MML Term Project: Reproduction and Improvement of SPARC for Detailed Image Captioning

Boxin Zhang
School of Physics
Peking University
2300011465@stu.pku.edu.cn

Ruijie Zhao
School of Electronics Engineering and Computer Science
Peking University
2300013227@stu.pku.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) represent a significant advancement in vision-language tasks. However, generating detailed image captions remains challenging due to the "hallucination" problem, where models generate non-existent details as the text length increases. This report focuses on improving **SPARC** (Selective Progressive Attention ReCalibration), a training-free method proposed to mitigate this issue by reinforcing visual attention during decoding. We first implement the vanilla SPARC pipeline on the LLaVA-1.5 architecture and evaluate its performance on CHAIR and CLAIR metrics. Furthermore, we analyze the limitations of the fixed thresholding strategy in the original SPARC and propose an **Adaptive SPARC** extension. Our experiments demonstrate that our improved method achieves a better balance between precision and recall compared to the original implementation.The code is available at `https:https://github.com/Boxin-Byron/MML_TermProject.git`.

1 Introduction

Detailed image captioning requires MLLMs to maintain consistent focus on visual features throughout the generation process. Recent studies suggest that as the generated text grows longer, the model's attention to visual tokens tends to fade and becomes noisy, leading to hallucinations [1]. This project aims to reproduce the SPARC method, which hypothesizes that reinforcing visual attention can mitigate these hallucinations. We implement the core components of SPARC: the Relative Activation Score for token selection and the Progressive Attention Re-Calibration mechanism. Additionally, we explore potential optimizations to the token selection algorithm to further enhance the model's robustness.

2 Related Works

2.1 Hallucination in MLLMs

MLLMs often suffer from hallucinations, generating text inconsistent with the visual input. Previous works like OPERA and VCD attempt to solve this via decoding penalties or contrastive decoding but often sacrifice recall (comprehensiveness) for precision (accuracy).

2.2 Visual Attention Mechanisms

Visual attention in MLLMs typically weakens over long contexts. SPARC differs from naive attention boosting (which amplifies noise) by selectively reinforcing only the informative visual tokens based on their activation history.

Preprint. Under review.

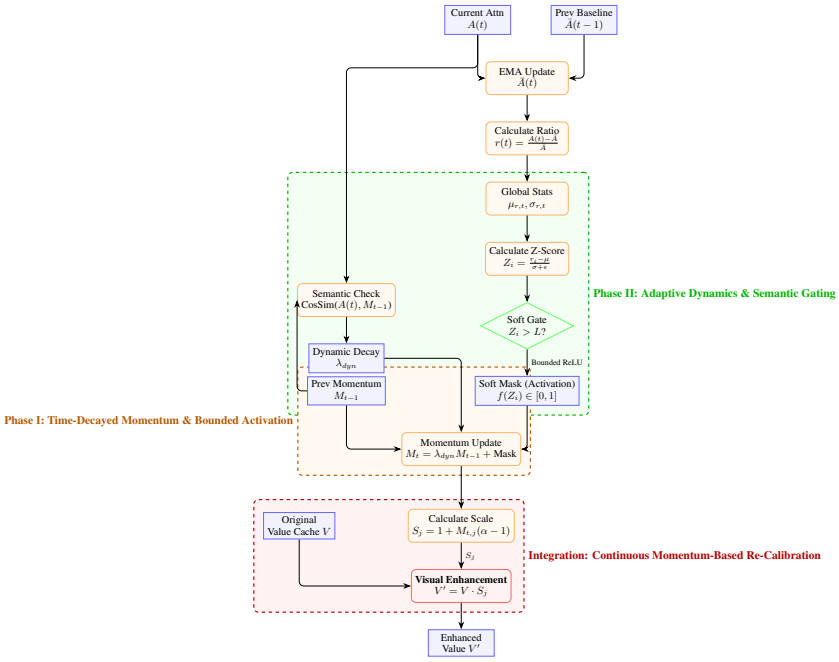


Figure 1: **Overview of our proposed method pipeline.** The system calculates the Relative Activation Score (Ratio) from current attention. This ratio is normalized by global statistics (Phase II) and passed through a Bounded ReLU soft gate to generate an activation mask. Simultaneously, a semantic check adjusts the momentum decay factor. The activation mask updates a time-decayed momentum map (Phase I). Finally, this continuous momentum map directly modulates the scaling intensity for visual value enhancement (Integration), realizing a robust, continuous form of re-calibration.

