

Supplemental Material for:

CAAC: Confidence-Aware Attention Calibration to Reducing Hallucinations in Large Vision-Language Models

Paper ID 2870

Anonymous Submission

1 Further Experimental Details

1.1 Baselines.

We compare CAAC with four training-free, inference-time approaches:

- **VCD** (Visual Contrastive Decoding) [3]: each token is decoded twice—once with the original image and once with a perturbed copy—and words whose likelihood collapses under perturbation are down-weighted.
- **AvisC** [4]: first performs *attentional vision calibration* by masking high-attention outlier tokens, then applies contrastive decoding to suppress visually ungrounded candidates.
- **M3ID** [1]: re-scores candidate tokens with a lightweight image-guided gradient signal, promoting those whose gradients align with visual features and filtering hallucinations.
- **OPERA** [2]: augments beam search with an *over-trust penalty* and a *retrospection-allocation* term, penalising tokens that receive insufficient cumulative attention from the image.

1.2 Implementation Details.

For the baselines, we adopt the hyperparameter settings reported in their respective papers to ensure consistency. For CAAC, we set the smoothing parameter β to 0.7 for LLaVA, 0.3 for LLaVA-NeXT, and 0.5 for InstructBLIP. The maximum scaling factor for AAR is set to $\lambda_{\max} = 1.5$ for LLaVA and LLaVA-NeXT, and 2 for InstructBLIP, with $\lambda_{\min} = 1.0$ and $p_{thr} = 0.25$. Also, the number of layers with VTC is set to the first ten decoder layers for all three models. Data and code are available at <https://github.com/mehrdadfazli/CAAC/>.

1.3 Hardware and runtime.

All experiments were conducted on a server equipped with $4 \times$ NVIDIA H100 (40 GB) GPUs and 512 GB of system RAM. We evaluate all models in 16-bit floating-point precision using `transformers` v4.47 from Hugging Face.

To ensure a fair comparison, we used our own implementations of the contrastive decoding (CD) baselines and CAAC, adapted for the Hugging Face version of LLaVA-NeXT. For runtime analysis, we selected 20 images from the CHAIR benchmark and paired each with the prompt:

Table 1: Mitigation Methods GPU Runtime per Token (ms)

Mitigation Methods	GPU Runtime per Token (ms)
VCD	298
M3ID	217
AvisC	298
CAAC (ours)	171

"please describe the image in detail." We collected the generated captions and measured the total GPU runtime and the total number of generated tokens.

We then computed the average GPU time per token for each baseline and CAAC. The results are reported in table 1. As shown, CAAC is considerably faster than the baselines due to its limited number of second forward passes per token.

1.4 CAAC hyper-parameters.

We use the following values, selected via the grid search: $\beta = 0.7$ for LLaVA-1.5, $\beta = 0.3$ for LLaVA-NeXT, and $\beta = 0.5$ for InstructBLIP; $\lambda_{max} = 1.5$, $\lambda_{min} = 1.0$; confidence threshold $p_{thr} = 0.25$; VTC applied to the first 10 decoder layers (out of 32). For the discriminative task (POPE) with InstructBLIP, we set $\lambda_{min} = 0$, $\lambda_{max} = 1.8$, and applied VTC on the first five layers of the decoder. This setting optimized the performance of InstructBLIP for discriminative tasks while ensuring that the scale factor is greater than 1 when $p < P_{thr}$.

1.5 Baseline implementations and settings.

Implementations for LLaVA-1.5 and InstructBLIP. We utilize official repositories and recommended configurations for all baselines:

- **OPERA** [2]: We use the official implementation¹ with the following settings: `beam_size=5`, `num_cands=5`, `scale_factor=50`, $\alpha = 1$, $\beta = 5$, and $r = 15$.
- **Contrastive Decoding Baselines:** Implemented using the official `AvisC` repository² with the authors' recommended hyperparameters:
 - **VCD** [3]: $\alpha = 1$, $\beta = 0.1$, $\gamma = 0.1$
 - **AvisC** [4]: $\alpha = 2.5$ (LLaVA) / 3.0 (InstructBLIP), $\gamma = 0.5$, $\lambda = 1$
 - **M3ID** [1]: $\lambda = 0.2$

Implementations for LLaVA-NeXT. Since official implementations for these baselines are not available for LLaVA-NeXT, we adapted the contrastive decoding methods to work with the Hugging Face Transformers version of LLaVA-NeXT using the `AvisC` codebase. For each baseline, we tested 5–10 hyperparameter configurations to identify the most effective settings within our computational budget.

