FoAReD: Foreground-guided Attention Reallocation Decoding for Comprehensive Image Captioning

Anonymous authors
Paper under double-blind review

Abstract

Despite their strong image captioning capabilities, VLMs often fail to capture fine-grained details in complex scenes. We find that this issue is associated with insufficient utilization of visual information during autoregressive decoding, where attention is dominated by redundant background and sink tokens. To address this issue, we propose FoAReD, a training-free image captioning framework that enhances visual information utilization through attention contrastive decoding. Specifically, FoAReD employs foreground-guided attention reallocation to construct a contrastive attention condition. During autoregressive decoding, the contrastive and original attention conditions are contrasted to enhance the generation preference toward informative visual details. Experiments on CompreCap and CaptionQA across multiple VLMs show that FoAReD consistently improves both caption comprehensiveness and accuracy.

1 Introduction

Image captioning is a task bridging computer vision and natural language (Vinyals et al., 2015; Xu et al., 2015; Rennie et al., 2017; Anderson et al., 2018; Johnson et al., 2016). In recent years, large vision-language models (VLMs) (Liu et al., 2023; Dai et al., 2023; Wang et al., 2025; Bai et al., 2025) have advanced image captioning with stronger visual understanding and language generation capabilities. These models can capture rich image semantics and generate fluent natural language descriptions.

Beyond general image semantics, comprehensive image captioning requires models to capture fine-grained visual details in complex scenes (Lu et al., 2025a; Yang et al., 2026). Existing methods mainly improve fine-grained image captioning through training-based or training-free strategies. Training-based methods (Li et al., 2024; 2025; Zeng et al., 2025; Xing et al., 2025) introduce extra supervision or specialized training strategies to enhance image captioning, but require additional training costs. Training-free methods (Xing et al., 2026; Pi et al., 2024; Peng et al., 2025) avoid additional training costs, yet often rely on additional reasoning steps or external visual expert models (Huang et al., 2025; Kirillov et al., 2023; Liu et al., 2024b). Although effective, existing methods improve captioning by introducing external supervision or guidance. This motivates us to investigate whether the visual representations already encoded in VLMs can be more effectively utilized to improve fine-grained captioning.

In VLMs, image captioning is performed through an autoregressive decoding process, where visual tokens contribute to next-token prediction through cross-attention. To examine how effectively VLMs utilize their internal visual information, we analyze attention distribution over visual tokens during caption generation. As illustrated in Figure 1, early layers attend more to foreground regions containing richer details, while deeper layers increasingly allocate attention to redundant background tokens. This phenomenon is consistent with prior studies on attention sink effects (Xiao et al., 2024; Kang et al., 2025). This suggests that the caption generation process may not fully utilize the visual information encoded in VLMs.

Based on this observation, we propose **FoAReD**, a training-free image captioning framework that incorporates **Fo**reground-guided **A**ttention **Re**allocation **D**ecoding. Unlike previous methods, FoAReD improves caption generation by better utilizing the visual information already encoded in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

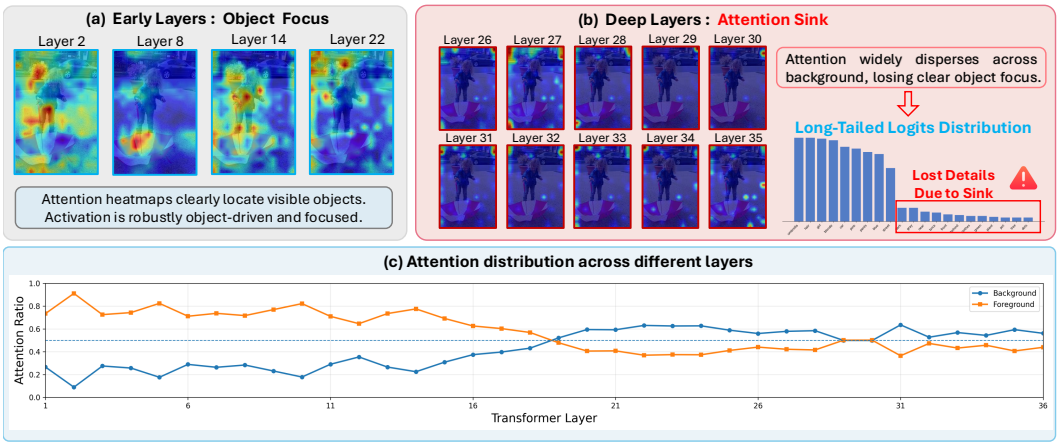


Figure 1: Layer-wise analysis of attention distribution in Qwen3-VL. (a) Attention in early Transformer layers is clearly concentrated on foreground objects. (b) Attention in deep layers disperses across background area. (c) Average foreground and background attention ratios across Transformer layers on the CompreCap dataset, showing the layer-wise variation of attention distribution.

VLMs, without updating model parameters or introducing external visual expert models. Specifically, FoAReD employs a novel contrastive decoding strategy that contrasts next-token distributions obtained under the original and contrastive attention. To obtain the contrastive attention, FoAReD introduces foreground-guided attention reallocation, which consists of three stages: visual token classification, head classification, and inverse reallocation. visual token classification classifies visual tokens into foreground, background and sink tokens. head classification selects heads suitable for reallocation based on their attention allocation patterns. Finally, inverse reallocation reallocates attention from sink and background tokens toward under-attended foreground tokens. The resulting attention is then used for attention contrastive decoding.

Our main contributions are summarized as follows.

- We propose **FoAReD**, a framework for image captioning by improving visual information utilization during autoregressive decoding, without external visual expert models and additional computational costs.
- We introduce a novel contrastive decoding strategy that contrasts prediction logits obtained under the original and contrastive attention. By amplifying token preferences induced by attention reallocation, FoAReD encourages the model to better utilize under-attended foreground visual information.
- Extensive experiments on CompreCap and CaptionQA across multiple VLMs demonstrate the effectiveness of FoAReD.