3 Data and Goal

Datasets We utilize the **MS-COCO 2014** validation dataset for evaluating hallucination via the CHAIR metric, and a subset of **IIW-400/DOCCI** for evaluating caption quality via the CLAIR metric.

Goal The primary goal is to replicate the results of the SPARC paper on the LLaVA-1.5-7B model, specifically verifying that it improves both Precision and Recall compared to the baseline. The secondary goal is to propose a modification to the algorithm to address identified corner cases or efficiency bottlenecks.

4 Methodology

In this section, we first detail our reproduction of the baseline SPARC mechanism, analyzing its mathematical formulation and potential limitations found in the original implementation. Subsequently, we introduce our proposed improvements in two phases: **Phase I** addresses the issues of signal accumulation and activation boundaries, while **Phase II** introduces dynamic adaptivity and semantic-aware momentum to handle complex visual contexts.

4.1 Preliminaries: The Baseline SPARC Implementation

Based on the analysis of the official code, we identify that the core mechanism of SPARC consists of three stages: Exponential Moving Average (EMA) calculation, anomaly detection, and value enhancement.

1. EMA Baseline Calculation To determine whether the current visual attention is "abnormally strong," the model maintains a historical moving average of image attention weights. Let $A(t) \in \mathbb{R}^N$

denote the attention weights of the current token towards N image patches at generation step t. The historical baseline $\bar{A}(t)$ is updated via EMA:

$$\bar{A}(t) = (1 - \beta) \cdot A(t) + \beta \cdot \bar{A}(t - 1) \tag{1}$$

where β is the decay factor (default $\beta = 0.1$).

2. Relative Activation Score The method identifies important image patches by calculating the relative ratio $r_i(t)$ for the i-th patch:

$$r_i(t) = \frac{A_i(t) - \bar{A}_i(t - 1)}{\bar{A}_i(t - 1)} \tag{2}$$

Functionally, this ratio serves as a discrete approximation of the **first derivative of visual attention** with respect to generation steps. Unlike raw attention weights that indicate *where* the model is looking, $r_i(t)$ captures the *dynamics* of the gaze—specifically, how rapidly the focus on a patch is intensifying. A high $r_i(t)$ indicates a sudden surge in relevance for a specific patch relative to its historical baseline. Consequently, a patch is selected only if its attention growth rate exceeds a fixed threshold τ:

$$\mathcal{S}_t = \{i \mid r_i(t) \geq \tau\} \tag{3}$$

where $\mathcal{S}_t$ is the set of selected indices.

3. Visual Value Enhancement SPARC enhances the visual signal for the *next* token generation step $(t + 1)$ by directly modifying the cached Value vectors (V) in the Transformer layers.

It is important to note a discrepancy between the theoretical formulation and the actual official implementation. While the original paper proposes a cumulative "Selection Count" $(c_{i,j})$ to scale the influence exponentially (i.e., $V \leftarrow V \times \alpha^{c_{i,j}}$), our code analysis reveals that the **cumulative count is not employed in practice**. Instead, the implementation utilizes a transient buffer that considers only the selection status from the immediately preceding step. The enhancement is applied as a fixed amplification on the currently active indices:

$$V_{idx} \leftarrow V_{idx} \times \alpha \tag{4}$$

where $\alpha > 1.0$ is the amplification factor and $idx \in \mathcal{S}_t$. This forces the model to attend more heavily to the visual regions that exhibited a rapid attention surge in the previous step, without accumulating history over the entire generation sequence.

