

KVSmooth: Mitigating Hallucination in Multi-modal Large Language Models through Key-Value Smoothing

Supplementary Material

A. The Detailed Derivation of EMA Smoothing on Hidden States

During inference, this growth manifests as increasingly large token-wise logit moments: the model state oscillates violently, drifts away from the visual prior, and deviates from the true image information through sudden jumps. To keep the decoding trajectory ideally smooth, we formalize the following hypothesis:

$$h_t = h_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_p^2). \quad (16)$$

Consequently, the transition prior is

$$P(h_t|h_{t-1}) = N(h_t; h_{t-1}, \sigma_p^2) \propto \exp\left(-\frac{1}{2\sigma_p^2}\|h_t - h_{t-1}\|^2\right). \quad (17)$$

Under this prior, the maximum-a-posteriori (MAP) estimate of h_t is given by

$$\hat{h}_t = \operatorname{argmax}_{h_t} \left[\underbrace{\log P(o_t|h_t)}_{\text{likelihood}} + \underbrace{\log P(h_t|h_{t-1})}_{\text{smoothed prior}} \right], \quad (18)$$

where o_t denotes the raw hidden state observed at step t . Assuming both terms are Gaussian, we have

$$P(o_t|h_t) \propto \exp\left(-\frac{1}{2\sigma_o^2}\|o_t - h_t\|^2\right), \quad (19)$$

$$P(h_t|h_{t-1}) \propto \exp\left(-\frac{1}{2\sigma_p^2}\|h_t - h_{t-1}\|^2\right). \quad (20)$$

Substituting (19) and (20) into (18) yields

$$\hat{h}_t = \operatorname{argmax}_{h_t} \left[-\frac{1}{2\sigma_o^2}\|o_t - h_t\|^2 - \frac{1}{2\sigma_p^2}\|h_t - h_{t-1}\|^2 \right]. \quad (21)$$

Taking the derivative of (21) with respect to h_t and setting it to zero, we obtain:

$$\frac{\partial}{\partial h_t} \left[-\frac{1}{2\sigma_o^2}\|o_t - h_t\|^2 - \frac{1}{2\sigma_p^2}\|h_t - h_{t-1}\|^2 \right] = 0 \quad (22)$$

$$\frac{1}{\sigma_o^2}(o_t - h_t) - \frac{1}{\sigma_p^2}(h_t - h_{t-1}) = 0.$$

Solving for \hat{h}_t , we obtain

$$\hat{h}_t = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2} o_t + \frac{\sigma_o^2}{\sigma_p^2 + \sigma_o^2} h_{t-1}. \quad (23)$$

Defining

$$\lambda_t = \frac{\sigma_o^2}{\sigma_p^2 + \sigma_o^2}, \quad (24)$$

the estimate reduces to the exponential moving average (EMA) form

$$\hat{h}_t = (1 - \lambda_t)o_t + \lambda_t h_{t-1}. \quad (25)$$

Thus, when the likelihood is Gaussian in h_t , the MAP estimator coincides exactly with an exponential moving average.

B. Benchmark Details

We briefly describe the four benchmarks used for evaluation:

CHAIR. CHAIR [29] is a classic benchmark for measuring **object hallucination** in image captioning tasks. It defines two key metrics:

- **CHAIR_S** (Sentence-level Hallucination) — the proportion of captions that contain at least one hallucinated object.
- **F₁** — a balanced measure reflecting both the accuracy and completeness of generated captions. In CHAIR, *precision* quantifies the proportion of generated objects that correctly appear in the ground-truth annotations, while *recall* measures how many ground-truth objects are successfully mentioned in the generated captions. The F₁ score captures the overall trade-off between these two factors.

In this work, following [12], we randomly sample 500 images from the COCO 2014 validation set and prompt the LVLMs with “Please describe the image in detail.” The maximum output length is set to 512 tokens.

OPOPE. OPOPE [4] extends POPE [20] by transforming its interactive yes/no polling mechanism into an **offline evaluation**. It retains POPE’s three sampling strategies — *random*, *popular*, and *adversarial* — but checks whether sampled positive and negative objects appear in the model-generated image descriptions instead of interacting with the model. Following [4], we report the following metrics:

- **Accuracy** — overall correctness of object presence identification.
- **Precision** — proportion of correctly identified positive objects among all predicted positives.

- $F_{\beta=0.2}$ — defined as

$$F_{\beta} = (1 + \beta^2) \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \quad \beta = 0.2,$$

which reduces the impact of false negatives following [4]. All reported numbers are averaged over the three sampling methods (random, popular, and adversarial).

AMBER. AMBER [35] is a benchmark for evaluating hallucinations in vision-language models from both generative and discriminative perspectives. In this work, we focus on the generative task setting to assess model performance. The dataset encompasses 14 major object categories, featuring a balanced distribution that mitigates the significant long-tail issue. Compared to existing benchmarks, AMBER extends coverage to categories such as Nature, Architecture, and Street View, and provides richer annotations in others—for example, the Fruit category includes over a dozen common fruits, whereas prior datasets only cover three types.

