Diffusion-CAM: Faithful Visual Explanations for dMLLMs

Haomin Zuo[2], Yidi Li[3], Luoxiao Yang[4], Xiaofeng Zhang[1]†
[1]Department of Automation and Intelligent Sensing, Shanghai Jiao Tong University
[2]Sun Yat-sen University [3]Northwestern University
[4]Technion - Israel Institute of Technology
{framebreak@}sjtu.edu.cn

Abstract

While diffusion Multimodal Large Language Models (dMLLMs) have recently achieved remarkable strides in multimodal generation, the development of interpretability mechanisms has lagged behind their architectural evolution. Unlike traditional autoregressive models that produce sequential activations, diffusion-based architectures generate tokens via parallel denoising, resulting in smooth, distributed activation patterns across the entire sequence. Consequently, existing Class Activation Mapping (CAM) methods, which are tailored for local, sequential dependencies, are ill-suited for interpreting these non-autoregressive behaviors. To bridge this gap, we propose **Diffusion-CAM**, the first interpretability method specifically tailored for dMLLMs. We derive raw activation maps by differentiably probing intermediate representations in the transformer backbone, accordingly capturing both latent features and their class-specific gradients. To address the inherent stochasticity of these raw signals, we incorporate four key modules to resolve spatial ambiguity and mitigate intra-image confounders and redundant token correlations. Extensive experiments demonstrate that Diffusion-CAM significantly outperforms SoTA methods in both localization accuracy and visual fidelity, establishing a new standard for understanding the parallel generation process of diffusion multimodal systems. Code is available at https://github.com/ZzzzzZhhmm/Diffusion-CAM

1 Introduction

Multimodal Large Language Models (MLLMs) have fundamentally transformed Artificial Intelligence, enabling seamless cross-modal understanding and reasoning (Zhao et al., 2023). The dominant paradigm has long centered on autoregressive architectures—such as Qwen-VL series (Wang et al., 2024b; Bai et al., 2023; Team, 2025), alongside others (Alayrac et al., 2022; Liu et al., 2023; Achiam et al., 2023)—which generate text sequentially based on explicit attention mechanisms. However, a significant paradigm shift is underway. Recent innovations are pioneering diffusion-based architectures, exemplified by LaViDa (Li et al., 2025a,b,c, 2026), LLaDA-V (You et al., 2025; Nie et al., 2025), MMaDA (Yang et al., 2025), and Dream-VL (Ye et al., 2025a,b). Unlike autoregressive models that function as a "sequential chain," these diffusion MLLMs employ parallel masked diffusion, conceptualizing the entire sentence simultaneously. This shift enhances both generation speed and global coherence (Song and Ermon, 2019; Song et al., 2020). But as the community pivots toward these parallel architectures, a critical question arises: *Are our interpretability tools keeping pace?* Understanding model decisions is a prerequisite for trustworthy AI (Mosbach et al., 2024; Wang et al., 2024a; Pan et al., 2024). In the autoregressive realm, visual explanation is mature, spanning from gradient-based CAMs, which provide a natural way to spatially

Figure 1: Visual comparison of (A) [LLaVA-CAM] and (B) [Diffusion-CAM] (ours). Our method generates more focused and clearer activation maps.

† Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12085–12101
July 2-7, 2026

ground model predictions by weighting feature channels with target gradients, (Zhou et al., 2016; Selvaraju et al., 2017; Jiang et al., 2021; Chattopadhay et al., 2018; Li et al., 2025d; Chowdhury et al., 2024; Omeiza et al., 2019) to attention-based approaches (Abnar and Zuidema, 2020; Chefer et al., 2021; Ferrando and Voita, 2024; Wei and Zhang, 2024; Zhang et al., 2025a, 2026; Zhao et al., 2026; Chang et al., 2025). Methods like LLaVA-CAM (Zhang et al., 2024) and Token Activation Maps (TAM) (Li et al., 2025e) rely heavily on the sequential, attention-rich nature of these models to trace specific token generation. However, the very advantages of dMLLMs—**parallel generation and global planning**—pose a fundamental challenge to these frameworks. DMLLMs operate by progressively denoising a global context without explicit token-wise attention weights (Xu et al., 2023). Consequently, applying traditional CAM methods results in diffuse, non-specific heatmaps (as shown in Figure 1), failing to disentangle the model's decision process for specific objects.

To address this limitation, we propose **Diffusion-CAM**, a gradient-based visual explanation framework tailored to the diffusion MLLMs. Unlike existing CAM-style methods built around autoregressive token dependencies, Diffusion-CAM is designed for the common dMLLM setting in which images and prompts provide fixed multimodal conditioning, while response tokens are produced through iterative parallel denoising. Our key insight is that reliable visual attribution in this setting must be extracted from *structurally valid intermediate multimodal states* along the denoising trajectory, where image-grounded spatial information is still preserved and can be linked to the final prediction through gradients. Based on this principle, we first construct raw activation maps by tracing gradients from the final response back to these valid hidden features. We further find that such raw maps are intrinsically corrupted by stochastic spatial noise, diffuse background responses. To address these challenges, Diffusion-CAM introduces four complementary refinement modules, yielding more localized, faithful, and diffusion-specific visual explanations. Additionally, extensive experiments on COCO Caption (Chen et al., 2015) and GranDf (Rasheed et al., 2024) datasets confirm that Diffusion-CAM achieves superior localization accuracy and background suppression compared to state-of-the-art baselines.

Our main contributions are summarized as follows:

- We articulate the fundamental conflict between the global, parallel nature of emerging dMLLMs and the sequential assumptions of existing explanation methods.
- We propose the first comprehensive interpretation method for dMLLMs, introducing a specialized pipeline—comprising critical-step gradient extraction and dedicated post-processing modules—to achieve precise localization.
- We demonstrate that Diffusion-CAM significantly outperforms state-of-the-art baselines in both localization accuracy and explanation quality across multiple benchmarks, establishing a new standard for interpreting dMLLMs.

2 Related Work

Interpretation for Autoregressive MLLMs. Model interpretability spans mechanistic analysis (Elhage et al., 2021; Olsson et al., 2022; Liao et al., 2025; Nam et al., 2025), attribution methods (Sundararajan et al., 2017; Shrikumar et al., 2017), and visual explanation (Selvaraju et al., 2017; Zhou et al., 2016). With the prominence of autoregressive MLLMs like LLaVA (Liu et al., 2023, 2024), Gemini (Team et al., 2023), and others (Chen et al., 2024; Bai et al., 2023), visual explanation has become vital for tracing token generation. Current techniques range from gradient-based CAMs (Selvaraju et al., 2017; Jiang et al., 2021; Smilkov et al., 2017) and attention mechanisms (Chefer et al., 2021; Abnar and Zuidema, 2020; Montavon et al., 2019) to perturbation methods like LIME and SHAP (Ribeiro et al., 2016; Lundberg and Lee, 2017). While these approaches inherently rely on autoregressive dependencies (Vaswani et al., 2017), rendering them unsuitable for non-sequential architectures.

Challenges in Diffusion Architectures. The shift from AR MLLMs to dMLLMs (You et al., 2025; Ye et al., 2025a; Li et al., 2025a,b,c, 2026) creates a different interpretability setting. Instead of left-to-right decoding, dMLLMs generate responses by iterative masked denoising (Ho et al., 2020; Song et al., 2020) under fixed multimodal conditioning, so the token-level causal structure assumed by conventional CAM-style methods no longer directly applies. As a result, visual attribution must be extracted from intermediate mul-

timodal states that still preserve image-grounded spatial structure during denoising. Meanwhile, existing diffusion-interpretability methods for text-to-image generation, such as DAAM (Tang et al., 2023), don't directly transfer to multimodal reasoning, and diffusion-based global refinement can further amplify noise from feature redundancy, background dispersion, and unstable visual representations (Darcet et al., 2023; He et al., 2016; Li et al., 2025f; Balasubramanian et al., 2024). Our work is motivated by these diffusion-specific challenges.

3 Method

In this section, we first explain how we utilize the common structural and principleal characteristics of dMLLMs to adapt the CAM method to them. Secondly, we expound four modules in Diffusion-CAM specifically designed to improve the clarity and effectiveness of activation maps. Finally, we introduce four metrics designed to enable fine-grained evaluations of the explanations.

