t,i}^{\text{ITA}})$ ▷ Combined score
      $\text{RVA}_{t,i} \leftarrow \frac{S_{t,i}^{\text{COMB}} - \bar{B}_t}{1 - \bar{B}_t}$ ▷ Relative Visual Advantage
      $\lambda_t \leftarrow \alpha \cdot (1 - \bar{B}_t)^2$ ▷ Dynamic guidance strength
      $L'_{t,i} \leftarrow L_{t,i} \cdot \exp(\lambda_t \cdot \sigma(\text{RVA}_{t,i}))$ ▷ Apply logit adjustment
    end for
    Replace original logits L_t with adjusted logits L'_t
  end if
  $y_t \leftarrow \text{sample}(L_t)$ ▷ Sample next token based on adjusted logits
  $y \leftarrow y \oplus y_t$ ▷ Append new token to output sequence
  if y_t is end token **then**
    break
  end if
end for
return y

without requiring additional forward passes or fine-tuning, thus providing a computationally efficient and effective solution for hallucination mitigation.

A.3 Implementation Details.

In this section, we provide comprehensive implementation details of our Dynamic Logits Calibration (DLC) method, including model configuration, parameter settings, and computational workflows. These details are essential for reproducing our results and understanding the practical aspects of DLC.

We compute all visual alignment scores using the SigLIP model [52]. The historical baseline $\bar{B}_t$ is calculated over a sliding window of $N = 8$ past context scores, following a 3-step warm-up period. For the top-k candidates ($k = 50$) at each step, the combined score $S_{t,i}^{\text{COMB}}$ is computed. The adaptive modulation employs a maximum intervention strength $\alpha = 3$, and the standard sigmoid function $\sigma(\cdot)$ for logit adjustment. Our experiments were run on a single A100 GPU.

A.4 Further Details on Benchmarks

To comprehensively evaluate our DLC method, we utilized several benchmarks specifically designed to assess hallucination in vision-language models. This section provides additional details about these benchmarks, including dataset characteristics, evaluation protocols, and metrics interpretation.

(1) **CHAIR** [36] quantifies object hallucination in image captions by measuring the ratio of objects mentioned in captions but absent from ground-truth labels. It includes two variants: CHAIR_I (C_I), calculating hallucinated objects versus all generated objects, and CHAIR_S (C_S), assessing sentences with hallucinations versus total sentences. These two metrics are defined as follows::

$$C_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, C_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|} \tag{10}$$

Lower scores indicate better hallucination suppression.

(2) **GPT-4 Assisted Evaluations** [55] function as a crucial complement to conventional assessment methodologies. While traditional metrics such as CHAIR effectively quantify hallucinations at the object-existence level, they demonstrate inadequacy in detecting more fine-grained hallucination types, including positional, relational, and attribute-based inaccuracies. To address this limitation, GPT-4 assisted evaluation employs detailed object-level descriptions from the Visual Genome (VG) [20] dataset as ground-truth reference, leveraging GPT-4's sophisticated capabilities for precise hallucination detection and classification. The complete prompt template is provided in the Appendix A.12.

(3) **GPT-4o assisted evaluations**[2] [48] further conduct a comprehensive analysis of hallucinations and text quality in open-ended generation tasks. Harnessing GPT-4o's sophisticated capabilities, [48] established a numerical assessment protocol utilizing a 0-10 scale, which evaluates two essential dimensions: Accuracy-the extent to which responses align with image content without fabrication, and Detailedness-the degree of informational richness present in the responses.

(4) **LLaVA-Bench-in-the-Wild** [30] is designed to evaluate LVLMs' proficiency in addressing complex tasks and their adaptability to novel domains. We perform targeted case studies on this dataset to demonstrate the efficacy of our proposed DLC.

A.5 Sensitivity Analysis of Key Hyperparameters: α and N

The performance of our DLC method is influenced by two key hyperparameters: the modulation strength α and the context window size N. In this section, we conduct a comprehensive sensitivity analysis to understand how these parameters affect the model's hallucination mitigation capability and generation quality. Experiments were conducted using a range of values for these parameters across three distinct LVLMs, with performance assessed by the CHAIR benchmark [36] (max new tokens: 512 for all ablation studies).