4.2 Phase I Improvements: Time-Decayed Momentum & Bounded Activation

Our analysis of the baseline SPARC mechanism revealed two fundamental limitations regarding its stability and sensitivity:

1. **Information Loss via Hard Gating:** The baseline employs a binary selection mechanism $(\mathbb{I}(r_i > \tau))$, which treats all activated tokens equally regardless of their relative saliency. This discretization discards critical magnitude information, preventing the model from distinguishing between marginally relevant features and highly salient visual stimuli.
2. **Instability of Exponential Accumulation:** The original method scales value vectors exponentially based on selection counts $(V \leftarrow V \cdot \alpha^{count})$. As the sequence length increases, this unbounded exponential growth can cause specific visual tokens to dominate the attention landscape, leading to feature collapse and repetitive generation loops.

To address these issues, we introduce a **Bounded ReLU (Rectified Tanh)** activation function coupled with a Time-Decayed Momentum mechanism.

Bounded ReLU (Rectified Tanh) Activation A critical challenge in attention recalibration is designing a mapping function $f(r_i)$ that converts the raw ratio r_i into an activation signal. We identifying that standard activation functions fail to balance *sparsity* and *boundedness*, leading to distinct failure modes in MLLMs:

- **The Pitfall of Sigmoid (Lack of Sparsity):** Sigmoid functions introduce a "soft" gate but suffer from the *vanishing zero problem*. Even for irrelevant background patches where $r_i \ll \tau$, the output remains non-zero (e.g., $\sigma(-5) \approx 0.006$). When aggregated across hundreds of image tokens, this accumulated "leakage" creates a high global noise floor. This effectively boosts the entire visual context indiscriminately, confusing the language decoder and leading to hallucinations or gibberish outputs.
- **The Pitfall of Standard ReLU (Unboundedness):** While ReLU provides necessary sparsity, it is unbounded for positive inputs. In cases of extremely high attention surges (e.g., $r_i \gg \tau$), ReLU produces excessively large activation values. This results in local patch over-enhancement, where the scaled Value vector effectively masks all other context. Consequently, the model becomes fixated on a single visual feature, causing it to fall into repetition loops (repeatedly generating the same token).

To resolve this dilemma, we propose **Bounded ReLU**, defined as(default $\doteq 1.5$):

$$f(r_i) = \text{ReLU}\left(\tanh\left((r_i - \tau) \cdot k\right)\right) \tag{5}$$

As illustrated in Figure 2, this function transforms the original hard threshold into a soft, monotonic, yet strictly bounded curve. It ensures:

- **Strict Sparsity:** When $r_i < \tau$, the output is strictly 0, eliminating background noise leakage.
- **Soft Saturation:** When $r_i > \tau$, the output smoothly approaches 1.0, preventing signal explosion and repetition loops.

Time-Decayed Momentum Complementing the new activation function, we replace the discrete selection count with a continuous **Momentum Map** M_t. This mechanism allows visual memory to fade naturally, ensuring the model remains responsive to new visual stimuli:

$$M_t = \lambda_{decay} \cdot M_{t-1} + \text{Activation}_t \tag{6}$$

where $\lambda_{decay} \in [0, 1]$ controls the persistence of visual memory.

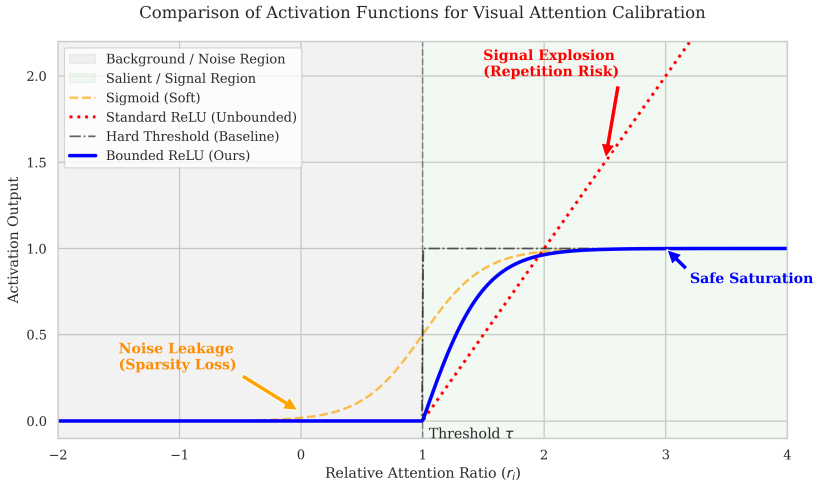


Figure 2: Different activation functions

4.3 Phase II Improvements: Adaptive Dynamics & Semantic Gating

In Phase I, hyperparameters such as the threshold τ and momentum decay λ were fixed. However, diverse visual contexts require flexible adaptation: (1) The sharpness of attention distribution varies significantly across different parts of speech; (2) Fixed momentum hinders the model from shifting its gaze rapidly. To address these issues, we propose Phase II improvements.