3.1 CAM Adaptation for dMLLMs

Conventional gradient-based visual explanation techniques (Selvaraju et al., 2017; Zhang et al., 2024) were primarily developed for autoregressive vision-language models (Liu et al., 2023; Bai et al., 2023). In autoregressive architectures, the model predicts tokens sequentially, with each token t_i conditioned on all preceding tokens: $p(t_i \mid t_{<i}, \mathbf{I})$. This left-to-right factorization yields a clear gradient path from the target output to the visual features, making CAM computation relatively direct.

Recent dMLLMs (Li et al., 2025c; You et al., 2025; Yang et al., 2025; Li et al., 2025a, 2026, 2025b), despite differing in task scope and implementation details, share a common generative mechanism: image features and textual prompts serve as fixed multimodal conditioning, while the response tokens are generated through *iterative masked denoising* rather than next-token prediction. This changes the attribution problem fundamentally. Instead of tracing gradients through a single autoregressive decoding path, CAM for dMLLMs must identify a denoising step whose intermediate hidden states still preserve the image-conditioned spatial structure required for visual grounding. Accordingly, we adapt CAM to this shared conditional masked-diffusion interface rather than to any model-specific decoding heuristic.

Formally, for autoregressive models, CAM computation follows:

$$G_k = \frac{\partial y^c}{\partial A_k} \tag{1}$$

For diffusion models, we instead compute gradients from the final response score to image-grounded features at a *valid conditioning step*:

$$G_c^{(s)} = \frac{\partial \mathcal{L}_{\text{final}}}{\partial A_c^{(s)}}, \qquad \mathcal{L}_{\text{final}} = \sum_{t \in \mathcal{T}} \mathbf{z}_{\text{final}}[t] \tag{2}$$

where $\mathcal{T}$ denotes the selected answer-token indices, and $A_c^{(s)}$ represents the c-th feature channel at denoising step s. In our implementation, the attribution step is *not hard-coded*. Instead, we select it using a dynamic feasibility criterion: a denoising step is valid only if the hooked hidden-state sequence still contains the full image-token span required for image-region extraction. For example, under LaViDa's Prefix-DLM inference (Li et al., 2025c), this criterion is satisfied only at the earliest conditioning step $s = 0$ B; for other dMLLMs, the same rule naturally extends to whichever step(s) remain structurally valid.

We introduce three adaptations to make CAM compatible with this setting.

(1) Model-aware feature extraction. We register a forward hook at an intermediate transformer block and retain its gradient during backpropagation from the final prediction score. This layer is selected from a layer sweep as a stable trade-off between spatial grounding and multimodal semantic integration. Rather than assuming a fixed timestep a priori, we extract features only from denoising steps that satisfy the above feasibility condition.

(2) Dynamic image-span localization. Because dMLLM hidden states contain mixed multimodal tokens, image features must be localized dynamically rather than assumed to occupy fixed positions. We store multimodal packing metadata in `info4cam`, from which we parse the image-token boundaries. Given the full hidden sequence $\mathbf{F} \in \mathbb{R}^{L \times D}$, we extract image features as:

$$\begin{aligned} \mathbf{A}_{\text{img}} &= \mathbf{F}[\mathcal{I}_{\text{img}}] \in \mathbb{R}^{(H \times W) \times D}, \\ \mathcal{I}_{\text{img}} &= \{i \in \mathbb{N} : N_{\text{base}} \le i < N_{\text{base}} + H \times W\}, \end{aligned} \tag{3}$$

where N_{base} is the base text/prompt offset recovered from `info4cam`. The extracted token sequence is then reshaped to $\mathbb{R}^{D \times H \times W}$ to form the spatial feature map.

(3) Diffusion-CAM generation. On the valid image-region feature map, we apply a Grad-CAM-style aggregation. Specifically, we spatially average the gradients to obtain channel weights

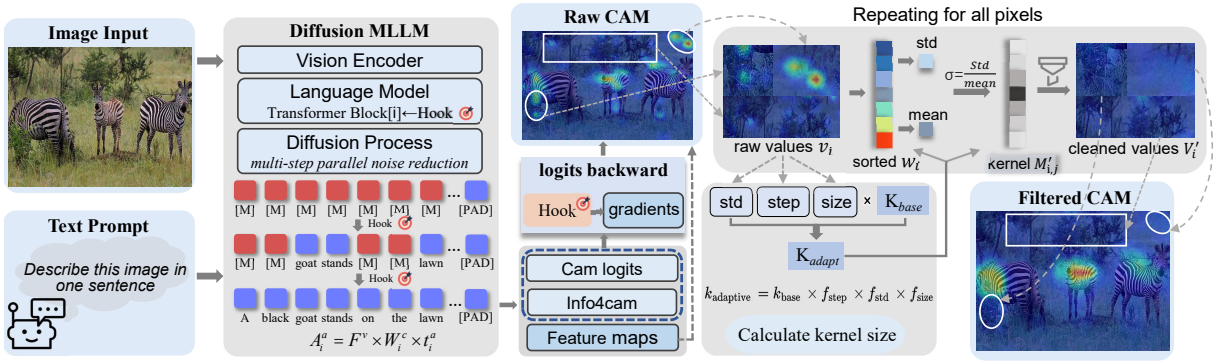

Figure 2: **Illustration of the Raw CAM generation and Adaptive Denoising. (Left) Raw CAM Generation:** We capture intermediate visual features and gradients from the LaViDa backbone during the denoising process via hook registration. **(Right) Adaptive Kernel Denoising:** A dynamic kernel $\mathbf{K}_{adapt}$, calibrated by feature statistics (e.g., σ), drives a Rank Gaussian Filter to suppress stochastic noise and architectural artifacts, transforming noisy activations into precise, semantically coherent heatmaps.

$$w_c = \frac{1}{H \times W} \sum_{i,j} G_c^{(s)}(i,j),$$
$$M_{\text{base}} = \text{ReLU}\left(\sum_c w_c \cdot A_c^{(s)} \right). \tag{4}$$

This yields the baseline CAM before the subsequent refinement modules. In this way, the proposed adaptation is tied to the shared masked-denoising structure of dMLLMs, rather than to the implementation details of a single model.

3.2 Four Modules within Diffusion-CAM

3.2.1 Adaptive Kernel Denoising

Transformer-based multimodal models exhibit high-frequency architectural artifacts in activation maps due to the discrete nature of self-attention. To mitigate this, we introduce a **Adaptive Kernel Denoising** module. This approach synergizes the robustness of order statistics with Gaussian weighting, dynamically calibrating filtering parameters based on diffusion properties.

Dynamic Kernel Determination. Standard fixed-size filters fail to accommodate varying noise profiles across different timesteps. As shown in the right of Figure 2, we propose a dynamic strategy scaling the receptive field k_{adaptive} via three governing factors:

$$k_{\text{adaptive}} = \lfloor k_{\text{base}} \cdot \mathcal{F}_{\text{step}}(S) \cdot \mathcal{F}_{\text{std}}(\sigma_{\mathbf{M}}) \cdot \mathcal{F}_{\text{size}}(H) \rfloor_{\text{odd}} \tag{5}$$

where $\lfloor \cdot \rfloor_{\text{odd}}$ yields the nearest odd integer. $\mathcal{F}_{\text{step}}$ scales with denoising steps S, as longer trajectories produce diffuse semantic patterns requiring larger aggregation windows. $\mathcal{F}_{\text{std}}$ reacts to spatial variance $\sigma_{\mathbf{M}}$, widening the kernel to suppress aggressive peaks in high-noise scenarios, while $\mathcal{F}_{\text{size}}$ adjusts for resolution H to ensure scale invariance.

Rank-Weighted Gaussian Filtering. To remove artifacts without blurring semantic boundaries, we employ a Rank-Weighted Gaussian Filter. Unlike spatial convolution, we sort activation values within a local window $\Omega_{i,j}$ to form an ordered set $\mathcal{V}^{\text{sorted}}$. The filtered value $M'_{i,j}$ is computed as:

$$M'_{i,j} = \sum_{n=1}^{k^2} v_n \cdot \mathcal{G}_n(\mu, \sigma_{rank}) \tag{6}$$

Here, $\mathcal{G}_n$ applies normalized Gaussian weights to the **rank index** n rather than spatial distance. This mechanism effectively suppresses outliers occupying extreme ranks while preserving the dominant semantic signal, offering superior robustness over standard linear filtering.