Following [35], we adopt four metrics for evaluation:

- **CHAIR** — object hallucination rate in responses.
- **Cover** — proportion of objects mentioned in the response relative to annotated objects, reflecting faithful object coverage.
- **Hal** — proportion of responses containing hallucinations.
- **Cog** — measures the extent to which hallucinated objects align with common human cognitive biases.

Object HalBench. Object HalBench (ObjHal) [29] is a diverse benchmark for evaluating **object hallucination robustness** under different prompt styles. It assesses models on 300 image–text pairs using eight varied prompts, providing a stable and comprehensive evaluation. Here, GPT-4 [27] is used to extract visible objects from the captions generated by the model.

The benchmark reports four main metrics:

- **CHAIR_S** — Image-level hallucination rate: proportion of images containing at least one hallucinated object.
- **CHAIR_{SR}** — Similar to CHAIR_S but excludes sentences without any object words; measures hallucination among sentences containing at least one MSCOCO object word.
- **CHAIR_I** — Proportion of hallucinated objects among all generated object words.
- **CoCoNum** — Number of captions containing at least one COCO object.

C. Efficiency Comparisons

To further assess the computational efficiency and resource usage of our approach, we compare its average per-caption

Table 5. Comparison of token generation efficiency and resource usage across different methods on LLaVA-1.5 (CHAIR benchmark). Our method delivers faster inference and lower memory cost than other training-free hallucination mitigation methods.

Method	Avg. Time (s/caption)	Peak Memory (MB)	Latency (ms/token)	Throughput (token/s)
Baseline	3.36	<u>14629.21</u>	31.24	32.06
PAI	6.68	15003.25	56.44	17.73
OPERA	34.62	19531.97	313.78	3.43
SPARC	4.13	14644.61	35.20	28.44
Ours	<u>3.61</u>	14625.06	<u>34.33</u>	<u>29.18</u>

inference time, token-level latency, throughput, and memory consumption with several representative hallucination mitigation methods, as summarized in Table 5.

As illustrated in the table, our method achieves a favorable trade-off between performance and efficiency. It maintains a comparable runtime and memory footprint to the baseline model while significantly reducing hallucinations. Notably, compared with more complex attention redistribution approaches such as SPARC and PAI, our approach requires substantially less computation time and memory overhead. This demonstrates that our smoothing mechanism is lightweight and can be seamlessly integrated into existing vision-language models without compromising efficiency.

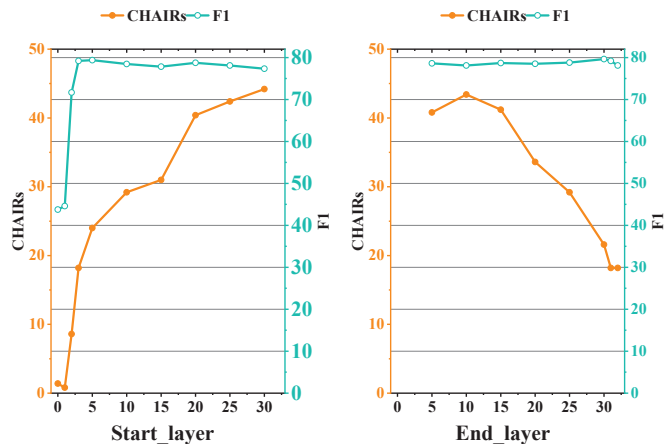


Figure 6. Sensitivity analysis of the layer range for KVSmooth on LLaVA-1.5 (CHAIR benchmark). Left: sensitivity to start layer (L_{start}); Right: sensitivity to end layer (L_{end}).

D. Parameter Sensitivity

To evaluate the robustness of our approach, we conduct a sensitivity analysis on key hyperparameters across three MLLMs, including the EMA decay rate λ_{ref} , which determines the strength of the smoothing effect, the start layer