Table 4: Impact of Modulation Strengths α on CHAIR Metrics for Different LVLMs and Decoding Strategies

Method	α	LLaVA-1.5		InstructBLIP		MiniGPT-4	
		$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Greedy Search	1.0	**33.8**	**9.85**	**47.9**	**13.2**	**32.6**	10.3
	2.0	34.1	10.0	48.1	13.5	33.7	10.1
	3.0	34.8	10.1	49.6	13.5	33.0	**9.89**
	4.0	35.3	9.87	50.5	14.2	33.5	10.0
	5.0	36.8	10.5	51.0	14.4	33.9	10.6
Nucleus Sampling	1.0	41.8	12.7	55.4	15.4	34.6	11.1
	2.0	40.3	11.4	53.2	14.8	34.0	**10.6**
	3.0	38.4	10.9	51.8	15.2	**32.7**	10.7
	4.0	**37.5**	**10.8**	50.8	14.7	33.6	10.7
	5.0	**37.5**	10.9	**50.6**	**14.5**	33.6	10.9

First, we analyze the effect of varying modulation strengths α. The optimal setting for α is contingent upon the specific LVLM and the decoding strategy employed. As detailed in Table 4, under Greedy Search, lower α values, particularly $\alpha = 1.0$, tend to be more effective for LLaVA-1.5

[2]We utilize GPT-4o for assisted evaluation, as it offers enhanced capabilities over previous vision-capable models like GPT-4V, which is also being phased out by OpenAI.

Table 5: Impact of Window Size N on CHAIR Metrics for Different LVLMs and Decoding Strategies

Method	N	LLaVA-1.5		InstructBlip		MiniGPT4	
		$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Greedy Search	4	**34.4**	**9.45**	52.2	14.6	35.4	11.1
	8	35.1	9.72	49.0	**12.8**	33.4	10.3
	12	**34.4**	9.51	48.0	13.5	33.2	9.85
	16	37.0	9.84	**47.0**	13.2	**31.0**	**9.69**
Nucleus Sampling	4	38.6	11.1	51.4	**14.3**	**31.0**	10.2
	8	39.4	**10.9**	51.0	**14.3**	32.0	11.2
	12	38.6	11.5	54.8	14.8	32.4	**10.1**
	16	**37.6**	11.4	**50.4**	15.0	34.0	12.6

and InstructBLIP. For MiniGPT-4 under Greedy Search, competitive results are observed at both $\alpha = 1.0$ and $\alpha = 3.0$. Conversely, when employing Nucleus Sampling, moderate to higher α values (typically ranging from 3.0 to 5.0) generally yield improved CHAIR scores, with LLaVA-1.5 and MiniGPT-4 showing optimal performance around $\alpha = 4.0$ and $\alpha = 3.0$ respectively, while InstructBLIP benefits from α values up to 5.0. Therefore, while a global value of $\alpha = 3.0$ is adopted for the main experiments in this paper, these findings underscore that tailoring α to specific model architectures and decoding methodologies can yield further performance enhancements.

Next, we examine the impact of the window size N. The results, presented in Table 5, indicate varied responses based on the model and decoding strategy. Under Greedy Decoding, both InstructBLIP and MiniGPT-4 generally deliver improved performance with larger window sizes. Conversely, for LLaVA-1.5 using Greedy Decoding, the influence of N is less definitive, suggesting that smaller to moderate N values might be more advantageous. Nucleus Sampling exhibits distinct responses to variations in N: MiniGPT-4 achieves its best performance with a small window, with larger N values leading to a decline; LLaVA-1.5 shows more fluctuating results; and InstructBLIP, which generally performs suboptimally with Nucleus Sampling, does not yield significant improvements with changes in N. Thus, while a uniform window size of $N = 8$ is currently utilized in our main experiments, our findings highlight that tailoring N to specific models and decoding methodologies can substantially boost performance.

In summary, both the modulation strength α and the window size N are influential hyperparameters for DLC. While default values are used for general demonstration in the main body of this paper, the ablation studies presented here clearly indicate that optimal performance is achieved when these parameters are carefully tuned according to the specific LVLM and the selected decoding strategy.

A.6 Robustness Analysis Across Different CLIP Models

To analyze the robustness of our DLC method and its dependence on the underlying CLIP model, we conducted comprehensive experiments using different CLIP backbone architectures. This analysis addresses concerns about the method's reliance on a specific vision-language alignment model.

Table 6 presents the performance of DLC when integrated with various CLIP variants, including the original CLIP models, SigLIP, and the recent FG-CLIP model.

Table 6: **Robustness Analysis Across Different CLIP Models.** CHAIR_S and CHAIR_I scores (lower is better) for DLC when using different CLIP backbones. Best results are highlighted in **bold**.