3.2.2 Distribution-Aware Confidence Gating

Different diffusion steps and image contents manifest varying statistical properties in activation maps. Uniform processing strategies often induce artifacts in high-variance scenarios while failing to denoise low-variance ones effectively. To mitigate this stochasticity, we propose **Distribution-Aware Confidence Gating (DACG)**, as illustrated in Figure 3(a).

Instead of rigid thresholding, DACG dynamically calibrates the confidence boundary τ_{conf} by analyzing the global activation statistics (mean $\mu_{\mathbf{M}}$ and standard deviation $\sigma_{\mathbf{M}}$). We formulate a dynamic mapping function $\mathcal{F}(\cdot)$ that determines the optimal quantile α for filtering:

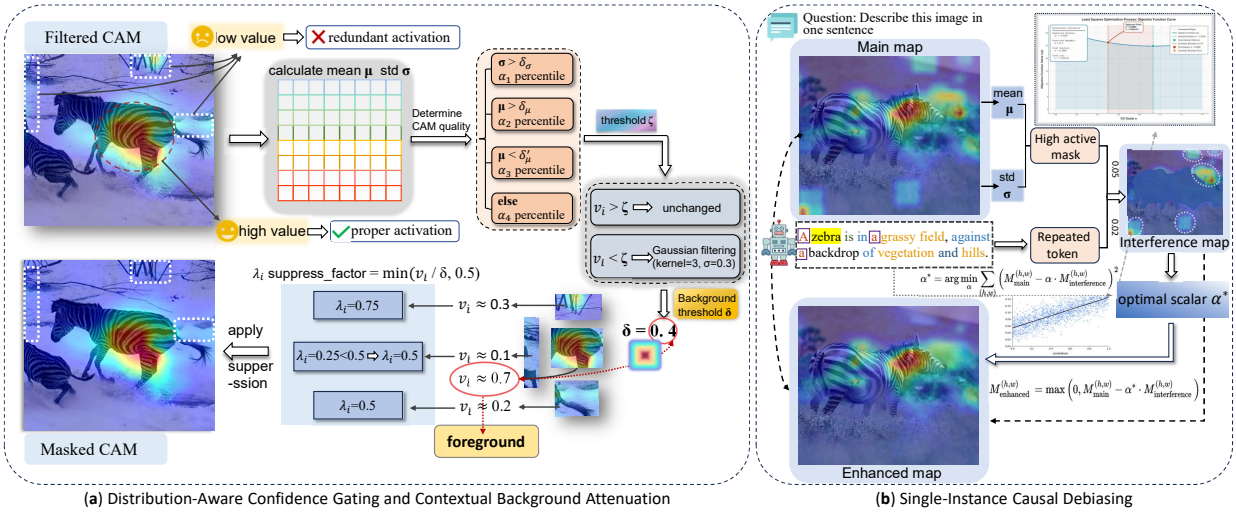

Figure 3: Illustration of the other three modules of Diffusion-CAM. **Distribution-Aware Confidence Gating:** We calculate an adaptive threshold and process regions differently. **Contextual Background Attenuation:** identifies the foreground and background and calculates suppression factors. **Single-Instance Causal Debiasing:** uses repeated tokens and abnormally high activation mask to make heatmap clearer.

$$\begin{aligned} \tau_{\text{conf}} &= \text{Quantile}(\mathbf{M}, \alpha), \\ \alpha &= \mathcal{F}(\mu_{\mathbf{M}}, \sigma_{\mathbf{M}}) \end{aligned} \tag{7}$$

Specifically, α adaptively shifts between stricter bounds (e.g., 90th percentile) for high-variance distributions ($\sigma_{\mathbf{M}} > \delta_\sigma$) to suppress noise, and relaxed bounds (e.g., 75th percentile) for high-brightness maps ($\mu_{\mathbf{M}} > \delta_\mu$) to preserve signals. This statistical alignment ensures robustness across diverse generation scenarios.

Crucially, we implement a **selective processing mechanism**. Rather than applying uniform denoising, we partition the activation map into high-confidence ($\mathbf{M}_{hc}$) and low-confidence ($\mathbf{M}_{lc}$) zones based on τ_{conf}:

$$\begin{aligned} \mathbf{M}_{hc} &= \{(i,j) \mid M(i,j) > \tau_{\text{conf}}\} \\ \mathbf{M}_{lc} &= \{(i,j) \mid M(i,j) \leq \tau_{\text{conf}}\} \end{aligned} \tag{8}$$

Then we apply mild Gaussian Denoising (kernel=3, σ=0.3) to $\mathbf{M}_{lc}$ to reduce salt-and-pepper noises, while $\mathbf{M}_{hc}$ retains its raw structural integrity. This dual-path strategy effectively disentangles semantic structures from background noise without over-smoothing salient regions.

3.2.3 Contextual Background Attenuation

In dMLLMs, background regions often accumulate residual signals from multiple denoising steps, leading to boundary blurring. To address this, as shown in Figure 3(a), we introduce **Contextual Background Attenuation (CBA)**, which leverages a multi-scale statistical ensemble to define a robust separation boundary.

We construct a composite background threshold τ_{bg} by aggregating complementary statistical descriptors:

$$\tau_{\text{bg}} = \sum_k w_k \cdot \mathcal{S}_k(\mathbf{M}) \tag{9}$$

where $\mathcal{S}(\mathbf{M})$ represents a vector of statistical metrics including the median of non-zero activations, global distribution quantiles, and peak activation references. For example, let $\mathbf{M} \in \mathbb{R}^{H \times W}_{\geq 0}$ be the normalized CAM map, and $\mathbf{M}_{>0} = \{\mathbf{M}_{i,j} \mid \mathbf{M}_{i,j} > 0\}$. We define the descriptor vector:

$$\mathcal{S}(\mathbf{M}) = [S_1, S_2, S_3, S_4], \tag{10}$$

with $\mathcal{S}_1 = \text{median}(\mathbf{M}_{>0}), \mathcal{S}_2 = Q_{0.60}(\mathbf{M}), \mathcal{S}_3 = \mathbb{E}[\mathbf{M}_{>0}], \mathcal{S}_4 = \max(\mathbf{M})$.

The weights w_k balance the contribution of conservative estimation and strong signal anchoring, ensuring the threshold adapts to the image's specific contrast profile.

For identified background regions $\mathcal{B} = \{(i,j) \mid \mathbf{M}_{i,j} < \tau_{\text{bg}}\}$, we employ a **progressive soft-attenuation** function instead of hard truncation:

$$\mathbf{M}'_{i,j} = \mathbf{M}_{i,j} \cdot \max\left(\gamma, \frac{\mathbf{M}_{i,j}}{\tau_{\text{bg}}}\right), \quad \forall (i,j) \in \mathcal{B} \tag{11}$$

Here, γ (set to 0.5) serves as a retention lower bound. This soft-masking approach creates a smooth gradient transition, effectively suppressing background noise while preventing the introduction of artificial hard edges common in power-law suppression, thereby enhancing the foreground-background contrast ratio naturally.

3.2.4 Single-Instance Causal Debiasing

Interference refers to activation in the heatmap that is weakly related to the intended visual evidence,such as from tokens unrelated to the key answers and from prompt tokens. Standard ECI (Li et al., 2025e) relies on cross-image frequency analysis to identify confounders, rendering it inapplicable to the single-instance inference paradigm of dMLLMs. To bridge this gap, we propose **Single-Instance Causal Debiasing (SICD)**, grounded in the *Hypothesis of Linguistic Economy*.