We attempted to implement OPERA for LLaVA-NeXT using Transformers v4.47, which supports LLaVA-NeXT; however, we were unable to finish the implementation successfully due to its heavy reliance of OPERA's modified beam search on the transformers' internal module that has

¹<https://github.com/shikiw/OPERA>

²<https://github.com/sangminwoo/AvisC>

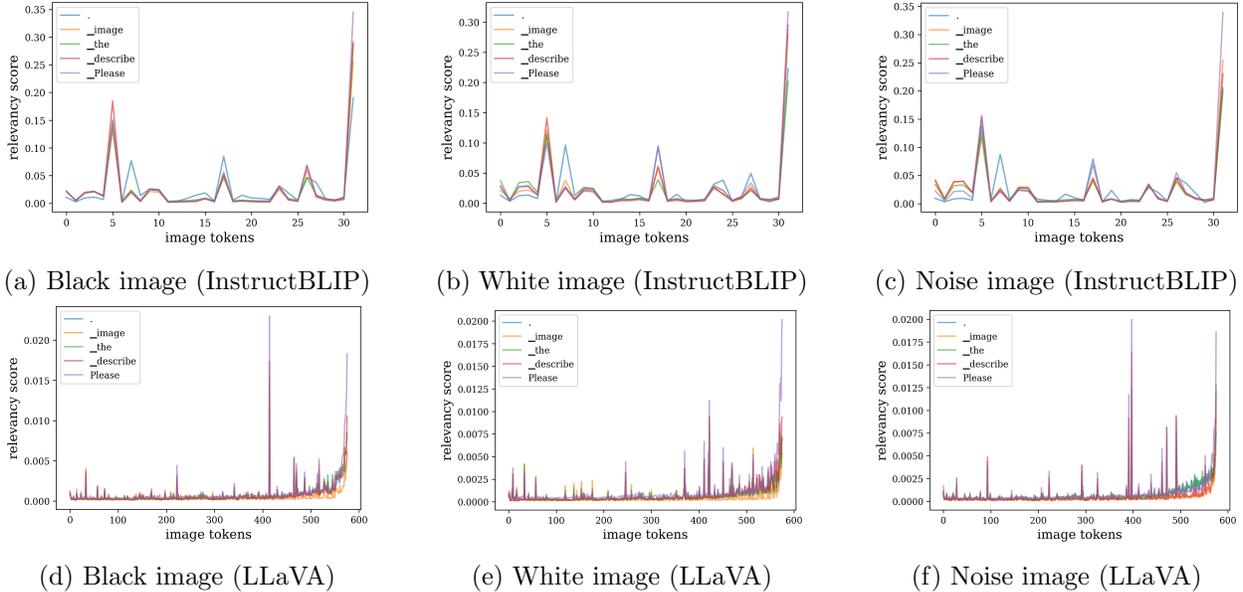


Figure 1: Distribution of relative image relevancy scores from InstructBLIP and LLaVA given plain (a-c) black, (d-f) white, and (g-i) noise images with the query “Please describe the image.” The distributions of relevancy scores are nearly identical regardless of the reference input image, supporting the robustness of the VTC module.

been substantially updated since the 4.29 version. Thus, we sourced the results from [5] for CHAIR and POPE benchmarks.

Hyperparameter Search. The following configurations yielded the best performance for each baseline on LLaVA-NeXT:

- **VCD:** $\alpha \in \{1.0, 1.5, 2.0\}$, noise step $\in \{500, 750\}$; best: $\alpha = 1.5$, noise step=750.
- **M3ID:** $\lambda \in \{0.1, 0.2, 0.4, 0.6\}$; best: $\lambda = 0.1$.
- **AvisC:** $\alpha \in \{2.0, 2.5, 3.0\}$, $\gamma \in \{0.5, 0.8\}$, $\lambda \in \{1.0, 1.5\}$; best: $\alpha = 2.5$, $\gamma = 0.5$, $\lambda = 1.0$.