CLIP Backbone	LLaVA-1.5		InstructBLIP	
	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$
CLIP-ViT-Large-Patch14	40.4	11.4	51.8	14.7
CLIP-ViT-Large-Patch14-336	39.7	10.6	51.2	14.4
SigLIP-SO400M-Patch14-384	39.4	10.9	51.0	14.3
FG-CLIP-Large [46]	**37.8**	**10.4**	**49.8**	**13.9**

The results demonstrate that while stronger CLIP models generally lead to better performance, DLC maintains consistent improvements across all tested architectures. This robustness indicates that our method is not brittlely tied to a specific CLIP variant, but rather benefits from the general principle of visual-textual alignment assessment. The superior performance with FG-CLIP, which specializes in fine-grained visual-textual alignment, particularly validates our approach's potential for further enhancement as CLIP models continue to evolve.

A.7 Extension to Visual Question Answering Tasks

While DLC was initially designed for open-ended descriptive generation tasks, we have successfully extended it to Visual Question Answering (VQA) domains through simple modifications. By treating the user's query content as already generated context, DLC demonstrates excellent performance on VQA tasks that require semantic-level reasoning.

We validated this extension on the popular POPE benchmark [27], which evaluates object existence in images through yes/no questions across three different setups: random, popular, and adversarial. The results are presented in Table 7.

The results demonstrate the effectiveness of DLC, validated by its successful extension from open-ended generation to structured VQA scenarios.

Table 7: **Performance on POPE Benchmark.** Results show DLC's effectiveness when extended to VQA tasks. Accuracy (Acc) and F1 scores are reported for different setups. Best results are highlighted in **bold**.

Method	Setup	LLaVA-1.5		InstructBLIP	
		Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑
Sample	Random	83.90	84.23	83.13	83.43
ICD		85.33	85.48	82.46	83.68
VCD		82.50	83.20	83.60	83.69
SID		87.56	87.10	84.36	84.43
DLC+SigLIP (Ours)		87.83	88.14	88.93	88.81
DLC+FG-CLIP (Ours)		**88.87**	**88.88**	**89.53**	**89.32**
Sample	Popular	80.93	81.85	77.20	78.78
ICD		82.16	82.80	74.20	77.66
VCD		79.86	81.13	77.13	78.68
SID		84.16	84.11	77.10	78.98
DLC+SigLIP (Ours)		84.37	85.23	**82.73**	**83.69**
DLC+FG-CLIP (Ours)		**85.33**	**85.84**	82.53	83.42
Sample	Adversarial	75.57	77.83	74.80	77.02
ICD		76.33	78.39	72.03	76.15
VCD		75.10	77.66	75.06	76.99
SID		**79.60**	80.43	76.03	78.07
DLC+SigLIP (Ours)		77.00	79.68	79.03	80.82
DLC+FG-CLIP (Ours)		78.70	**80.68**	**79.57**	**81.10**

A.8 Comprehensive Component Ablation Studies

To thoroughly validate the importance of each component in our proposed DLC framework, we conducted comprehensive ablation studies. We systematically evaluated the impact of removing individual components: Current Contextual Token Alignment (CCTA), Isolated Token Alignment (ITA), and Relative Visual Advantage (RVA).

The ablation experiments were conducted using GPT-4o assisted evaluation on the COCO dataset, with models generating detailed descriptions for 500 randomly selected images. Performance is assessed on two key dimensions: Correctness (accuracy of visual descriptions) and Detailedness (richness of information).

Table 8: **Comprehensive Ablation Study Results.** GPT-4o evaluation scores (0-10 scale, higher is better) demonstrating the importance of each DLC component. Best results are highlighted in **bold**.

Settings	LLaVA-1.5		InstructBLIP	
	Correctness↑	Detailedness↑	Correctness↑	Detailedness↑
w/o CCTA	6.22	6.65	5.79	6.20
w/ CCTA	**7.00**	**6.72**	**6.60**	**6.46**
w/o ITA	6.22	6.61	5.84	6.30
w/ ITA	**6.97**	**6.72**	**6.63**	**6.52**
w/o RVA	6.35	6.51	5.83	6.25
w/ RVA	**7.00**	**6.69**	**6.71**	**6.60**

The results clearly demonstrate that each component contributes significantly to the overall performance of DLC.

A.9 Scalability to Larger Model Architectures

To validate the effectiveness of our method on larger-scale models, we conducted comprehensive experiments on 13B-parameter architectures, specifically evaluating DLC on LLaVA-1.5-13B and InstructBLIP-13B using both the CHAIR benchmark and GPT-4o assisted evaluation.