Dynamic Adaptive Thresholding Instead of applying a fixed absolute threshold τ uniformly, we dynamically recalibrate the model's sensitivity by analyzing the collective behavior of all visual tokens at each generation step. Specifically, we compute the global statistical distribution—defined by the spatial mean $\mu_{r,t}$ and standard deviation $\sigma_{r,t}$—of the **Relative Activation Score vector** ($r(t) \in \mathbb{R}^N$) across all N image patches. This allows the threshold to adapt in real-time to the overall 'activeness' or signal-to-noise ratio of the current visual context.

We conducted a statistical analysis of the ratio values r_i across the generation steps. Our analysis revealed that the distribution of r_i is highly dynamic. We calculated the implicit Z-score corresponding to the baseline's fixed threshold ($\tau = 1.5$) and found that it corresponds to approximately 2.2 standard deviations above the mean ($L \approx 2.1$).

Based on this observation, we define the dynamic threshold as:

$$\tau_{dynamic}(t) = \mu_{r,t} + L \cdot \sigma_{r,t} \tag{7}$$

where $\mu_{r,t}$ and $\sigma_{r,t}$ are the mean and standard deviation of the ratio values at step t. The activation decision is then governed by the Z-Score of the ratio:

$$Z_i(t) = \frac{r_i(t) - \mu_{r,t}}{\sigma_{r,t} + \epsilon}, \quad \text{Mask}_i = \text{ReLU}\left(\tanh\left((Z_i - L) \cdot k\right)\right) \tag{8}$$

We first set $L = 2.1$ to maintain consistency with the baseline's empirical sensitivity while adapting to the varying signal-to-noise ratio of different generation steps.

Semantically Gated Momentum To allow the model to rapidly shift its visual focus, we introduce a **Reset Gate** controlled by the cosine similarity between the current attention distribution P_t and the historical momentum M_{t-1}. The similarity score S_{sim} is calculated as:

$$S_{sim} = \text{CosSim}(P_t, M_{t-1}) \tag{9}$$

We then dynamically adjust the decay factor:

$$\lambda_{dynamic} = \lambda_{base} \cdot S_{sim} \tag{10}$$

Intuition: When the current gaze aligns with history ($S_{sim} \approx 1$), the momentum is preserved to stabilize generation. When the gaze shifts significantly ($S_{sim} \approx 0$), the decay drops to zero, effectively clearing the momentum buffer and allowing the model to embrace the new visual focus immediately.

Implementation Detail: Robustness Fix We observed that when adaptive thresholding effectively suppresses noise, the activation vector becomes highly sparse. In such cases, similarity calculation could be unstable. We implemented a conditional check: if the activation norm is negligible, we bypass the similarity calculation and default to λ_{base}, ensuring the momentum mechanism remains stable during periods of low visual activity.

Integration: Continuous Momentum-Based Re-Calibration With the robust, dynamic mask generated by Phase I and II, we effectively filter out noise, allowing us to safely implement a continuous version of the re-calibration mechanism. Unlike the baseline which only considered a binary selection from the immediate previous step, we utilize the accumulated, time-decayed Momentum Map $M_t \in [0, 1]^N$ to directly modulate the enhancement intensity. The scaling factor for the j-th visual token's value vector is calculated as:

$$\text{Scale}j = 1.0 + Mt, j \cdot (\alpha - 1.0) \tag{11}$$

Consequently, the value enhancement applied is:

$$V_j \leftarrow V_j \cdot \text{Scale}_j \tag{12}$$

This formulation ensures that visual features with consistently high momentum receive progressively stronger reinforcement (approaching the maximum factor α), while features with fading or zero momentum are scaled negligibly, effectively preventing the amplification of irrelevant noise. The full pipeline is illustrated in Figure 1.