2 Related Work

Image Captioning. Image captioning aims to generate natural language descriptions from visual inputs (Vinyals et al., 2015; Xu et al., 2015; Rennie et al., 2017; Anderson et al., 2018; Johnson et al., 2016). Recent advances in VLMs have substantially improved image captioning. Existing methods for improving VLM-based image captioning can be broadly categorized into training-based and training-free methods. Training-based methods improve captioning through additional supervision or specialized optimization strategies, including large-scale grounded captioning data (Li et al., 2024; 2025), preference optimization (Zeng et al., 2025), and reinforcement learning (Xing et al., 2025). Although effective, these methods require additional training resources and optimization. To avoid additional training costs, recent studies explore training-free methods for enhancing image captioning. These methods enhance captioning through inference-time modifications, such as additional guidance (Xing et al., 2026) or contextual information (Pi et al., 2024; Peng et al., 2025). These methods often require extra reasoning steps or visual expert models such as object detectors

(Huang et al., 2025), segmentation models (Kirillov et al., 2023), or grounding models (Liu et al., 2024b). Different from exsiting methods, FoAReD requires no additional training costs, guidance or contextual information and relies solely on the model's internal representations.

Attention Sink and Attention Reallocation. Generating detailed image captions requires the model to accurately attend to informative visual tokens during autoregressive decoding. However, attention distributions in VLMs may become biased toward tokens with limited semantic information. This phenomenon is called attention sink (Xiao et al., 2024; Darcet et al., 2024), which is potentially due to massive activations in specific hidden-state dimensions (Sun et al., 2024; Cancedda, 2024). Recent methods reallocate attention from sink tokens to other visual tokens (Kang et al., 2025; Liu et al., 2026; Lu et al., 2025b; Tang et al., 2026). However, these methods do not distinguish foreground from background during reallocation. As a result, the reallocated attention may still be assigned to background tokens with limited details, making them less suitable for fine-grained image captioning.

Contrastive Decoding for VLMs. Contrastive Decoding (CD) (Li et al., 2023) is a training-free decoding paradigm that improves text generation by contrasting next-token distributions under different conditions. Recent studies extend CD to VLMs to improve visual-conditioned text generation. Since image captioning is a typical visual-conditioned text generation task, decoding strategies that improve the utilization of visual information are also important for generating detailed captions. Existing methods construct contrastive objectives from noised images (Leng et al., 2024), perturbed instructions (Wang et al., 2024), modality-specific predictions (Liu et al., 2024a), or internal model layers (Wu et al., 2026), but mainly target hallucination reduction. In contrast, FoAReD constructs contrastive distributions from attention reallocation for fine-grained image captioning, encouraging better utilization of detailed foreground visual information.

3 METHOD

We propose **FoAReD**, a framework which integrates attention contrastive decoding for image captioning. Figure 2 shows the overview of FoAReD.

Image captioning with VLMs generates a caption $Y = \{y_1, \ldots, y_T\}$ for an image I and text prompt x in an autoregressive manner, where T denotes the number of generated tokens. The image is encoded into visual tokens $\mathcal{V}$ through a image encoder E and a projector P: $\mathcal{V} = P(E(I)) = \{v_i\}_{i=1}^{N}$, where N denotes the total number of visual tokens. Meanwhile, the prompt x is tokenized as: $X = \text{Tokenizer}(x)$. We consider a LLM parameterized by θ, and at each decoding step t, the next token is predicted based on the $\mathcal{V}$, X, and previously generated tokens. Formally, the generation process can be expressed as:

$$y_t \sim p_\theta(y_t|\mathcal{V}, X, y_{<t}) \propto \exp(\text{logit}_\theta(y_t|\mathcal{V}, X, y_{<t})), \tag{1}$$

where y_t denotes the token generated at step t, and $y_{<t}$ represents the previously generated token sequence.

3.1 ATTENTION CONSTRASTIVE DECODING

Contrastive Decoding (Li et al., 2023) aims to mitigating hallucinations by contrasting next-token distributions obtained under different conditions. Specifically, at decoding step t, given two next-token distributions under contrastive condition c_A and original condition c_B, Contrastive Decoding computes the contrastive objective as:

$$\Delta\text{logit}_\theta(y_t|\mathcal{V}, X, y_{<t}) = \text{logit}_\theta^{c_A}(y_t|\mathcal{V}, X, y_{<t}) - \text{logit}_\theta^{c_B}(y_t|\mathcal{V}, X, y_{<t}). \tag{2}$$

In VLM-based image captioning, visual tokens contribute to next-token prediction through the attention mechanism. However, the attention distribution may become biased toward background or sink tokens with limited information. Attention reallocation can alleviate this bias by adjusting the attention distribution. Through reallocation, we obtain a new attention distribution $\tilde{A}$, which serves as the contrastive attention condition to the original attention A. Based on these two attention conditions, we construct the contrastive objective. Specifically, at each decoding step t, we compute:

$$\Delta\text{logit}_\theta(y_t|\mathcal{V}, X, y_{<t}) = \text{logit}_\theta^{\tilde{A}}(y_t|\mathcal{V}, X, y_{<t}) - \text{logit}_\theta^{A}(y_t|\mathcal{V}, X, y_{<t}).$$