2 Image Attention Skew

Visual-Token Calibration (VTC) relies on a single “reference” image to derive its calibration vector. A natural concern is whether the choice of that reference—white canvas, black canvas, or random noise—affects the resulting adjustment. To test this, we feed each meaningless image to the LVLm together with the fixed query “Please describe the image.” and compute the relative image-relevancy for all query tokens. Figure 1 shows the results for InstructBLIP and LLaVA. As one can see from the relative image relevancy plots, the choice of the reference image for calibration has no meaningful impact on the calibration vectors.

3 Hyperparameter Analysis

In this subsection, we evaluate the impact of key parameters in the CAAC framework on its performance, focusing on the most influential ones due to limited computational resources. For

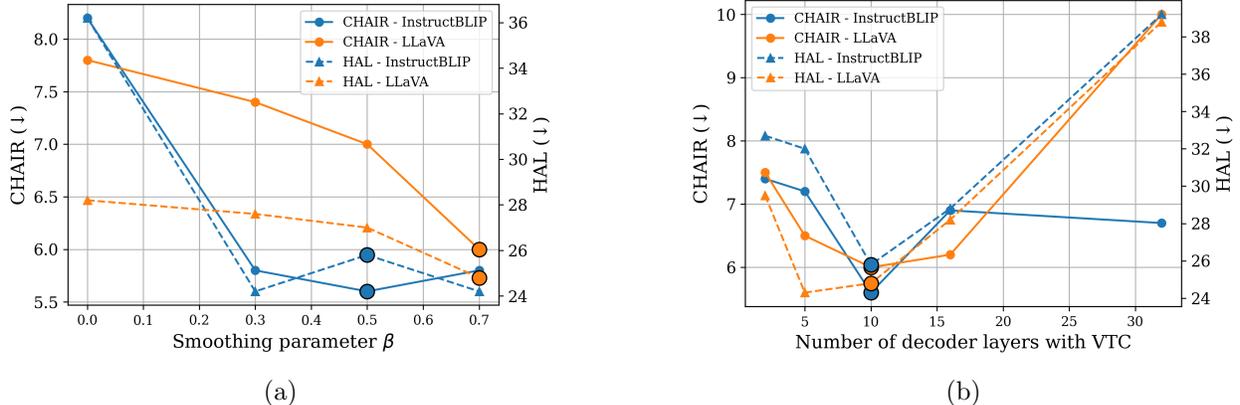


Figure 2: The impact of (a) smoothing parameter β and (b) the number of decoder layers with VTC on the hallucination rate for LLaVA (orange) and InstructBLIP (blue) models. As shown, intermediate values of β , $\tilde{0}.3$ to 0.7 , and the first third of the decoder layers are the optimal parameters for the CAAC framework.

the Adaptive Attention Re-Scaling (AAR) module, we set the confidence threshold $p_{\text{thr}} = 0.25$, as hallucinatory token frequency increases noticeably when the logit probability drops below this value. We also selected $\lambda_{\text{max}} = 1.5$, since values above 2 impair response fluency and coherence. Furthermore, applying AAR to all decoding layers proved optimal, yielding consistent and coherent outputs based on experimental results.

For the Visual-Token Calibration (VTC) module, applying it to all layers often produced incoherent or truncated sequences, likely due to significant changes in attention distribution causing information loss in later layers. We thus examined the effect of varying the number of layers, from the first 2 to all 32 decoder layers. The best performance, with minimal hallucination rates, was observed when VTC was applied to the first 10 layers, as shown in fig. 2b.

We also assessed the smoothing parameter β , testing values from 0 to 0.7. A value of $\beta = 1$ mirrored the issues seen with all-layers VTC application and was excluded. Intermediate values (0.3 to 0.7) reduced hallucination rates most effectively, with $\beta = 0.5$ and $\beta = 0.7$ yielding the best results for InstructBLIP and LLaVA, respectively (fig. 2a). These observations suggest that modest early-to-mid-layer calibration can encourage a balanced attention to the image while allowing the later layers to promote semantically relevant tokens.

4 Qualitative Evaluation

Additional examples of CAAC’s performance compared to baseline mitigation methods on the AMBER dataset are presented in fig. 3 and fig. 4. These cases further illustrate CAAC’s ability to reduce hallucinations across varied image-instruction pairs, with hallucinated content marked in red.