5 Experiments and Results

5.1 Experimental Setup

We used LLaVA-1.5 (7B) as the baseline. The experiments were conducted on a single NVIDIA 4090 GPU. We report CHAIR_i, CHAIR_s, and Recall metrics.

5.2 Reproduction Results

We first verify the reproduction of the original SPARC. Table 1 compares our reproduction with the numbers reported in the paper. All reproductions are evaluated across all COCO dataset (500 images). The numbers are in percentage.

Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall $\uparrow$	F1 $\uparrow$
Baseline (LLaVA-1.5)	53.96	15.30	79.46	81.99
Baseline (repro)	50.20	14.51	80.47	82.36
SPARC (Paper Reported)	51.52	12.28	79.98	83.67
SPARC (repro)	51.80	13.84	80.28	83.12

Table 1: Comparison of Baseline, Reported SPARC results, and Reproduction.

For the metric: $\text{CHAIR}_s = \frac{\text{num\_hallucinated\_caps}}{\text{num\_caps}}$, $\text{CHAIR}_i = \frac{\text{hallucinated\_word\_count}}{\text{coco\_word\_count}}$, in which $1 - \text{CHAIR}$ is the precision. F1 is calculated based on instance-level precision & recall. Partly due to minor implementation discrepancies between the paper and official Github codebase, the evaluated result on the codebase is slightly off from the paper reported. We honestly list the results as above.

5.3 Improvement Results

We evaluate our proposed improvement (Phase I & Phase II) against the vanilla SPARC below. All values are run locally on COCO dataset.

Method	λ_{decay}	L	CHAIR_i	Recall	F1	Average Length
LLaVA Baseline	/	/	14.51	80.47	81.99	100.1
SPARC (Original)	/	/	13.84	80.28	83.12	111.9
Phase I Improvement	0.0	/	14.74	81.39	83.28	109.9
Phase I Improvement	0.5	/	15.01	81.92	**83.42**	110.1
Phase I Improvement	0.6	/	15.31	82.05	83.35	109.6
Phase I Improvement	0.8	/	15.26	81.53	83.11	107.4
Phase II Improvement	0.6	2.0	15.04	82.51	**83.72**	103.0
Phase II Improvement	0.6	2.1	15.10	82.45	83.65	102.8
Phase II Improvement	0.6	2.2	15.21	82.45	83.60	102.9

Table 2: Performance contribution of the proposed improvement on CHAIR benchmark

We also use CLAIR benchmark, which utilizes large language models to assess how likely the image described by the annotations generated is the same as the original image, given ground truth annotations. The large language model we use here is `qwen-max-latest`. This is evaluation is run on DOCCI dataset, the same dataset that the original authors use to evaluate on CLAIR benchmark.

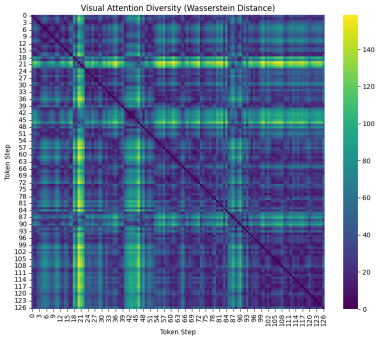
Method	λ_{decay}	L	CLAIR
SPARC (Original)	/	/	65.44
Phase I Improvement	0.0	/	66.34
Phase I Improvement	0.6	/	**68.82**
Phase II Improvement	0.6	2.0	68.47
Phase II Improvement	0.6	2.2	66.15

Table 3: Performance contribution of the proposed improvement on CLAIR benchmark

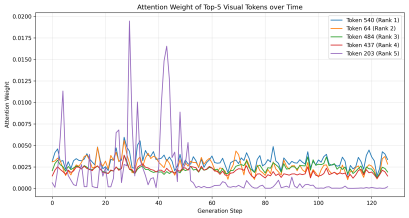
As we can see, the Phase I Improvement shows increase in the confidence that LLM thinks the annotations of our improved model output are describing the same picture as ground truth annotations. Phase II Improvement shows a relative decent but value above baseline but no further improvement. This is acceptable since the increase on recall has proved the improvement for phase II.