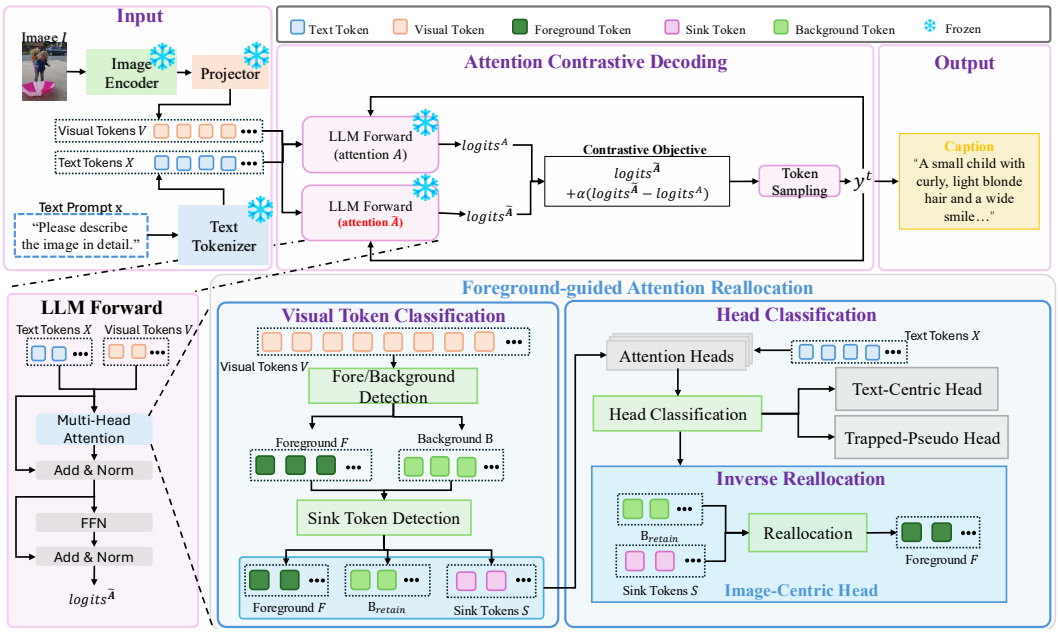


Figure 2: Overview of FoAReD. FoAReD enhances image caption through attention contrastive decoding, which contrasts next-token distributions obtained under the original attention A and contrastive attention $\tilde{A}$. To construct $\tilde{A}$, FoAReD performs foreground-guided attention reallocation with three stages: visual token classification, head classification, and inverse reallocation. It reallocates attention from sink and background tokens toward under-attended foreground tokens to improve the utilization of visual information.

This contrastive objective captures the changes in token predictions induced by attention reallocation, highlighting visual information that may be underutilized under the original attention distribution A.

The final next-token distributions are obtained by amplifying this contrastive objective:

$$\text{logit}_\theta^{ACD}(y_t|\mathcal{V}, X, y_{<t}) = \text{logit}_\theta^{\tilde{A}}(y_t|\mathcal{V}, X, y_{<t}) + \alpha\Delta\text{logit}_\theta(y_t|\mathcal{V}, X, y_{<t}). \tag{3}$$

where α controls the strength of the contrastive objective. Then the next token can be sampled according to:

$$y_t \sim p_\theta^{ACD}(y_t|\mathcal{V}, X, y_{<t}) \propto \exp(\text{logit}_\theta^{ACD}(y_t|\mathcal{V}, X, y_{<t})),$$

By contrasting the next-token distributions from the original and contrastive attention, FoAReD strengthens the token preferences induced by attention reallocation. This enables the model to better utilize its internal visual representations and generate captions with richer details.

3.2 Foreground-guided Attention Reallocation

To obtain the reallocated attention $\tilde{A}$, we introduce foreground-guided attention reallocation. Attention reallocation adjusts the original attention distribution by reducing attention assigned to less informative tokens and enhancing informative visual tokens. It has been explored in Kang et al. (2025) for general vision-language tasks such as VQA. It consist of three stages: *sink token detection*, *head classification*, and *attention reallocation*. However, directly applying them to image captioning is insufficient. First, *sink token detection* does not distinguish foreground and background tokens, causing reallocated attention may be assigned to background tokens with limited visual information. Therefore, FoAReD introduces *visual token classification* to identify foreground and background toksn before sink token detection. Second, *head classification* identifies image-centric heads based on attention over all visual tokens, while caption generation relies more on foreground visual details. Thus, FoAReD selects heads according to their attention allocation toward foreground

tokens. Third, *attention reallocation* follows the original attention distribution, which may reinforce already highly attended tokens. FoAReD therefore introduces *inverse reallocation* to assign larger increments to under-attended foreground tokens.

3.2.1 VISUAL TOKEN CLASSIFICATION

Low-texture or repetitive background regions, such as sky, walls or roads usually tend to produce weaker feature representations and visual tokens from these regions tend to exhibit lower L2 norm values of visual feature. Therefore, we rank visual tokens by their L2 norm values and heuristically select tokens with the values as background tokens: $q_i = \|v_i\|_2$, $\mathcal{I}_B = \mathrm{TopK}(-q_i)$, $\mathcal{B} = \{v_i | i \in \mathcal{I}_B\}$, where q_i denotes the L2 norm value of the i-th visual token, $\mathcal{I}_B$ denotes the indices of selected background candidates and B denotes the detected background tokens The remaining tokens are treated as foreground tokens: $\mathcal{F} = \mathcal{V} \setminus \mathcal{B}$.

We further employ pruning to remove redundant background tokens to reduce unnecessary background competition. Specifically, we adopt the token pruning strategy from Fang et al. (2026) and apply it only to the background tokens. The pruning process partitions the background tokens into two subsets: $\mathcal{B}retain$, the retained representative background tokens, and $\mathcal{B}prune$, the removed redundant tokens. Then he remaining visual token set is defined as: $\mathcal{V}' = \mathcal{V} \setminus \mathcal{B}_{prune} = \mathcal{F} \cup \mathcal{B}_{retain}$.