References

- [1] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-Modal Hallucination Control by Visual Information Grounding.

- [2] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation.
- [3] 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.
- [4] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models, May 2024. arXiv:2405.17820 [cs].
- [5] Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. Mitigating Object Hallucinations in Large Vision-Language Models via Attention Calibration, February 2025. arXiv:2502.01969 [cs].



Please describe the image.



InstructBLIP: The image features a close-up of a brown horse sniffing the face of a woman in the grassy field. ... while the horse's black muzzle is placed on top of her forehead. ... There are **a number of trees** visible in the background, providing ...

OPERA: The image features a woman standing in a grassy field next to a brown horse ... possibly nuzzling or sniffing her. ... There are **several trees** visible in the background. ...

VCD: The scene depicts a young woman in a field, standing next to a brown horse. The horse is ... In the background, there **are some trees** and grassy areas, and the environment appears serene and peaceful.

M3ID: The image features a young woman standing in a grassy field with a horse standing nearby. ... There are **two horses** visible in the scene, both positioned close to the woman, one near the center and another closer to the left edge. ...

AvisC: The scene features a woman standing next to a brown horse ... there are **multiple horses** scattered across the field, further showcasing the farm environment. Some horses are positioned ... while others are placed further away ... each other.

CAAC (Ours): The scene features a woman standing in a grassy field, with a horse standing close to her. The horse is positioned near the woman, with its head resting on her shoulder. The woman appears to be petting the horse, creating a warm and friendly atmosphere. The scene is set in a pasture, providing a natural and serene backdrop for the interaction between the woman and the horse.

(a) InstructBLIP: Case 1



Please describe the image.



InstructBLIP: A lone young man is gliding down a sandy slope, enjoying a day on the beach. ... The man is balancing confidently on the beach while enjoying a moment of leisure. The **beach area** in the background is visible, offering ...

OPERA: The image features a man riding a snowboard down a **snow-covered hill**. He is wearing a **blue shirt** and shorts, and he appears to be enjoying himself as he glides down the slope. ...

VCD: The image depicts a man on a skateboard riding down a sandy hill. ... skateboard is visible in the lower half of the image. ... In the background, **a few trees are visible**, likely on the beach ...

M3ID: The image features a person, likely a man, riding down a large sandy slope ... navigating through the **snow-covered surface**. ... **Several snowboards** are scattered throughout the scene. ... **There are six snowboards visible**, spread across the snowy hilltop. ...

AvisC: In the scene, a person is snowboarding down a sandy hill on a sunny day. ... There are at least **nine skis** scattered throughout the scene, perhaps representing **additional snowboarders or left behind equipment** ... the scene showcases the winter fun on the sandy slope ...

CAAC (Ours): The image features a man riding a snowboard down a sandy hill. He is standing on the sandy slope, enjoying the thrill of the ride. The man is wearing down the hill, making the most of the slope's incline. The sky in the background is clear and blue, indicating a sunny day.

(b) InstructBLIP: Case 2



Please describe the image.



InstructBLIP: An image captures a man speaking to his **colleague** at a desk. ... In front of the laptop, there is a chair that has just caught the audience's attention. ... The two seem to be having a light-hearted conversation.

OPERA: The image features a man wearing a suit and tie, sitting at a desk with an open laptop in front of him. ... In the background, **there are two other laptops** visible, one on ... These laptops are not being used by the man sitting at the desk. ...

VCD: The image features a man sitting at a desk with an open laptop computer on it. ... There are **several chairs** placed throughout the scene. ... Additionally, there are **several cups** on the table, possibly used during the presentation. ...

M3ID: The image features a man sitting at a desk with a silver laptop computer open on his lap. ... There are at least **two chairs** placed near the desk. ... Additionally, **two pens** can be spotted, possibly for writing ...

AvisC: There is a man wearing a suit sitting at a table, excitedly using an open Apple laptop on the surface. ... In addition to the man and laptop, there are **two chairs** present near the table. ... **Two televisions** can also be spotted in the scene. ...

CAAC (Ours): The image features a man sitting at a desk, holding a laptop computer. He is wearing a suit and tie, and appears to be excited about something. The man is sitting in front of the laptop, with his hands on the keyboard, making a gesture or expression towards the screen.