A.9.1 CHAIR Benchmark Results on 13B Models

Table 9 presents the CHAIR evaluation results for 13B models across different maximum token settings. The results consistently demonstrate that DLC provides substantial improvements over nucleus sampling baselines on both model architectures.

The results show remarkable improvements across both token length settings

Table 9: **CHAIR Evaluation Results on 13B Models.** Comparison of nucleus sampling baseline and DLC on larger model architectures. Lower CHAIR scores indicate better performance (fewer hallucinations). Best results are highlighted in **bold**.

Methods	Max Token	LLaVA-1.5-13B		InstructBLIP-13B	
		$CHAIR_S\downarrow$	$CHAIR_I\downarrow$	$CHAIR_S\downarrow$	$CHAIR_I\downarrow$
Nucleus Sampling	64	23.20	8.16	27.80	10.22
DLC (Ours)	64	**18.20**	**5.53**	**24.20**	**7.75**
Nucleus Sampling	512	53.20	15.07	64.00	19.57
DLC (Ours)	512	**33.00**	**8.80**	**55.60**	**15.37**

A.9.2 GPT-4o Assisted Evaluation on 13B Models

To complement the CHAIR evaluation, we conducted GPT-4o assisted evaluation focusing on correctness and detailedness of generated descriptions. The results are presented in Table 10.

The GPT-4o evaluation demonstrates substantial improvements in both correctness and detailedness.

Table 10: **GPT-4o Evaluation Results on 13B Models.** Performance comparison on correctness and detailedness metrics. Best results are highlighted in **bold**.

Settings	LLaVA-1.5-13B		InstructBLIP-13B	
	Correctness↑	Detailedness↑	Correctness↑	Detailedness↑
Nucleus Sampling	5.25	6.05	4.26	5.06
DLC (Ours)	**7.75**	**6.77**	**7.00**	**6.36**

A.10 DLC Performance with Various Decoding Strategies

Our proposed DLC demonstrates compatibility with various mainstream decoding strategies. To comprehensively evaluate its effectiveness, in addition to its performance with common strategies such as nucleus sampling and greedy sampling, we conducted further experiments on the LLaVA-1.5 [30] model using the MSCOCO dataset [28] and the CHAIR benchmark [36]. These experiments explored other decoding methods and parameters, specifically: Top-p sampling (p=0.9), Top-k sampling (k=50), and Top-k sampling (k=50) with varying temperatures (t=1.5 and t=0.8). The results, presented in Figure 8, consistently indicate that regardless of the sampling method employed, DLC effectively reduces hallucinations and enhances overall model performance. This substantiates the robustness and effectiveness of our proposed DLC across diverse decoding strategies.