We observe that human descriptions typically employ referential grouping for semantic entities (e.g., "two sheep" rather than "a sheep... a sheep"), whereas repetition is predominantly exhibited by **syntactic function words** (e.g., "a", "the"). Furthermore, the phenomenon of language economy has also been verified in some studies (Meister et al., 2021; Gwak et al., 2025; Hao and Kaiser, 2025; Jaeger and Levy, 2006). For instance, in daily communication, humans have many ways to express the same meaning. Rational speakers will subconsciously follow the Uniform Information Density (UID) assumption when organizing sentences, and try to avoid *peaks* and *valleys* of information density in the sentences, such as omitting the relative pronoun "that". In the VQA visualization activation maps for MLLMs, unlike semantic nouns that ground specific objects, these high-frequency functional tokens generate diffuse, non-specific activations that act as background noise. Consequently, as shown in Figure 3(b), we reformulate interference $\mathcal{I}$ by targeting these syntactic redundancies alongside statistical anomalies:

$$\mathcal{I} = \omega_{\text{rep}} \cdot \mathbf{M}_{\text{rep}} + \omega_{\text{out}} \cdot \mathbf{M}_{\text{out}} \tag{12}$$

where $\mathbf{M}_{\text{rep}}$ captures the diffuse activations from repeated functional tokens, and $\mathbf{M}_{\text{out}}$ isolates spatial outliers exceeding statistical bounds (e.g., $\mu_{\mathbf{M}} + 2\sigma_{\mathbf{M}}$). ω_{rep} and ω_{out} are balancing coefficients.

To determine the optimal intervention strength λ^*, we employ a constrained least-squares objective:

$$\lambda^* = \arg \min_{\lambda \in [0, \lambda_{\max}]} \|\mathbf{M} - \lambda \cdot \mathcal{I}\|_F^2 \tag{13}$$

The final corrected map is obtained via $\mathbf{M}_{\text{clean}} = \text{ReLU}(\mathbf{M} - \lambda^* \cdot \mathcal{I})$. This module operates as a conditional gate, triggering only when the map exhibits pathological traits (high variance/skewness), thus preserving the integrity of non-repetitive, semantically rich descriptions.

3.3 Evaluation Metrics

We propose a three-metric evaluation framework tailored for diffusion models' smooth, gradually distributed activations (Li et al., 2025e).

Target Localization (Obj-IoU): Measures spatial overlap between predicted regions $\mathcal{P}$ (obtained via Otsu thresholding) and ground truth $\mathcal{G}_c$ for object class c:

$$\text{Obj-IoU} = \max_c \frac{|\mathcal{P} \cap \mathcal{G}_c|}{|\mathcal{P} \cup \mathcal{G}_c|} \tag{14}$$

Foreground-Background Contrast: Quantifies background suppression effectiveness, where M_{fg} and M_{bg} denote mean activations within and outside ground truth regions:

$$R_{\text{c}} = \frac{\mathbb{E}[M_{\text{fg}}]}{\mathbb{E}[M_{\text{bg}}] + \epsilon} \tag{15}$$

Target Activation Concentration: Measures activation focus on target regions, where $\mathcal{G}$ represents ground truth pixel locations:

$$C_{\text{t}} = \frac{\sum_{(i,j) \in \mathcal{G}} M(i,j)}{\sum_{(i,j)} M(i,j) + \epsilon} \tag{16}$$

F3-Score: We employ harmonic mean to ensure balanced performance, as it is more sensitive to lower values and penalizes imbalanced development:

$$\text{F3-Score} = 3 \Big/ \left(\frac{1}{\text{Obj-IoU}} + \frac{1}{\min\left(\frac{R_{\text{c}}}{20}, 1\right)} + \frac{1}{C_{\text{t}}} \right) \tag{17}$$

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our approach on datasets featuring both textual descriptions and pixel-level annotations. Our primary benchmark is the COCO Caption dataset (40,504 images) (Chen et al., 2015). Since interpretability methods only require inference, we directly utilize the validation set. All images include segmentation masks for quantitative evaluation. We additionally report results on the GranDf dataset (1,000 images) (Rasheed et al., 2024), which provides fine-grained object localization annotations. Both datasets employ manually annotated masks to ensure reliability.

Implementation Details Our experiments were conducted on an H20 GPU. The baseline LLaVA-CAM (Zhang et al., 2024) employs a fixed threshold of 0.4 for feature selection, while our approach omits this step. The hyperparameters δ_σ, δ_μ,δ'_σ in

Method	COCO Caption				GranDf			
	Obj-IoU(%)	Contrast	Concen.	F3-Score	Obj-IoU(%)	Contrast	Concen.	F3-Score
LLaVA-CAM	20.02	2.18×	41.22	18.08	18.16	1.41×	70.80	14.22
Grad-CAM	19.93	2.04×	41.91	17.43	17.82	1.48×	75.42	14.67
TAM	15.21	2.51×	40.10	17.61	20.39	1.36×	67.31	14.23
Diffusion-CAM (Ours)	**30.10**	**2.58×**	**51.41**	**23.04**	**28.41**	**2.02×**	**86.14**	**20.53**

Table 1: Comprehensive comparison with state-of-the-art methods on COCO Caption (Chen et al., 2015) and GranDf datasets (Rasheed et al., 2024), and the best baseline results are underlined. The F3-Score (%) reflects the overall performance, calculated as the harmonic mean of the three core metrics.

Config.	Obj-IoU(%)	Contrast	Concen.	F3-Score
Baseline	20.12	2.19×	41.42	18.16
+ Denoising	25.85	2.28×	44.01	20.16
+ Gating	21.22	2.20×	47.94	18.88
+ Attenuation	21.42	2.41×	42.63	19.59
+ Debiasing	24.11	2.29×	44.31	19.82
Full	**30.10**	**2.58×**	**51.44**	**23.04**

Table 2: Ablation study on COCO Caption (Chen et al., 2015). Each row toggles one module on top of the baseline.

Method	Layer	Obj-IoU(%)	Contrast	F3-Score
LLaVA-CAM	10	20.02	2.18×	18.07
	20	13.01	1.44×	11.51
	25	16.09	1.19×	10.56
	29	17.16	1.00×	10.04
Diffusion-CAM	10	30.10	2.58×	23.04
	20	29.62	1.58×	15.44
	25	22.37	1.22×	12.23
	29	19.08	1.03×	10.86

Table 3: Performance comparison of Diffusion-CAM and LLaVA-CAM across different target layers on COCO Caption (Chen et al., 2015).

DACG are set to 0.22, 0.35 and 0.25, which are derived from empirical observation of the activation distribution. For IoU computation and performance evaluation, we tag each word's part of speech using the pos_tag function from the NLTK Python package and employ standard Otsu threshold binarization. Furthermore, our results are the average values obtained from multiple experiments.

4.2 Quantitative Results

Ablation Studies and Effectiveness Analysis Table 2 reports the results of COCO Caption (Chen et al., 2015) ablation experiments with only one Diffusion-CAM module activated on the baseline model. Specifically, the *Adaptive Kernel Denoising* alone lifts F3-Score by 2.0% through suppressing salt-and-pepper noise; *Distribution-Aware Confidence Gating* raises concentration from 41.42% to 47.94% through selective refinement and the *Contextual Background Attenuation* increases foreground–background contrast from 2.19× to 2.41×, while preserving target integrity. Using *Causal Debiasing* is enabled the Obj-IoU reaches 24.11%. Results illustrate each module can make different improvements to the three metrics. Although evaluated independently, the modules are complementary: when all are activated simultaneously the F3-Score rises by 4.88%.

Comparison with SoTA Methods Table 1 presents comprehensive comparisons between Diffusion-CAM and existing methods: LLaVA-CAM (Zhang et al., 2024), TAM (Li et al., 2025e) and Grad-CAM (Selvaraju et al., 2017). Our approach surpasses all competitors in all four metrics.

The comparative analysis demonstrates the superiority of our method across different datasets and evaluation metrics. On COCO Caption dataset (Chen et al., 2015), Diffusion-CAM achieves an F3-Score of 23.04%, representing a 4.96% improvement over the strongest baseline LLaVA-CAM (Zhang et al., 2024). The 30.1% Obj-IoU indicates exceptional target localization capability, while the 2.58× contrast ratio and 51.4% concentration validate effective background suppression and activation focusing. On the more challenging GranDf (Rasheed et al., 2024) benchmark, our method maintains superior performance with an F3-Score of 20.53%, exceeding Grad-CAM (Selvaraju et al., 2017) by 5.86%. These results confirm the robustness and generalization capability of our method across varying dataset complexities.