Then, we further detect visual sink tokens $\mathcal{S}$ following Kang et al. (2025). $\mathcal{S} \subseteq \mathcal{V}'$,

Token pruning changes the visual token sequence by removing redundant tokens, which may break their original 2D spatial positions if they are directly re-indexed. Therefore, FoAReD preserves the original position IDs of retained tokens. Let $\mathbf{P}$ denote the original position IDs and $\mathcal{I}_{\mathcal{V}'}$ denote the indices of retained visual tokens. The corresponding position IDs after pruning are obtained as: $\mathbf{P}' = \mathbf{P}[\mathcal{I}_{\mathcal{V}'}]$. This preserves the original spatial correspondence among the remaining visual tokens.

3.2.2 HEAD CLASSIFICATION

In image captioning, foreground tokens usually contain the visual details required for generating comprehensive captions, while background regions often provide limited information. Therefore, attention heads that focus more on foreground tokens are more suitable for attention reallocation. Instead of measuring attention distribution over all visual tokens, FoAReD identifies Image-Centric heads according to their attention assigned to foreground tokens. Meanwhile, heads whose visual attention is dominated by sink tokens are identified as Trapped-Pseudo heads and excluded from attention reallocation.

Specifically, let $A^{\ell}_{h,t,j}$ denote the attention weight from decoding token t to visual token j in head h. We first calculate the total visual attention: $V^{\ell}_h = \sum_{j \in \mathcal{V}'} A^{\ell}_{h,t,j}$.

FoAReD measures the attention allocation toward foreground tokens by computing the foreground attention ratio of each head:

$$R^{\ell}_h = \frac{\sum_{j \in \mathcal{F} \setminus \mathcal{S}^{\ell}} A^{\ell}_{h,t,j}}{V^{\ell}_h + \epsilon}. \tag{4}$$

Heads with sufficient foreground attention are classified as Image-Centric heads. Meanwhile, heads dominated by sink tokens are classified as Trapped-Pseudo heads and excluded from reallocation. More results about Trapped-Pseudo heads can be found in Appendix 6.2.

3.2.3 INVERSE REALLOCATION

FoAReD performs inverse attention reallocation. Specifically, FoAReD extracts attention from visual sink tokens and retained background tokens, and reallocates it to foreground tokens. And FoAReD adopts an inverse allocation strategy, assigning larger increments to foreground tokens with lower original attention weights. Specifically, within the Image-Centric heads, we define the extraction set as: $\mathcal{E} = \mathcal{S} \cup \mathcal{B}_{\mathrm{keep}}$,

The extracted attention weight is computed as: $\Omega = p \sum_{j \in \mathcal{E}^{\ell}} A^{\ell}_{h,t,j}$, where p denotes the predefined extraction ratio.

The extracted attention is then reallocated to the foreground tokens $\mathcal{F}$. To implement inverse reallocation, FoAReD assigns larger attention increments to foreground tokens with lower original attention weights. Specifically, the reallocated attention is computed as:

$$\tilde{A}^{\ell}_{h,t,i} = A^{\ell}_{h,t,i} + \Omega \frac{1 - A^{\ell}_{h,t,i}}{\sum_{u \in \mathcal{F}} (1 - A^{\ell}_{h,t,u})}, \qquad i \in \mathcal{F}. \tag{5}$$

The inverse weighting term $(1 - A^{\ell}_{h,t,i})$ assigns larger increments to foreground tokens that are insufficiently attended in the original attention distribution. Finally, the reallocated attention distribution $\tilde{A}^{\ell}$ is re-normalized before computing the attention output.

Through foreground-guided inverse attention reallocation, FoAReD reallocates attention from background and sink tokens toward under-attended foreground regions, improving the utilization of visual representations for caption generation.

4 Experiments

4.1 Setup and implementation Details

We compare FoAReD with three categories of related methods: captioning methods, including Image Textualization (IT) (Pi et al., 2024), Patch Matters (PM) (Peng et al., 2025), and ScaleCap (Xing et al., 2026); attention-based methods, including VAR (Kang et al., 2025) and VisFlow (Tang et al., 2026); and contrastive decoding methods, represented by VCD (Leng et al., 2024). For VAR, VisFlow, and VCD, which were originally developed for other tasks, we evaluate their effectiveness in the image captioning setting. Experiments are conducted on CompreCap, CaptionQA, AIM-500, and AMBER (Lu et al., 2025a; Yang et al., 2026; Li et al., 2021; Wang et al., 2023). To evaluate cross-model generalization, we further test FoAReD on Qwen3-VL-4B (Bai et al., 2025), LLaVA-v1.6-7B (Liu et al., 2023), and InternVL3-2B (Zhu et al., 2025). All datasets and model checkpoints are used under their respective licenses and are not redistributed.

4.2 Main results

Captioning Performance. Table 1 reports the captioning performance of different methods. Compared with captioning methods, including IT, PM, and ScaleCap, FoAReD achieves more consistent improvements across different VLMs. These methods mainly improve caption quality by introducing additional visual cues or reasoning steps, while FoAReD directly improves the utilization of visual representations already encoded in VLMs. The consistent gains in 'Unified' scores across Qwen3-VL-4B, LLaVA-v1.6-7B, and InternVL3-2B demonstrate the effectiveness of enhancing visual information utilization. Compared with attention-based methods such as VAR and VisFlow, FoAReD obtains stronger improvements across multiple captioning metrics. Since FoAReD performs attention reallocation with distinguishing informative foreground tokens from background and sink tokens. This enables more effective utilization of visual details. The results suggest that effective attention reallocation requires not only redistributing attention, but also accurately identifying informative visual tokens and transferring the resulting visual preference to the final generation process. The results highlight the importance of allocating attention to more informative foreground tokens. Compared with contrastive decoding method VCD, FoAReD achieves better captioning performance with comparable inference cost. Different from VCD, which constructs contrastive objective by adding noise to visual input, FoAReD constructs contrastive objective from foreground-guided attention reallocation. The consistent improvements demonstrate that incorporating foreground-aware visual preferences provides a more effective contrastive objective for image captioning.