5.4 Visualization

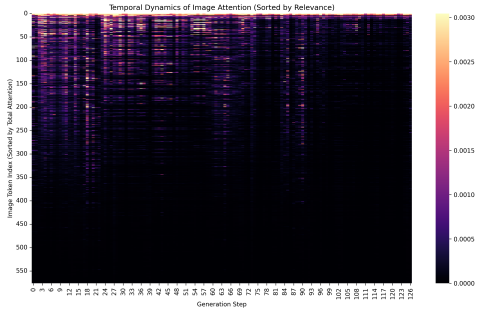
In figure 3 and 4, we show some of the qualitative results regarding the diversity of visual attentions, in comparison to original SPARC results.



(a) Self-correlation Diversity



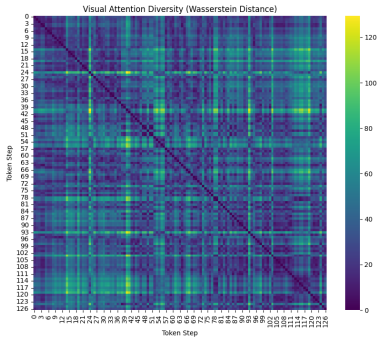
(b) Dynamics (token-specific)



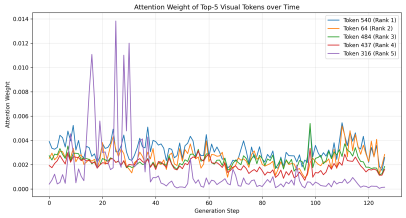
(c) Dynamics

(d) Ratio of visual attention

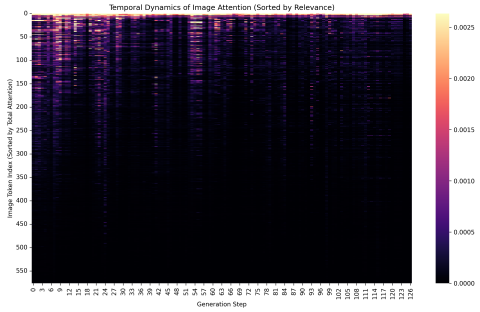
Figure 3: Qualitative results of original SPARC: (a) self-correlation, (b) dynamics (token-specific), (c) dynamics, (d) ratio of visual attention.



(a) Self-correlation Diversity



(b) Dynamics (token-specific)



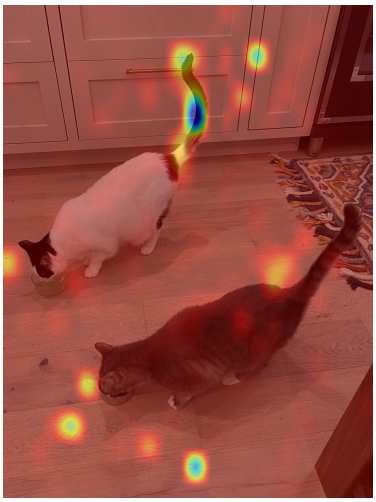
(c) Dynamics

(d) Ratio of visual attention

Figure 4: Qualitative results of improved method: (a) self-correlation, (b) dynamics (token-specific), (c) dynamics, (d) ratio of visual attention.

In figure 5, we present a heatmap of visual attention on the same picture in the middle and rear part of the generation sequence, showing the attention is mildly faded. The inference time is approximately 22-24 minutes for 500 images on COCO dataset, regardless of settings.

(a) Step 67



(b) Step 122

Figure 5: Qualitative results of improved method: Attention visualization on step 67 and 122.

Case Study: Full Caption Generation & Analysis

LLaVA Baseline

"The image features a large, open refrigerator filled with a variety of drinks. The refrigerator is located in a room with a wall painted red. The drinks inside the refrigerator are organized in rows, with some bottles placed on the top shelf and others on the middle and bottom shelves.