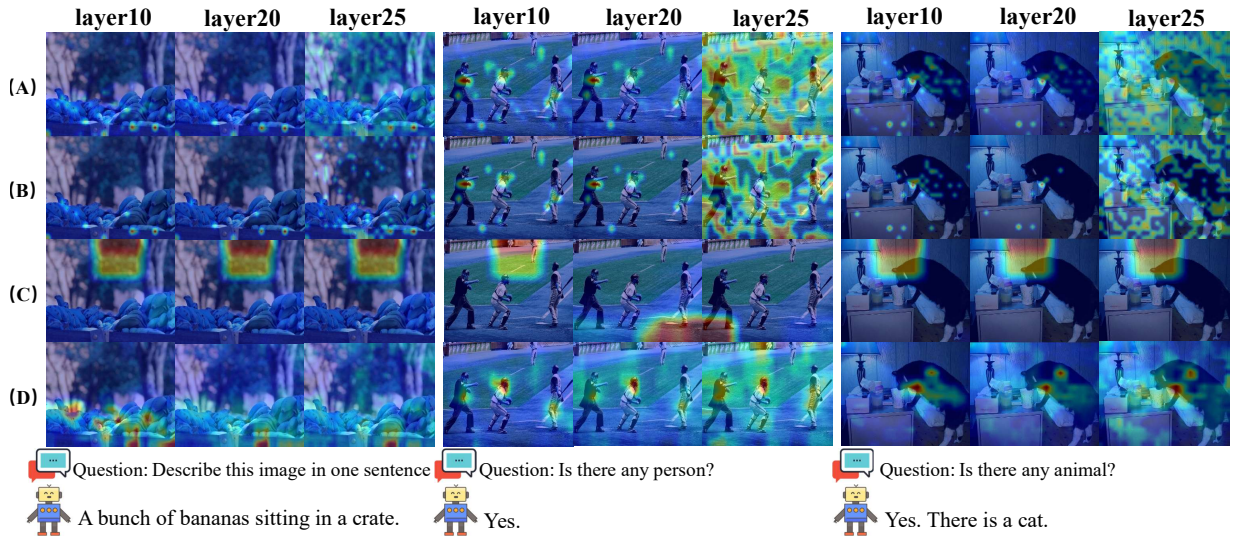

Figure 4: Visual comparison between our method (D) and SoTA approaches including (A) [Grad-CAM], (B) [LLaVA-CAM], and (C) [TAM] at different Transformer layers. Diffusion-CAM produces more precise activation maps compared to the baselines, exhibiting robust localization ability across all visualized layers.

Applicability and Scalability Diffusion-CAM exhibits broad applicability through its design leveraging core properties of diffusion architectures, making it applicable to various dMLLMs. Notably, as shown in Table 3, our method outperforms method LLaVA-CAM (Zhang et al., 2024)(which gets the highest score on layer 10 in the SoTA methods) across different target layers. It demonstrates the applicability of our method to different transformer layers. Simultaneously, the framework demonstrates excellent scalability through four independent modules that can be selectively applied to meet different scenario requirements.

4.3 Qualitative Results

Localization Performance Figure 4 and Figure 6 demonstrate the superior localization capability of our method across diverse scenarios. Diffusion-CAM produces more focused and clear activation maps that effectively highlight target objects while suppressing background noise.

Our approach consistently exhibits excellent localization performance across various contexts. In multi-object scenes, Diffusion-CAM successfully focuses on the most relevant regions while maintaining global context awareness. The activation maps display significantly reduced attention to uninformative background areas and substantially lower noise levels, with improved emphasis on target objects compared to baseline methods, demonstrate enhanced capability to distinguish between relevant and irrelevant regions.

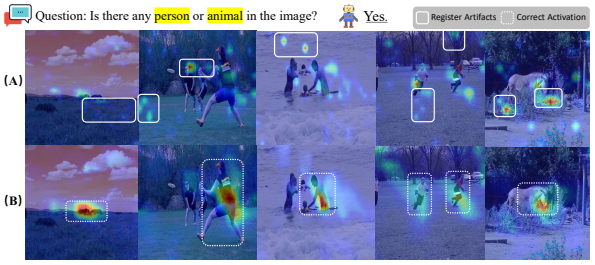

Figure 5: Visual comparison regarding artifacts elimination. (A) [Original CAM], (B) [Diffusion-CAM].

Visualization Comparison Across Target Layers Figure 4 presents CAM visualizations generated by our method across different target layers alongside comparative baseline results. Experimental findings indicate that our approach significantly outperforms baseline methods at all tested target layers. At layer[10], our method produces the most precise object localization, with activation regions more concentrated on target objects and minimal activation in irrelevant areas.

As layer depth increases, activation patterns gradually become more diffuse because the image tokens receive substantially less attention than prompts and that image-token information flow converges in shallow layers and disperses in deep layers, with "cliff layers" after which image tokens become highly redundant. (Zhang et al., 2024)while our method maintains effective target object localization. Crucially, even at deeper layers (layer[25]), it continues to generate clearer, more focused CAMs than baseline approaches, demon-

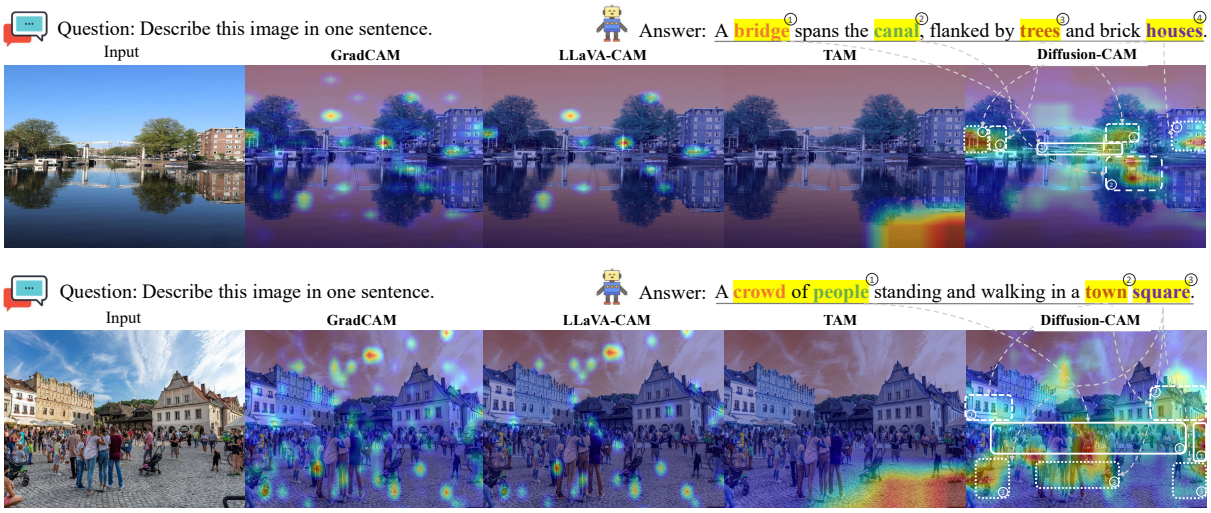

Figure 6: Visual comparison between our method and SoTA methods. Diffusion-CAM demonstrates superior global semantic alignment and also suppresses some noise in the baseline methods

strating robust adaptability to different layer features in diffusion models. This inter-layer consistency further validates that Diffusion-CAM effectively handles the unique global activation characteristics inherent to diffusion architectures.

Suppression of Architectural Artifacts. As illustrated in Figure 5, the original activation maps without any processing frequently exhibit high-magnitude activations in irrelevant background regions, such as the sky or corners. These anomalies align with the "register artifacts" phenomenon inherent to Vision Transformers, where the model repurposes low-information background tokens to store global context (Darcet et al., 2023). These high-norm outliers manifests as false-positive noise in gradient-based explanations. In contrast, by incorporating our specialized modules, Diffusion-CAM effectively eliminates these high-frequency spikes. The resulting heatmaps demonstrate a clean background with activation concentrated solely on the semantic target, confirming our framework's robustness in mitigating the architectural artifacts of the underlying ViT encoder.

Global Characteristics of Diffusion Models Figure 6 illustrates the global characteristics of dMLLMs. The visualization reveals that, when describing landscapes or scenes without a specific target, Diffusion-CAM can more accurately focus on key tokens (such as nouns) in the response. Our method successfully enhances the quality of these global activations through targeted interventions, demonstrating the effectiveness of our global enhancement strategy.