Although some methods achieve better performance on the 'Embodied' category of CaptionQA, this may be attributed to the characteristics of this category. Images in this category are mainly complex traffic scenes, where important semantic cues can also appear in background regions, such as surrounding vehicles and traffic lights. Since FoAReD reduces background token redundancy through pruning, this strategy may be less beneficial in such scenarios. Nevertheless, FoAReD

Table 1: Performance comparison with other training-free methods on CompreCap and CaptionQA. The best and second-best performance scores are highlighted in **bold** and underline, respectively. * indicates that the model is trained on the CapRL-2M dataset (Xing et al., 2025) and is included only as a reference, not for comparative analysis.

Method	Latency (s/image)	CompreCap				CaptionQA	
		Object	Attribute (0∼5)	Relation (0∼5)	Unified	Natural	Embodied
Qwen3 VL 4B							
Base	22.14	81.10	2.89	2.95	64.05	83.19	79.05
+CapRL*	-	86.1	3.07	3.09	67.71	88.73	84.62
+IT	51.60	79.83	2.82	2.90	62.93	86.15	82.23
+PM	99.64	81.94	2.94	3.00	65.10	84.35	**83.88**
+ScaleCap	135.90	79.65	2.91	2.99	64.17	85.40	83.04
+VAR	32.27	82.80	2.95	2.99	65.27	82.93	81.10
+VisFlow	54.74	81.63	2.88	2.95	64.14	83.19	78.69
+VCD	45.09	83.32	**2.96**	2.97	65.29	85.34	82.57
+Ours	43.86	**85.25**	2.95	**3.01**	**66.05**	**86.75**	83.17
LLaVA v1.6 7B							
Base	11.74	68.22	2.42	2.66	55.27	66.04	64.61
+CapRL*	-	75.87	2.71	2.86	60.85	67.92	66.75
+IT	13.55	68.62	2.43	2.68	55.64	66.72	64.66
+PM	27.69	64.81	2.24	2.65	53.14	65.56	64.74
+ScaleCap	37.04	66.98	2.38	2.36	54.84	66.49	**66.04**
+VAR	15.75	68.79	2.43	2.70	55.80	66.49	65.26
+VisFlow	17.22	68.65	2.43	2.65	55.31	63.35	61.55
+VCD	24.56	70.11	2.46	**2.75**	56.74	66.72	65.13
+Ours	22.91	**70.15**	**2.49**	2.72	**57.18**	**66.87**	65.58
InternVL3 2B							
Base	7.45	70.56	2.50	2.87	58.10	74.82	73.16
+CapRL*	-	80.52	2.87	2.97	63.99	76.01	74.59
+IT	11.83	70.31	2.48	2.87	57.92	75.10	73.33
+PM	31.79	72.46	2.58	2.89	59.28	74.77	73.50
+ScaleCap	33.82	68.57	2.42	2.92	57.42	75.33	73.59
+VAR	10.39	73.57	2.61	2.91	59.91	**75.81**	**74.08**
+VisFlow	12.17	71.66	2.52	2.88	58.54	73.23	73.41
+VCD	13.75	71.38	2.56	2.88	58.73	74.95	73.27
+Ours	13.24	**80.33**	**2.82**	**2.94**	**63.40**	**75.81**	73.62

achieves consistent overall improvements across models and benchmarks. Additional qualitative captioning examples are provided in Appendix 6.3.

Quantitative Evaluation of Visual Token Classification. We evaluate the heuristic visual token classification across different backbones on AIM-500. 'Pre.' denotes background-token precision, while 'Mis.' denotes the foreground-to-background misclassification rate. As shown in Table 2, 'Pre.' reaches 0.894, 0.795, and 0.800 on Qwen3-VL, LLaVA-v1.6, and InternVL3, respectively. 'Mis.' reaches 0.058, 0.099, and 0.061. These results show that visual token classification based on L2-norm of visual token features can provide a practical approximation for identifying background candidates across different VLMs.

Comparison with a Larger Model. We compare Qwen3-VL-4B with FoAReD against the larger Qwen3-VL-8B model. As shown in Table 3, Qwen3-VL-4B with FoAReD achieves a 'Unified' score of 66.05, outperforming Qwen3-VL-8B, which obtains 65.23. FoAReD also achieves higher scores on 'Object', 'Attribute', and 'Relation'. These results suggest that FoAReD can provide greater benefits than simply increasing model size for comprehensive image captioning.

Hallucination Evaluation. We evaluate whether generating more fine-grained details increases hallucination on AMBER. As shown in Table 4, FoAReD consistently reduces hallucination across all three backbones. The hallucination score decreases from 18.4 to 14.7 on Qwen3-VL, from 58.9

Table 2: Heuristic visual token classification.

Model	Pre. (↑)	Mis. (↓)
Qwen3 VL	0.894	0.058
LLaVA v1.6	0.795	0.099
INternVL3	0.800	0.061

Table 3: Comparison between FoAReD and a larger-model baseline.

Model	Obj.	Attr. (0∼5)	Rel. (0∼5)	Unified
Qwen3 VL 4B +Ours	85.25	2.95	3.01	66.05
Qwen3 VL 8B	83.27	2.93	3.00	65.23

Table 4: Hallucination evaluation across different VLM backbones. Lower scores indicate fewer hallucinations.

Model	Base	Ours
Qwen3	18.4	14.7
LLaVA v1.6	58.9	55.0
InternVL3	34.9	23.2

Table 5: Comparison with Length-Controlled Prompting.

Method	Obj.	Attr. (0∼5)	Rel. (0∼5)	Unified
Long Prompt	82.59	2.94	2.99	65.27
Ours	85.25	2.95	3.01	66.05

Table 6: Ablation Study on the contribution of each component.

Disabled component	Unified
VTC	65.62
FAR	65.77
ACD	65.70
None	66.05

Table 7: Ablation study of background pruning under different attention extraction ratios.