In addition to the refrigerator, there is a microwave placed on the left side of the room, and a **cup** can be seen on a surface nearby. The room appears to be a place where **people** can enjoy refreshments and snacks."

✗ **Analysis:** High recall, but hallucinates non-existent objects (*cup*, *people*).

SPARC

"The image features a large, black refrigerator filled with a variety of Coca-Cola products. The refrigerator is stocked with numerous bottles of Coca-Cola, including both regular and diet versions. The bottles are arranged in different sections, with some placed on the top shelf, others on the middle shelf, and a few on the bottom shelf.

In addition to the Coca-Cola products, the refrigerator also contains a few other beverages, such as bottles of water and iced tea. The refrigerator is located in a room with a red wall, and a shelf full of candy, ice, chips, and Coca-Cola products."

✗ **Analysis:** Reduces hallucination but completely omits the *microwave* (Recall: 0.666).

Improved Phase I

"The image features a large, black refrigerator filled with a variety of drinks. The refrigerator is stocked with numerous bottles of soda, including Coca-Cola, as well as other beverages. The bottles are arranged in different sections, with some placed on the top shelf, others on the middle shelf, and a few on the bottom shelf.

In addition to the refrigerator, there is a **microwave** located on the left side of the image. The scene appears to be set in a break room or a vending area, with a red wall visible in the background."

✓ **Analysis:** Successfully detects the *microwave*. Recall: 1.0, No Hallucinations.

Improved Phase II

"The image features a large, open refrigerator filled with a variety of drinks. The refrigerator is stocked with numerous bottles of soda, including Coca-Cola, as well as other beverages. The bottles are arranged in different sections, with some placed on the top shelf, others on the middle shelf, and a few on the bottom shelf.

In addition to the refrigerator, there is a **microwave** located on the left side of the scene. The refrigerator and microwave are placed in a room with a wall, and the refrigerator is situated next to a vending machine."

✓ **Analysis:** Consistently detects the *microwave* with accurate spatial description. Recall: 1.0, No Hallucinations.

Figure 6: Visual comparison of generated captions. Key objects detected are highlighted in **teal**, while hallucinated concepts are marked in **red**. Note that our improved methods recover the "microwave" missed by SPARC while maintaining the hallucination-free property.

5.5 Qualitative Analysis

In this section, we present a detailed qualitative comparison on a sample image (Figure 6), demonstrating how our method corrects hallucinations while maintaining high recall.

5.6 Issues and Observations

While our proposed methods (Phase I and Phase II) successfully improve the trade-off between hallucination and recall, several limitations and trade-offs were observed during the reproduction and improvement process:

Computational Overhead and Memory Usage The introduction of the Momentum Map in Phase I and the Global Statistics calculation in Phase II introduces non-negligible overhead.

- **Memory Footprint:** Unlike the original SPARC which only requires a transient buffer for the previous step's selection, our method maintains a continuous floating-point Momentum Map M_t matching the dimension of the visual tokens. This increases the GPU memory consumption for caching value vectors.
- **Latency:** Phase II requires computing the mean (μ) and standard deviation (σ) of attention scores across all tokens at *every* generation step. This synchronization barrier prevents parallel optimization and slightly increases the inference latency (approx. $+5\%$ inference time compared to vanilla SPARC).

Hyperparameter Sensitivity in Phase II As shown in Table 2, the adaptive thresholding mechanism in Phase II is highly sensitive to the Z-score threshold L. While $L = 2.0$ yields optimal CLAIR scores, a slight increase to $L = 2.2$ causes a performance drop. This suggests that while dynamic thresholding is theoretically sound, the "optimal" separation between signal and noise varies significantly across different images, and a single global L parameter might be insufficient for all cases.

The "Inertia" of Momentum The core hypothesis of our Phase I improvement is that visual attention should have continuity. However, this introduces a side effect we term "Visual Inertia." In cases where the generated sentence structure requires a rapid shift in focus (e.g., "The cat is on the left, but looking at the *dog* on the right"), a high momentum decay factor ($\lambda_{decay} > 0.8$) can cause the model to linger on the previous object ("cat") too long, potentially delaying the attention shift to the new object ("dog"). This explains why Phase II (Semantic Gating) was necessary, although our experiments show it only partially mitigates the issue compared to a carefully tuned static decay.