5 Conclusion

We presented **Diffusion-CAM**, the first interpretability framework that bridges the explanatory gap for dMLLMs. By extracting features from the critical conditioning step and employing Adaptive Kernel Denoising, Single-Instance Causal Debiasing and other post-processing modules, our method showcases high-quality visualization results. Comprehensive evaluations confirm its superiority over autoregressive baselines in both localization accuracy and noise suppression. This work establishes a rigorous foundation for interpreting dMLLMs, paving the way for future more in-depth and comprehensive research.

6 Limitations

Despite the promising performance of Diffusion-CAM in interpreting dMLLMs, we acknowledge several limitations that point towards future research directions:

Abstract Concept Localization. Second, while Diffusion-CAM demonstrates superior precision in grounding concrete physical objects, it exhibits limitations when interpreting abstract or emotional concepts. For prompts involving intangible attributes (e.g., love," guilt," or "atmosphere"), the model often diffuses attention across the entire scene rather than converging on specific visual regions. This reflects an inherent *spatial-semantic gap*: abstract concepts in MLLMs may rely on global holistic representations that lack distinct spatial coordinates, making them difficult to visualize through strictly gradient-based localization maps.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2024. Decomposing and interpreting image representations via text in vits beyond clip. *Advances in Neural Information Processing Systems*, 37:81046–81076.

Shuochen Chang, Xiaofeng Zhang, Qingyang Liu, and Li Niu. 2025. D3-tom: Decider-guided dynamic token merging for accelerating diffusion mllms. *arXiv preprint arXiv:2511.12280*.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Townim Faisal Chowdhury, Kewen Liao, Vu Minh Hieu Phan, Minh-Son To, Yutong Xie, Kevin Hung, David Ross, Anton Van Den Hengel, Johan W Verjans, and Zhibin Liao. 2024. Cape: Cam as a probabilistic ensemble for enhanced dnn interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11072–11081.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Nelson Elhage and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. *arXiv preprint arXiv:2403.00824*.

Minju Gwak, Guijin Son, and Jaehyung Kim. 2025. Revisiting the uniform information density hypothesis in llm reasoning traces. *arXiv preprint arXiv:2510.06953*.

Hailin Hao and Elsi Kaiser. 2025. Uniform information density and syntactic reduction: Revisiting* that*-mentioning in english complement clauses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21980–21994.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30:5875–5888.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. 2025a. Lavidao: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. 2025b. Sparse-lavida: Sparse multimodal discrete diffusion language models. *arXiv preprint arXiv:2512.14008*.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025c. Lavida: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*.

Shufan Li, Yuchen Zhu, Jiuxiang Gu, Kangning Liu, Zhe Lin, Yongxin Chen, Molei Tao, Aditya Grover, and Jason Kuen. 2026. Lavida-r1: Advancing reasoning for unified multimodal diffusion language models. *arXiv preprint arXiv:2602.14147*.

Xingjian Li, Qiming Zhao, Neelesh Bisht, Mostofa Rafid Uddin, Jin Yu Kim, Bryan Zhang, and Min Xu. 2025d. Diffcam: Data-driven saliency maps by capturing feature differences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10327–10337.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. 2025e. Token activation map to visually explain multimodal llms. *arXiv preprint arXiv:2506.23270*.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. 2025f. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409.

Chonghua Liao, Ruobing Xie, Xingwu Sun, Haowen Sun, and Zhanhui Kang. 2025. Exploring forgetting in large language model pre-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2112–2127, Vienna, Austria. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 963–980.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on nlp. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 3078–3105.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. 2025. Causal head gating: A framework for interpreting roles of attention heads in transformers. *arXiv preprint arXiv:2505.13737*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Catherine Olsson and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. 2019. Smooth grad-cam++: An enhanced inference level visualization technique. *arXiv preprint arXiv:1908.01224*.

Xu Pan, Aaron Philip, Ziqian Xie, and Odelia Schwartz. 2024. Dissecting query-key interaction in vision transformers. *arXiv preprint arXiv:2405.14880*.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. 2023. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, and 1 others. 2024a. Knowledge mechanisms in large language models: A survey and perspective. *arXiv preprint arXiv:2407.15017*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jinfeng Wei and Xiaofeng Zhang. 2024. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*.

Jiacheng Ye, Shansan Gong, Jiahui Gao, Junming Fan, Shuang Wu, Wei Bi, Haoli Bai, Lifeng Shang, and Lingpeng Kong. 2025a. Dream-vl & dream-vla: Open vision-language and vision-language-action models with diffusion language model backbone. *arXiv preprint arXiv:2512.22615*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025b. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025a. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. pages 3512–3534.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024. From redundancy to relevance: Information flow in lvlms across reasoning tasks. *arXiv preprint arXiv:2406.06579*.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and 1 others. 2026. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*.

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. 2025b. Finer-cam: Spotting the difference reveals finer details for visual explanation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9611–9620.

Qiyan Zhao, Xiaofeng Zhang, Shuochen Chang, Qianyu Chen, Xiaosong Yuan, Xuhang Chen, Luoqi Liu, Jiajun Zhang, Xu-Yao Zhang, and Da-Han Wang. 2026. Context tokens are anchors: Understanding the repetition curse in dmllms from an information flow perspective. In *The Fourteenth International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

A Background on CAM-based Interpretability

Class activation mapping (CAM) (Zhou et al., 2016) is a post-hoc visual explanation technique that highlights the image regions most responsible for a target prediction. Let a classifier be denoted by $f : \mathcal{X} \to \mathbb{R}^C$, which maps an input image $\mathbf{x} \in \mathcal{X}$ to class logits $\mathbf{y} \in \mathbb{R}^C$. At a chosen layer, the network produces K feature maps $\mathbf{A} = \{A_k\}_{k=1}^K$ with $A_k \in \mathbb{R}^{H \times W}$. CAM explains class c by forming a weighted combination of these feature maps:

$$L^c = h\left(\sum_k \alpha_k^c A_k\right), \tag{1}$$

where α_k^c measures the contribution of channel k to class c, and $h(\cdot)$ is typically chosen as ReLU to retain positive evidence. In the original CAM formulation, when the classifier is built on global average pooling, the class logit can be written as

$$y^c = \sum_k w_k^c \, \mathrm{GAP}(A_k) + b^c, \tag{2}$$

so the channel importance is directly given by the classifier weight, i.e., $\alpha_k^c = w_k^c$ (Zhou et al., 2016). Grad-CAM (Selvaraju et al., 2017) generalizes this idea to broader architectures by replacing classifier weights with gradient-based importance:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_k^{ij}}, \tag{3}$$

where $Z = H \times W$. Intuitively, CAM-style methods answer two coupled questions: *what* target score should be explained, and *how* intermediate feature channels should be weighted to recover spatial evidence.

Subsequent CAM variants mainly refine one of these two components. Score-CAM (Wang et al., 2020) replaces gradient-based weighting with forward confidence changes to reduce gradient noise; LayerCAM (Jiang et al., 2021) exploits spatially local importance from intermediate layers for finer localization; and Finer-CAM (Zhang et al., 2025b) further changes the explanation target from an isolated class score to a contrastive target, emphasizing discriminative details relative to visually similar concepts. More recently, multimodal explanation methods such as LLaVA-CAM (Zhang et al., 2024) and Token Activation Map (Li et al., 2025e) extend this CAM paradigm to large vision-language models by treating textual outputs or concepts as explanation targets, while Smooth-CAM-style designs (Omeiza et al., 2019) improve stability through perturbation-based smoothing. Taken together, these methods establish CAM as a lightweight and flexible framework for spatially grounded explanation.

B Shared Structural Basis of Diffusion MLLMs and Our CAM Adaptation

Although recent diffusion MLLMs differ in emphasis, for example, LLaDA-V (You et al., 2025) adopts visual instruction tuning, Dream-V (Ye et al., 2025a)L highlights planning-oriented diffusion backbones, and LaViDa-O (Li et al., 2025a) / Sparse-LaViDa (Li et al., 2025b) / LaViDa-R1 (Li et al., 2026) extend the LaViDa (Li et al., 2025c) family toward unified generation, sparse inference, and stronger reasoning, their underlying mechanism is highly consistent. In all these models, images are first converted into visual embeddings or visual tokens through a vision encoder and projector, then combined with textual prompts as **fixed multimodal conditioning**, while the response is generated by **iterative masked denoising**. In other words, despite architectural variations, they all expose the same attribution interface: intermediate multimodal hidden states that support **image-conditioned prediction of masked response tokens**. This shared conditional masked-diffusion structure is the structural basis targeted by our method.