Setting	p=0.6	p=0.7	p=0.8
w/ pruning	66.05	65.69	65.40
w/o pruning	65.62	65.29	65.02

Table 8: Ablation study of the contribution of different attention extraction tokens.

Extraction source	Unified
Sink Token	65.79
Background Token	65.84
Both	66.05

Table 9: Ablation study of different attention reallocation strategy.

Reallocation strategy	Unified
Reinforcement	65.27
Average	65.48
Inverse	66.05

to 55.0 on LLaVA-v1.6, and from 34.9 to 23.2 on InternVL3. These results indicate that FoAReD improves caption comprehensiveness without sacrificing visual faithfulness.

Comparison with Length-Controlled Prompting. To verify that the improvement is not simply due to longer captions, we compare FoAReD with a length-controlled prompt baseline using Qwen3-VL, where both methods generate captions with comparable lengths (around 300 words). As shown in Table 5, FoAReD achieves a higher 'Unified' and 'Object' score, while achieving comparable 'Attribute' and 'Relation' scores. This result suggests that the improvement comes from more effective visual information utilization rather than caption length alone.

4.3 Ablation Study

We conduct ablation studies on CompreCap with Qwen3-VL-4B to analyze the contributions of different components and design choices in FoAReD. We further evaluate inference latency across three VLMs.

Inference Latency. Table 1 reports the inference latency of different methods across VLM backbones. Compared with captioning methods such as IT, PM, and ScaleCap, FoAReD introduces substantially lower inference latency, as it does not require additional visual expert models or multi-step reasoning steps. Compared with attention-based methods such as VAR and VisFlow, FoAReD introduces moderate additional cost due to contrastive decoding, while achieving better captioning performance. Compared with the contrastive decoding method VCD, FoAReD has comparable

inference latency, as both methods require additional decoding computation. Overall, FoAReD achieves a favorable trade-off between inference efficiency and captioning performance.

Complementarity of Background Pruning and Attention Reallocation. We evaluate the effect of background pruning under different extraction ratios p. As shown in Table 7, pruning consistently improves the Unified score, from 65.62 to 66.05 at $p = 0.6$, from 65.29 to 65.69 at $p = 0.7$, and from 65.02 to 65.40 at $p = 0.8$. These results show that increasing the reallocation ratio cannot fully compensate for the absence of pruning. Background pruning and attention reallocation therefore play complementary roles. Pruning removes redundant background tokens from the visual sequence, thereby reducing background competition, whereas reallocation changes the attention distribution over the remaining tokens. Thus, reallocation alone cannot eliminate the redundancy caused by retained background tokens, while pruning alone does not actively increase attention to overlooked foreground tokens.

Effectiveness of Each Component. Table 6 presents the component-wise ablation study of FoAReD, including visual token classification(VTC), foreground-guided attention reallocation(FAR), and attention contrastive decoding(ACD). Removing visual token processing causes the largest performance drop, reducing the 'Unified' score from 66.05 to 65.62, highlighting the importance of reducing redundant visual token competition. Foreground-guided attention reallocation and contrastive decoding also provide complementary gains. Removing them decreases the 'Unified' score to 65.77 and 65.70, respectively. These results verify that the three components jointly improve visual information utilization.

Choice of attention extraction sources. Table 8 compares different attention extraction strategies. Using either background or sink tokens improves performance, while jointly extracting from both achieves the best result, increasing the 'Unified' score to 66.05 compared with 65.84 and 65.79 for background-only and sink-only extraction. This demonstrates that background and sink tokens introduce complementary attention competition, and jointly removing both enables more effective reallocation toward foreground tokens.

Comparison of attention reallocation strategies. As shown in Table 9, we compare three attention reallocation strategies: reinforcement reallocation, which follows existing attention reallocation strategies by strengthening high-attention foreground tokens (Kang et al., 2025); average reallocation, which uniformly distributes attention as a simple comparison baseline; and inverse reallocation, which is proposed in FoAReD to prioritize under-attended foreground tokens. The results shows that compensating under-attended foreground tokens is more effective than further enhancing already dominant visual tokens, leading to more comprehensive captions.

5 Conclusion

In this paper, we propose FoAReD, a framework for improving image captioning by better utilizing the visual information already encoded in VLMs. We observe that VLMs may under-utilize encoded visual information during autoregressive decoding, as attention tends to concentrate on redundant background and sink tokens. FoAReD addresses this issue through attention contrastive decoding, which leverages foreground-guided attention reallocation to better utilize informative foreground tokens. Experiments on CompreCap and CaptionQA across VLMs with different architectures demonstrate that FoAReD consistently improves captioning performance. Overall, our results highlight that improving the utilization of existing visual representations within VLMs provides an effective direction for enhancing image captioning without additional computational costs.

Limitations

Despite its effectiveness, our method has several limitations. First, it requires access to intermediate hidden states and attention maps, which limits its direct applicability to closed-source VLMs. Second, self-ablation contrastive decoding introduces additional inference cost, as it requires an extra raw branch at each decoding step. Third, foreground enhancement may suppress useful background context, which can be important for accurately describing the overall scene.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4117–4126, 2025.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4565–4574, 2016.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 13547–13574. International Conference on Learning Representations, ICLR, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, 2023.

Xiangtai Li, Tao Zhang, Yanwei Li, Haobo Yuan, Shihao Chen, Yikang Zhou, Jiahao Meng, Yueyi Sun, Shilin Xu, Lu Qi, et al. Denseworld-1m: Towards detailed dense grounded caption in the real world. *arXiv preprint arXiv:2506.24102*, 2025.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556, 2024.

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

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Xu Liu, Guikun Chen, and Wenguan Wang. Sinktrack: Attention sink based context anchoring for large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19618–19627, 2025a.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025b.

Ruotian Peng, Haiying He, Yake Wei, Yandong Wen, and Di Hu. Patch matters: Training-free fine-grained image caption enhancement via local perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3963–3973, 2025.

Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: an automatic framework for creating accurate and detailed image descriptions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 108116–108139, 2024.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9439–9447, 2026.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. doi: 10.48550/arXiv.2311.07397.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

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*, pp. 15840–15853, 2024.

Yebo Wu, Han Jin, Zhijiang Guo, and Li Li. Spotlight and shadow: Attention-guided dual-anchor introspective decoding for mllm hallucination mitigation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13219–13233, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jinsong Li, Shuangrui Ding, Weiming Zhang, Nenghai Yu, et al. Scalecap: Scalable image captioning via dual-modality debiasing. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23741–23750, 2026.

Yu Zeng, Yukun Qi, Yiming Zhao, Xikun Bao, Lin Chen, Zehui Chen, Shiting Huang, Jie Zhao, and Feng Zhao. Enhancing large vision-language models with ultra-detailed image caption generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26703–26729, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

6 Appendix

6.1 Background token detection

We detect background patches based on the activation strength of visual patch features. Given the visual tokens extracted by the vision encoder, we compute the L2 norm of each patch feature and regard patches with relatively low feature norms as background candidates. Specifically, we rank all visual patches in ascending order according to their feature norms and select the lowest proportion as background patches. Figure 3 shows the detected background patches under different detection thresholds.

6.2 Discussion about Trapped-pseudo Heads

In our proposed FoAReD framework, attention reallocation is strictly confined to *Image-Centric* heads, explicitly bypassing *Trapped-Pseudo* heads. To empirically validate the necessity of this selective intervention, we conduct an ablation study where the same reallocation strategy is forcibly applied to Trapped-Pseudo heads across varying reallocation coefficients.

As illustrated in Figure 4, forcibly redistributing the attention of Trapped-Pseudo heads leads to a consistent performance degradation. Instead of yielding improvements, the intervention introduces a deleterious effect that generally amplifies as the reallocation coefficient increases.

This counter-intuitive phenomenon corroborates our hypothesis regarding the intrinsic mechanisms of large vision-language models. It suggests that Trapped-Pseudo heads do not simply "fail" by collapsing onto attention sinks; rather, they deliberately utilize these non-semantic sink tokens as functional anchors. These heads likely bear special structural responsibilities, such as aggregating



Figure 3: Visualization of background patch detection under different thresholds. Red boxes indicate the patches detected as background.

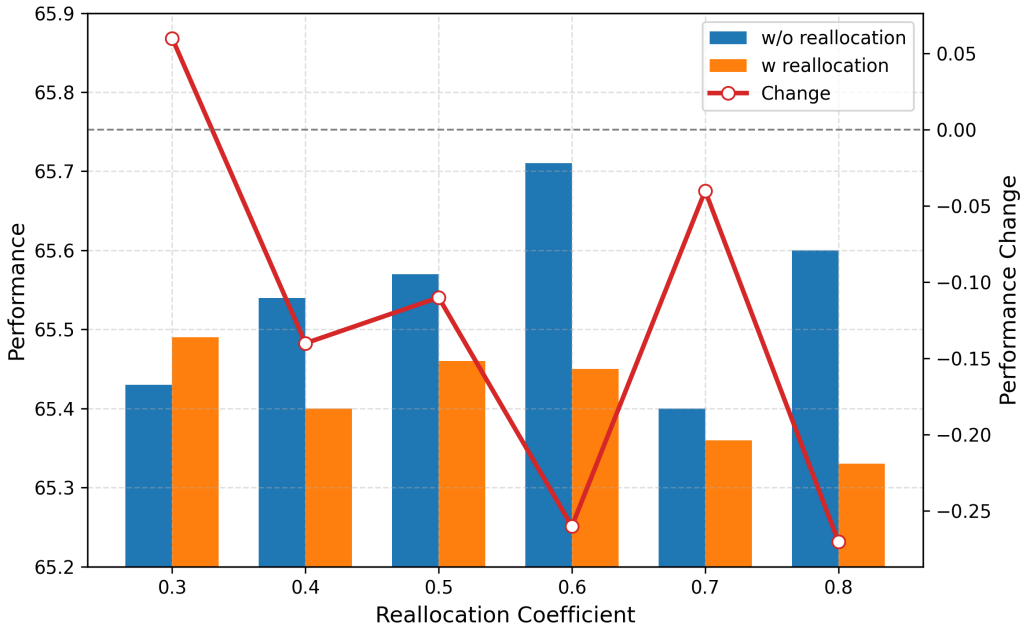


Figure 4: Impact of Intervening Trapped-Pseudo Heads Across Varying Reallocation Coefficients



Qwen-3-VL 4B:
The image shows a young child standing outside on a sidewalk or driveway, playing with a bright pink umbrella that gray is lying upside down on the ground. The umbrella has magenta panels, pale pink sections, and white polka dots inside, with its handle standing upright.
The child has curly blond hair partially covering their face and is wearing a blue long-sleeved shirt, gray pants, and dark shoes. They appear focused on the umbrella, reaching toward it as if trying to pick it up or adjust it.
In the background, there is a quiet street with a parked gray car, brick buildings, a low brick wall, a tree, and a potted plant. The scene feels like a candid outdoor moment, centered on the child's curiosity and play.

Qwen-3-VL 4B + FoAReD:
The image shows a young child standing outdoors on a sidewalk or driveway, interacting with an open pink umbrella that is lying upside down on the ground. The umbrella is bright magenta on the outside with pale pink panels and white polka dots on the inside. Its metal shaft is upright, and the child appears to be holding or reaching toward the handle.
The child has long, curly blond hair falling partly over their face and is wearing a blue long-sleeved top, gray pants, and dark shoes. Their posture suggests they are focused on the umbrella, possibly trying to lift it, play with it, or figure out how it works.
In the background, there is a quiet residential street with a gray car parked along the curb. Behind the car are brick buildings, a low brick wall, a tree trunk, and a large potted plant with dry, pale leaves. The scene feels candid and everyday, capturing a small moment of play or curiosity on a neighborhood sidewalk.