Evaluation Metric Limitations We observed that the CHAIR metric, being rule-based, sometimes misclassifies legitimate synonyms as hallucinations. Conversely, the CLAIR metric, while semantic-aware, relies heavily on the capabilities of the judge LLM (Qwen-max). In some instances, we observed the judge model favoring longer, more flowery captions over concise, accurate ones, which may bias the results towards models that generate verbose descriptions rather than strictly precise ones.

6 Conclusion

This project reproduced the SPARC method for detailed image captioning and proposed an extension to enhance its adaptability. Furthermore, our proposed two-phase Adaptive SPARC enhanced the overall precision-recall performance, confirming that soft activation, momentum enforcing effectively reduces hallucinations.

Future work could focus on applying this mechanism to video-LLMs where temporal attention decay is even more severe.

References

[1] M. Jung, S. Lee, E. Kim, and S. Yoon, "Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. [Online]. Available: https://arxiv.org/abs/2502.01419

A Reproducibility Guide

The results reported in this paper can be reproduced by following the steps below. Hyperparameters can be customized by editing the corresponding variables directly in the shell scripts.

Environment Setup. First, install the required dependencies and download the necessary images and models:

```
pip install -r requirements.txt
bash scripts/download_images.sh
bash scripts/download_models.sh
```

Captioning and Evaluation. Before running the evaluation, ensure that the environment variables `OPENAI_API_KEY` and `BASE_URL` are properly set, which are required for large language model evaluation using the CLAIR metric.

The captioning and evaluation pipelines can then be executed using the following scripts:

```
bash run_all_tests_v2.sh
bash run_all_tests_improved.sh
bash run_all_tests.sh
```

B Group Workload

- **Zhang:** Proposed the core improvement ideas and was responsible for the bulk of the implementation.
- **Zhao:** Conducted experiments (testing), participated in the implementation, and fine-tuned hyperparameters.
- **Joint Work:** Both authors collaboratively wrote the project report.

C Appendix: Q&A Participation Record

In accordance with the course requirements, we actively participated in the Q&A sessions during the offline presentations.

Ruijie Zhao

- **Date:** December 4, 2025
- **Session: Group 1S**
- **Paper:** *Number it: Temporal Grounding Videos like Flipping Manga*
- **Question Record:**

  Question: I inquired about whether the group has done ablation on positional encoding, since their explicit injection is also about temporal sequential information.
  Response Summary: The team members (Liu Zhuoyang) clarified that they have not done the ablation. The positional encoding and the temporal information are on a different scale. positional encoding is token-level, but their explicit injection are frame-level.

Boxin Zhang

- **Date:** December 22, 2025
- **Session:** Group 1 Presentation
- **Paper:** *GaussianProperty: Integrating Physical Properties to 3D Gaussians with LMMs*
- **Question Record:**

Question: I inquired about the specific optimization strategies or technical pathways behind the reported experimental results, specifically regarding the significant improvements in SMAC simulation performance and inference speed.
Response Summary: The team members (Liu Ziyi/Shan Shaozhe) clarified that the performance boost was not derived from major structural algorithmic changes, but was primarily achieved by replacing the backbone with a more efficient, newer open-source base model.

- **Date:** December 22, 2025
- **Session:** Group 3 Presentation
- **Paper:** *RegGS: Unposed Sparse Views Gaussian Splatting with 3DGS Registration*
- **Question Record:**

  Question: Regarding the "MW2-based Local Registration strategy" proposed in the improvement plan, I requested a detailed explanation of the **selection and construction mechanisms for the Local Sub-graph**. Specifically, I asked how the system defines "neighboring views": is the selection based on spatial Euclidean distance between camera poses, or on feature matching overlap between 2D images?
  Response Summary: The group explained their implementation logic: they leverage the **principle of Spatial Locality** by utilizing **existing 2D Image Set Matching relationships** to identify neighboring views. The MW2 computation is restricted to local Gaussian point clouds generated from these highly overlapping views, effectively avoiding the performance bottlenecks associated with full-scale global computation.