Attribution step	feat_len	img_end	Valid
$t = 0$ (Prefix)	579	551	PASS
$t = S/4$	64	551	FAIL
$t = S/2$	64	551	FAIL
$t = 3S/4$	64	551	FAIL
Aggregate over steps	–	–	N/A

Table 4: Per-step feasibility check for image-span CAM extraction under LaViDa's Prefix-DLM + KV-cached generation. A denoising step is considered valid only if the hooked hidden-state length satisfies `seq_len` $\geq$ `img_end`. Across all logged step-records ($6600 = 200$ images $\times$ (1 prefix $+S$ denoising steps)), only the earliest conditioning step passes, yielding a valid ratio of $200/6600 = 3.0\%$.

To make the model-aware extraction rule explicit, we report the per-step feasibility check

Attribution step	Valid	Contrast ↑	Concen. ↑	Obj-IoU ↑	F3-score ↑
Prefix ($t = 0$)	200/200	2.549 ± 0.126	50.69 ± 1.741	0.307 ± 0.024	0.234 ± 0.018
$t = S/4$	0/200	–	–	–	–
$t = S/2$	0/200	–	–	–	–
$t = 3S/4$	0/200	–	–	–	–
Aggregate over steps	–	–	–	–	–

Table 5: CAM quality at the valid attribution step. Since later denoising steps are infeasible for image-span CAM extraction under Prefix-DLM + KV cache, metrics are reported only for the valid prefix step.

and the CAM quality at the valid step under LaViDa's (Li et al., 2025c) Prefix-DLM + KV-cached generation. As shown in Table 4, under Prefix-DLM + KV-cached generation only the earliest conditioning step remains structurally valid for image-span CAM extraction. We therefore report CAM quality only at this valid step in Table 5.

This rule explains why LaViDa (Li et al., 2025c) typically yields only the earliest conditioning step as valid under Prefix-DLM + KV cache, while models that expose the full multimodal sequence at more denoising steps can admit multiple valid extraction steps. Therefore, our method is applicable to diffusion MLLMs in general: it relies only on the common masked-denoising mechanism and on the availability of at least one structurally valid intermediate multimodal state, both of which are shared across current dMLLM families.

C Additional Technical Analyses

C.1 Sensitivity Analysis of Distribution-Aware Confidence Gating

The proposed Distribution-Aware Confidence Gating (DACG) introduces several threshold-like parameters, including the branch-boundary values and branch-specific percentile settings. To evaluate whether the method is overly sensitive to these choices, we perform two targeted sensitivity studies on 200 COCO images with the prompt *"Describe this image in one sentence."* We report the same metrics as in the main paper, namely Contrast, Concentration, Obj-IoU, and F3-score, together with the branch routing ratios.

Branch-boundary sensitivity. We first vary the three branch-boundary parameters one at a time while keeping the remaining settings fixed at their default values. Specifically, the baseline setting is $\delta_\sigma = 0.22$, $\delta_\mu = 0.35$, and $\delta'_\mu = 0.25$, with branch percentiles $\alpha_1 = 90$, $\alpha_2 = 75$, $\alpha_3 = 80$, and $\alpha_4 = 85$. Tables 6, 7, and 8 report the detailed sweep results. Across all three sweeps, the performance varies smoothly without catastrophic degradation. For example, Obj-IoU changes within a narrow range for each sweep (0.201–0.215 for δ_σ, 0.199–0.215 for δ_μ, and 0.198–0.215 for δ'_μ), while F3-score remains within 0.192–0.196 throughout. We also observe that the branch routing ratios shift gradually and always sum to approximately 100%, confirming that each sample is deterministically assigned to exactly one branch. These results indicate that DACG is not brittle with respect to modest perturbations of its boundary parameters.

Percentile sensitivity. We then evaluate the sensitivity to the percentile parameters while fixing the branch boundaries to their baseline values. Table 9 summarizes the results. The percentile choices are likewise stable: Obj-IoU varies only from 0.197 to 0.215, and F3-score remains within 0.192–0.196 across the tested configurations. Importantly, in this setting the branch routing ratios remain unchanged because the partition boundaries are fixed; only the within-branch processing strength changes. This further supports our claim that DACG acts as a mild confidence-aware refinement rather than a brittle thresholding mechanism. Overall, the sensitivity study shows that DACG improves the baseline in a stable manner, and that its effectiveness does not hinge on a single narrowly tuned hyperparameter choice.

C.2 Efficiency and Computational Overhead

Since our framework contains multiple post-processing modules, an important practical question is whether these refinements substantially increase computational overhead. To answer this, we measure both end-to-end latency and per-module runtime on 200 images using an NVIDIA H20 GPU. We also report the peak GPU memory footprint. The key observation is that the dominant cost comes from the attribution pass itself, i.e., gener-

δ_σ	Obj-IoU ↑	Contrast ↑	Concen. ↑	F3-Score ↑	high_var	high_mean	low_mean	default
0.18	0.201 ± 0.021	2.29 ± 0.13	41.89 ± 2.044	0.192 ± 0.013	36.4%	14.1%	47.2%	3.3%
0.20	0.208 ± 0.017	2.38 ± 0.08	42.77 ± 2.041	0.194 ± 0.012	27.8%	16.6%	50.8%	4.8%
0.22	$\mathbf{0.215 \pm 0.015}$	$\mathbf{2.44 \pm 0.12}$	$\mathbf{43.21 \pm 2.037}$	$\mathbf{0.196 \pm 0.011}$	22.1%	19.3%	53.4%	5.2%
0.24	0.206 ± 0.019	2.39 ± 0.12	42.69 ± 2.039	0.193 ± 0.015	18.6%	21.0%	54.9%	5.6%
0.26	0.202 ± 0.025	2.30 ± 0.14	42.01 ± 2.049	0.192 ± 0.011	15.1%	22.3%	56.4%	6.2%

Table 6: δ_σ sweep for DACG branch-boundary sensitivity. (Baseline values are $\delta_\sigma = 0.22$, $\delta_\mu = 0.35$, $\delta'_\mu = 0.25$, with $\alpha_1 = 90$, $\alpha_2 = 75$, $\alpha_3 = 80$, and $\alpha_4 = 85$.)

δ_μ	Obj-IoU ↑	Contrast ↑	Concen. ↑	F3-Score ↑	high_var	high_mean	low_mean	default
0.28	0.199 ± 0.024	2.30 ± 0.15	41.75 ± 2.121	0.192 ± 0.037	14.7%	37.2%	45.2%	2.9%
0.32	0.205 ± 0.022	2.36 ± 0.18	42.41 ± 2.132	0.194 ± 0.034	19.2%	26.4%	49.8%	4.6%
0.35	$\mathbf{0.215 \pm 0.015}$	$\mathbf{2.44 \pm 0.12}$	$\mathbf{43.21 \pm 2.037}$	$\mathbf{0.196 \pm 0.011}$	22.1%	19.3%	53.4%	5.2%
0.38	0.207 ± 0.027	2.35 ± 0.19	42.60 ± 2.054	0.194 ± 0.031	25.6%	11.8%	57.2%	5.4%
0.42	0.201 ± 0.025	2.32 ± 0.19	42.01 ± 2.173	0.193 ± 0.039	29.3%	3.1%	61.7%	5.9%

Table 7: δ_μ sweep for DACG branch-boundary sensitivity.