(c) InstructBLIP: Case 3

Figure 3: Comparison of CAAC outputs with baseline methods for the InstructBLIP model. Hallucinations are highlighted.



Please describe the image.



LLaVA: The image features a woman standing in a grassy field. ... In the background, there are **two other people** visible, one standing further to the right and another person closer to the center of the scene. The overall atmosphere of the image ...

OPERA: The image features a woman standing next to a brown horse in a grassy field. The horse appears to be nuzzling or sniffing the woman's hair. ... The woman appears to be smiling, enjoying the moment with the horse.

VCD: The image features a woman standing in a grassy field, with a horse gently nuzzling her face. ... In the background, there is **a car** parked, likely belonging to the woman or someone else visiting the field. ...

M3ID: The image features a woman standing next to a brown horse in a grassy field. ... In the background, there is **a stop sign** that can be spotted, suggesting the presence of a nearby road or path.

AvisC: In the image, a young woman is standing outdoors next to a brown horse in a grassy field. ... There are **a few other people** visible in the scene. ... Additionally, there is **a chair** located towards the right side of the image. ...

CAAC (Ours): The image features a woman standing next to a horse in a grassy field. The horse is positioned close to the woman, with its head near her shoulder. The woman appears to be petting the horse, creating a warm and friendly atmosphere. The scene is set in a pasture, providing a natural and serene backdrop for the interaction between the woman and the horse.

(a) LLaVA: Case 1



Please describe the image.



LLaVA: The image features a young boy standing in a grassy field, holding a white soccer ball. ... The scene takes place in a residential area, with a house visible in the background. The yard is surrounded by trees. ...

OPERA: The image features a young boy standing in a grassy field, holding a **white frisbee**. ... playing a game of frisbee with friends or family. The boy is wearing swim trunks, which suggests that the weather **might be warm and suitable** for outdoor activities.

VCD: The image features a young boy standing in a grassy yard, holding a white soccer ball. He appears to be **posing for a picture**. ... The yard is surrounded by trees, providing a pleasant and natural setting ...

M3ID: The image features a young boy standing in a grassy yard, holding a white soccer ball. ... **Another person** is visible in the background. ... **A stop sign** is located near the yard, providing guidance for drivers. ...

AvisC: In the image, a shirtless young boy is standing in a grassy yard, holding a white soccer ball in his hands. ... Around the yard, there are several trees. ... **Another person** can be seen further back. ... Additionally, there are **two benches** ...

CAAC (Ours): The image features a young boy standing in a grassy field, holding a soccer ball. He appears to be enjoying his time outdoors, possibly playing a game of soccer. The field is surrounded by a fence, and there are a few trees in the background. The boy is the main focus of the scene, and his presence adds a sense of playfulness to the image.

(b) LLaVA: Case 2



Please describe the image.



LLaVA: The image features a small, white toilet with a **blue lid** and a **blue dispenser** attached to the wall above it. ... and the dispenser appears to be a **hand dryer**. ... as there is a **sink** visible in the background. The overall scene is clean and well-maintained.

OPERA: The image depicts a small bathroom with a white toilet positioned in the center of the room. ... Additionally, there is a **hand soap dispenser** mounted on the wall next to the toilet, providing convenience for users.

VCD: The image features a small, white toilet built into the floor of a tiled bathroom. ... Next to the toilet, there is a **hand soap dispenser** attached to the wall, ... with the tiled walls and floor adding to the overall aesthetic.

M3ID: The image features a small bathroom with a white toilet sitting on a tiled floor. The toilet has a **blue dispenser** attached to the wall, providing hand soap for users. ...

AvisC: The image features a small tiled bathroom. ... Next to the toilet, there is a **hand soap dispenser**. ... There is also a **sink** located towards the left side of the scene. ...

CAAC (Ours): The image features a small bathroom with a white toilet positioned in the corner of the room. The toilet is surrounded by a white tiled floor. There is a blue toilet paper holder placed above the toilet, adding a touch of color to the otherwise white bathroom. The bathroom appears to be in a hotel, as indicated by the presence of a toilet paper holder.

(c) LLaVA: Case 3

Figure 4: Comparison of CAAC outputs with baseline methods for the LLaVA model. Hallucinations are highlighted.