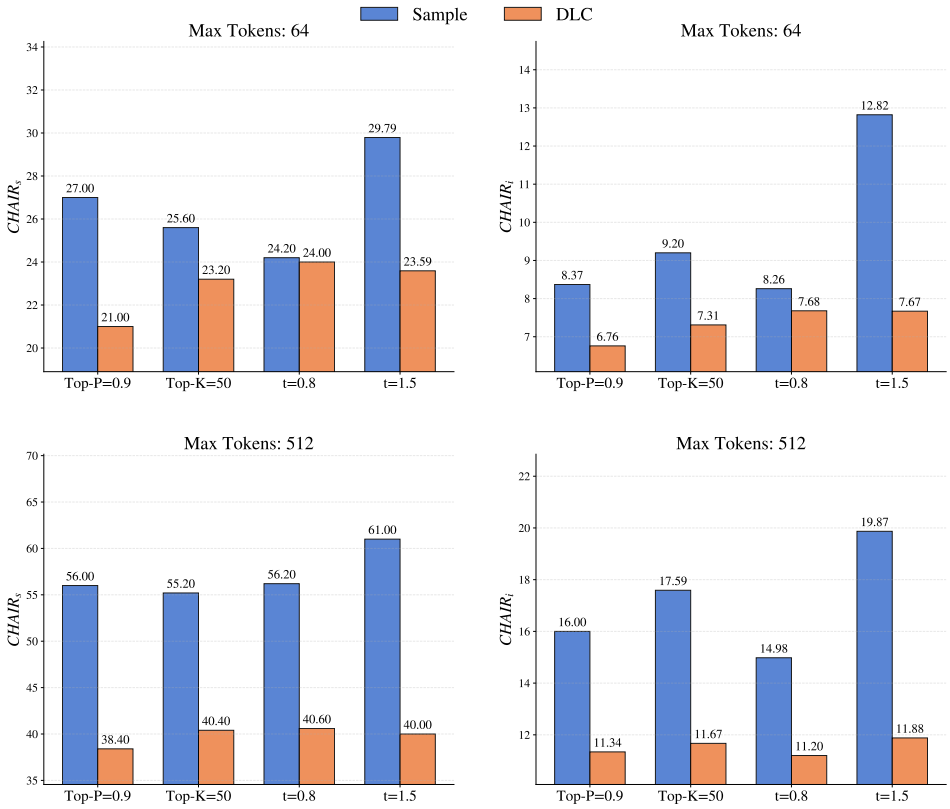

Figure 8: **Results of different decoding strategies.**

A.11 Discussion of Methodological Limitations and Social impacts.

In this section, we delineate two limitations inherent to the proposed DLC method:

Limited Applicability to Spatial and Coordinate-Level Tasks: While our extension to VQA tasks shows promising results on object existence verification (POPE benchmark [27]), DLC faces inherent limitations with tasks requiring precise spatial reasoning or coordinate-level grounding. Our method relies on CLIP's global semantic assessment, which lacks sensitivity to fine-grained spatial details or numerical coordinates. This limitation affects tasks such as: (1) object counting and spatial relationship reasoning (e.g., "How many red cars are to the left of the blue building?"), (2) precise coordinate grounding tasks like object detection with bounding box generation, and (3) referring expression comprehension requiring exact localization.

Dependence on Visual Content Quality: DLC's effectiveness is inherently tied to the quality and clarity of the input visual content. For images with poor quality, ambiguous visual information, the method's ability to provide reliable calibration signals may be compromised. In such cases, the visual

alignment scores from CLIP may not provide sufficiently discriminative information to guide token selection effectively, potentially limiting the method's benefits.

Social impacts. Dynamic Logits Calibration (DLC) holds significant promise for advancing the field of LVLMs. By offering a novel training-free decoding framework that dynamically aligns text generation with visual evidence, DLC effectively mitigates the pervasive issue of hallucination in LVLMs. This approach overcomes critical limitations of existing strategies, such as reliance on static constraints and the inefficiency of multiple forward passes, thereby enhancing the reliability and practical applicability of these models without incurring additional training costs. DLC not only provides an effective and efficient solution to a key challenge but also serves as an inspiration for the research community to explore new pathways for improving visual grounding and tackling dynamic vision-language alignment, fostering the development of more robust and trustworthy LVLMs.

A.12 Prompts for GPT-4 and GPT-4o Evaluation

GPT-4 Prompt

Please help me judge if the comment of this image is hallucination or correct.
I will give you a list of region description of a image. The format is [x1, y1, x2, y2]: region description, where [x1, y1, x2, y2] is the bounding box of the region. Highly overlapping bounding boxes may refer to the same object. This is the ground truth information of the image. Besides, I give you some factual information about the content of the image (which is 100% accurate). Your judgement should base on this information. However, this information only descibe the objects in the region of image, so it cannot descibe the subjective part of the image, e.g., atmosphere, style, emotion. In that case, you can return "Cannot judge".
Also, I will give you a list of comments of the image for you to judge if it is hallucination. Please give a judgement one by one along with the reason.

Your output should be:
Judgement:
1. hallucination or correct or cannot judge: <reason>
2. ...

Here are the region descriptions of the image:
{}

Factual Information:
{}

Here is the comment for you to judge (hallucination, correct, or cannot judge):
{}

Figure 9: **Prompts of GPT-4 for evaluations.**

GPT-4o Prompt

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:
1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinationsshould be given higher scores.
2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not countas necessary details.
Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

[Assistant 1]
{}
[End of Assistant 1]

[Assistant 2]
{}
[End of Assistant 2]

Output format:
Accuracy: <Scores of the two answers>
Reason:

Detailedness: <Scores of the two answers>
Reason:

Figure 10: **Prompts of GPT-4o for evaluations.**

A.13 Extended Comparative Results on LLaVA-Bench

In this section, we present additional comparative results of our DLC approach against other methods on the LLaVA-Bench [30]. These extended evaluations provide a more comprehensive understanding of our method's performance across various question types and scenarios.

Figure 11: **Case Study:** LLaVA-1.5 on LLaVA-Bench-in-the-Wild (Hallucinations Marked Red).