δ'_μ	Obj-IoU ↑	Contrast ↑	Concen. ↑	F3-Score ↑	high_var	high_mean	low_mean	default
0.20	0.198 ± 0.025	2.28 ± 0.22	40.81 ± 2.201	0.192 ± 0.012	26.0%	23.5%	43.6%	6.9%
0.23	0.204 ± 0.023	2.31 ± 0.19	42.07 ± 2.147	0.194 ± 0.013	23.4%	20.7%	49.9%	6.0%
0.25	$\mathbf{0.215 \pm 0.015}$	$\mathbf{2.44 \pm 0.12}$	$\mathbf{43.21 \pm 2.037}$	$\mathbf{0.196 \pm 0.011}$	22.1%	19.3%	53.4%	5.2%
0.28	0.209 ± 0.021	2.39 ± 0.17	42.61 ± 2.094	0.195 ± 0.015	19.2%	16.8%	60.2%	3.8%
0.30	0.202 ± 0.031	2.25 ± 0.24	41.98 ± 2.188	0.193 ± 0.013	14.9%	13.4%	69.3%	2.4%

Table 8: δ'_μ sweep for DACG branch-boundary sensitivity.

Setting	Obj-IoU ↑	Contrast ↑	Concen. ↑	F3-Score ↑	high_var	high_mean	low_mean	default
Dynamic (ours)	$\mathbf{0.215 \pm 0.015}$	$\mathbf{2.44 \pm 0.12}$	$\mathbf{43.21 \pm 2.037}$	$\mathbf{0.196 \pm 0.011}$	22.1%	19.3%	53.4%	5.2%
All $\alpha - 5$	0.201 ± 0.022	2.28 ± 0.26	41.77 ± 2.111	0.193 ± 0.008	22.1%	19.3%	53.4%	5.2%
All $\alpha + 5$	0.203 ± 0.017	2.32 ± 0.28	41.62 ± 2.102	0.193 ± 0.013	22.1%	19.3%	53.4%	5.2%
Fixed $\alpha = 85$	0.197 ± 0.019	2.23 ± 0.23	41.05 ± 1.966	0.192 ± 0.009	22.1%	19.3%	53.4%	5.2%

Table 9: Percentile sensitivity of DACG with branch boundaries fixed.

Stage	Time (ms)	Device	Complexity
Base attribution pass	4880.8	GPU	-
AKD	12.7	CPU	$O(HW \cdot k^2)$
DACG	6.9	CPU	$O(HW)$
CBA	5.6	CPU	$O(HW)$
SICD	10.2	CPU	$O(HW)$
Total post-processing	**35.4**	CPU	$O(HW)$

Table 10: Per-module runtime breakdown. The refinement modules are CPU-only and add negligible overhead relative to the attribution pass.

ation followed by the contrastive backward pass required for gradient-based CAM extraction. In contrast, all refinement modules operate only on the final 2D activation map and simple map statistics, and therefore do not introduce any additional model forward or backward passes. Table 11 reports the end-to-end runtime of the base Diffusion-CAM and its module variants, while Table 10 provides a stage-wise breakdown.

Quantitatively, the full pipeline requires 4914.4 ms per image, compared with 4880.8 ms for the base attribution pass, which means that all post-processing modules together add only 35.4 ms, i.e.,

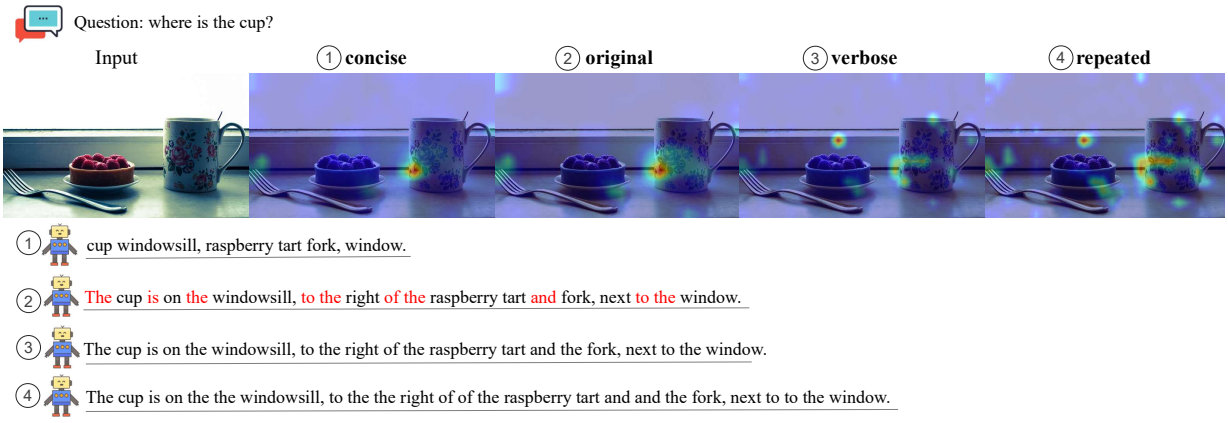

Figure 7: Visualization results of four caption variants under the same teacher-forced attribution protocol by Diffusion-CAM. It is evident that, following the addition of redundant function words, regions in the image unrelated to the answer keywords (e.g., "cup") began to exhibit high-intensity activation; furthermore, as the number of function words increased, these irrelevant activations became increasingly pronounced.

Method	Time (ms)	Peak Mem (GB)	F3-score	Notes
Diffusion-CAM (base)	4880.8	83.21	0.188	generate + contrastive backward
Base + AKD	4893.8	83.21	0.207	+ rank Gaussian filter
Base + DACG	4887.0	83.21	0.187	+ confidence gating
Base + CBA	4886.3	83.21	0.190	+ background attenuation
Base + SICD	4891.1	83.21	0.195	+ causal debiasing
Diffusion-CAM (full)	4914.4	83.21	0.225	all modules enabled

Table 11: End-to-end computational overhead per image on 200 images (NVIDIA H20).

Variant	Contrast ↑	Concen. ↑	Obj-IoU ↑	F3-score ↑
concise	2.3546 ± 0.1133	43.56 ± 1.968	0.2342 ± 0.0153	0.1942 ± 0.0071
original	2.1905 ± 0.1219	41.31 ± 2.029	0.2086 ± 0.0166	0.1846 ± 0.0085
verbose	2.1266 ± 0.1286	39.17 ± 2.031	0.1969 ± 0.0155	0.1829 ± 0.0087
repeated	2.0985 ± 0.1300	39.01 ± 2.027	0.1908 ± 0.0138	0.1811 ± 0.0092

Table 12: Controlled validation of the linguistic economy hypothesis on 200 COCO images. We compare four caption variants under the same teacher-forced attribution protocol. Increasing function-word redundancy leads to more diffuse and less object-specific CAMs.

0.7% of the total runtime. Moreover, the peak GPU memory remains unchanged at 83.21 GB across all variants, since the additional modules run on CPU and do not increase the memory footprint of the model itself. The per-module breakdown further shows that each individual module accounts for only a small fraction of total time. These results indicate that the proposed refinements provide improved explanation quality at negligible extra cost beyond the original attribution pass. We therefore view the framework as a practical interpretability tool whose main computational bottleneck remains the gradient-based CAM extraction itself, rather than the subsequent refinement modules.

D Linguistic Economy Hypothesis and Controlled Validation

Our Single-Instance Causal Debiasing module is motivated by the observation that redundant syntactic scaffolding and repeated function words often induce diffuse, non-specific activations in multimodal heatmaps. We emphasize that, in this paper, we do not use “linguistic economy” as a universal linguistic claim. Instead, we adopt it as an empir-

ical hypothesis about model explanations: when semantically equivalent captions are expressed with increasingly redundant function-word structure, the resulting CAMs tend to become less concentrated and less object-specific.

To test this hypothesis in a controlled manner, we conduct a teacher-forced language intervention study on 200 COCO images. For each image, we construct four caption variants derived from the same base description: *concise* (content words only), *original*, *verbose* (with additional function-word scaffolding), and *repeated* (where function words are duplicated once). We then run the same hooked activation-and-gradient attribution pipeline as in our main method, keeping the image, hook layer, target-score construction, and evaluation metrics fixed across all variants.

Table 12 reports the results: moving from concise to original to verbose to repeated captions consistently reduces Contrast, Concentration, Obj-IoU, and F3-score. In particular, as shown in Figure 7, repeated functional words lead to the most diffuse and least localized heatmaps. This result provides direct evidence that function-word redundancy can act as an interference source in diffusion-CAM attribution, and therefore supports the design motivation behind the Single-Instance Causal Debiasing module. In summary, this experiment does not attempt to prove a general theory of language production. Rather, it verifies, within our visual explanation setting, that syntactic redundancy is systematically associated with weaker spatial grounding and more diffuse activation patterns.