Figure 5: Qualitative Caption Comparison between Qwen-3-VL and FoAReD

global visual context, buffering cross-modal alignment, or maintaining the stability of the autoregressive logic flow. Consequently, forcibly diluting their attention away from these sink anchors disrupts their intended operations, confirming that semantic grounding intervention must be precisely targeted.

6.3 Visualization Examples of Captions

We provide more qualitative visualization examples showing that our method generates more detailed captions with richer attribute descriptions and more explicit object-relation information. The visualization is shown in Figure 5 and Figure 6.

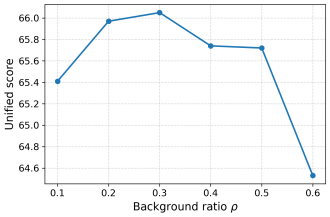


Qwen-3-VL 4B:
The image shows a man standing outdoors next to a traffic light display. He is wearing a white T-shirt, light gray shorts, and black flip-flops, and is facing the camera with a relaxed posture. Beside him is a traffic signal mounted on a pole, with the red light illuminated at the top. Below the signal is a sign that reads "AUSTRALIAN TRAFFIC LIGHT." The area around them is filled with dense green plants, broad leaves, small white flowers, rocks, and a few shrubs, giving the scene a garden-like or tourist-attraction setting. A red-and-yellow bollard is visible near the lower right corner.

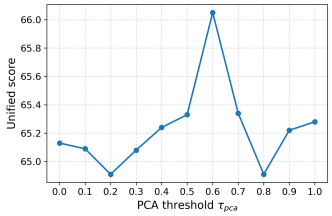
Qwen-3-VL 4B + FoAReD:
The image shows a man standing outdoors beside a traffic light display, surrounded by dense green tropical plants and flowering shrubs. He is positioned on the left side of the frame, wearing a white T-shirt, light gray shorts, and black flip-flops. He is standing casually on a gravel or sandy path, facing the camera.
To his right is a tall traffic signal mounted on a metal pole. The red light at the top is illuminated, while the amber and green lights below are unlit. Under the traffic light, there is a sign that reads **"AUSTRALIAN TRAFFIC LIGHT"**, suggesting this may be part of a tourist display, roadside attraction, or themed exhibit rather than a normal street intersection.
The background is filled with lush vegetation, including broad green leaves and small white flowers. There are also rocks, small plants, and a red-and-yellow bollard near the bottom right corner. Overall, the image looks like a casual travel photo taken in a garden-like outdoor setting, with the traffic light and sign as the main point of interest.

Figure 6: Qualitative Caption Comparison between Qwen-3-VL and FoAReD

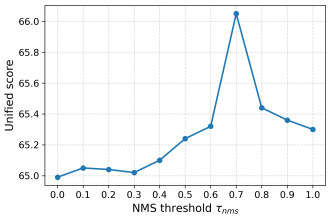
6.4 HYPERPARAMETERS



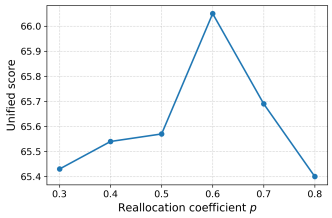
(a) Impact of different background patch detection threshold ρ



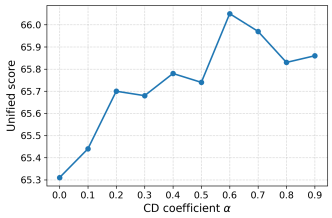
(b) Impact of different PCA energy threshold τ_{pca}



(c) Impact of different NMS similarity threshold τ_{nms}



(d) Impact of different reallocation coefficient p



(e) Impact of different CD coefficient α

Figure 7: Visualization of attention allocation.

Background ratio. Figure 7a reports the impact of the background patch ratio on model performance. The score increases from 65.41 to 66.05 as the ratio rises from 0.1 to 0.3, showing that a moderate number of low-activation background candidates helps reduce redundant visual attention. However, larger ratios gradually degrade performance, likely because informative patches are mistakenly treated as background. Overall, a ratio of 0.3 provides a favorable balance between background redundancy suppression and visual information preservation.

PCA coefficient. Figure 7b shows the effect of τ_{pca} on background token pruning. The score improves as τ_{pca} increases and reaches the best result at $\tau_{pca} = 0.6$, suggesting that retaining sufficient principal semantic energy helps preserve useful background context for caption generation. How-

ever, further increasing τ_{pca} leads to a slight drop, likely because redundant semantic groups are retained, making the pruned token set less compact. These results indicate that $\tau_{pca} = 0.6$ provides a favorable balance between semantic diversity preservation and redundancy suppression.

NMS coefficient. Figure 7c shows the effect of the NMS coefficient τ_{nms}, with the best performance achieved at $\tau_{nms} = 0.7$. A smaller value may remove semantically distinct background tokens, while a larger value may retain repetitive ones. This suggests that $\tau_{nms} = 0.7$ provides a favorable balance between semantic preservation and redundancy pruning.

Ratio of reallocation. We analyze the effect of the reallocation coefficient p. As shown in Figure 7d, increasing p from 0.3 to 0.6 improves the 'Unified' score, indicating that moderate reallocation helps redirect attention toward caption-relevant foreground regions. Larger values degrade performance, likely because they disturb the original attention distribution. Therefore, we set $p = 0.6$ by default.

CD coefficient. We analyze the effect of the contrastive decoding coefficient α. As shown in Figure 7e, increasing α from 0 to 0.6 improves the 'Unified' score, indicating that moderate contrastive decoding strengthens the effect of foreground-guided reallocation. Larger values degrade performance, likely because they disturb the original next-token distribution. Therefore, we set $\alpha = 0.6$ by default.