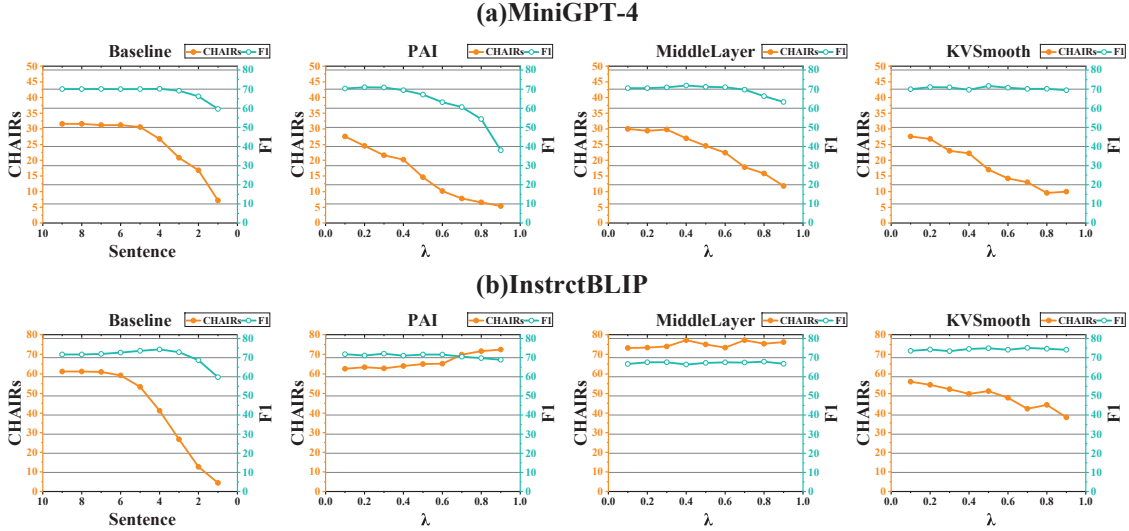


Figure 7. Sensitivity analysis of the hyperparameter λ_{ref} for KVSmooth based on MiniGPT-4 (a) and InstructBLIP (b) and comparisons of four methods in terms of the CHAIR_S-F₁ trade-off (CHAIR benchmark). Larger values of λ_{ref} lead to stronger smoothing and more effective hallucination mitigation. Moreover, our method consistently maintains a favorable balance between precision and recall: it reduces hallucinations while preserving a high F₁ score, demonstrating diverse and reliable object coverage across different smoothing strengths.

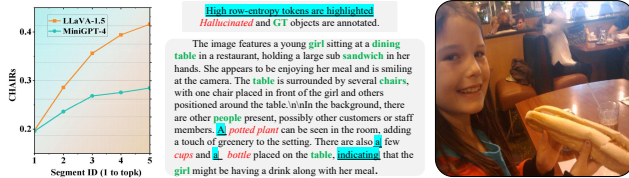


Figure 8. Temporal analysis. Left: quantitative results; right: illustrative examples.

(L_{start}) and end layer (L_{end}), which define the range of layers where EMA is applied, and the queue length (Q_{len}), denoting the maximum length of the queue. We vary each parameter while keeping the others fixed to examine its influence on overall performance.

D.1. Effect of the Reference Smoothing Coefficient

$$\lambda_{\text{ref}}$$

Specifically, we analyze the influence of the decay threshold range by varying λ_{ref} and evaluating KVSmooth on the CHAIR benchmark, measuring both CHAIR_S and F₁ metrics. As shown in Figure 5 and Figure 7, the CHAIR_S score consistently decreases as λ_{ref} increases, while the F₁ score remains nearly unchanged. This indicates that a larger λ_{ref} leads to stronger smoothing effects and more effective hallucination mitigation. Moreover, despite enhanced hallucination suppression, KVSmooth maintains a stable balance between precision and recall, demonstrating robustness across different smoothing strengths. **Reference**

coefficient(λ_{ref}). λ_{ref} is not selectively picked. The selected [0.5, 0.9] of λ_{ref} were designed to ensure a **fair comparison** with baselines under identical CHAIR_S levels. As a control variable, a higher λ_{ref} corresponds to a lower hallucination degree. For any unseen model, it can be flexibly adjusted based on the specific requirements for hallucination suppression. As shown in Manuscript Figure 5 and Appendix Figure 7, regardless of the targeted hallucination constraint, KVSmooth consistently achieves the **highest F₁** among all methods at equivalent CHAIR_S levels.

D.2. Effect of Start and End Layers ($L_{\text{start}}, L_{\text{end}}$).

To determine the optimal layer range for applying EMA updates, we conduct a two-dimensional layer sweep. First, we fix the end layer L_{end} at 31 and vary the start layer L_{start} from 0 to 30. Then, we fix L_{start} at 3 and vary L_{end} from 5 to 32. As shown in Figure 6, CHAIR_S decreases gradually as more layers are included in the EMA updates, demonstrating stronger hallucination suppression. However, applying EMA at the very early (0–2) or final (32) layers causes a noticeable drop in F₁, indicating that excessive smoothing in these regions impairs the model’s ability to represent real objects and degrades caption quality.

D.3. Effect of Queue length (Q_{len}).

Table 6 shows stable performance across $Q_{\text{len}} \in [10, 80]$. Longer queues slightly reduce CHAIR_S but may over-smooth late tokens as attention spreads over more tokens, thereby degrading F₁. Thus, we choose 15 as the recom-

Table 6. Sensitivity analysis of the $Qlen$ on LLaVA-1.5.

$Qlen$	10	15	20	30	40	50	60	70	80
$C_S \downarrow$	20.6	18.2	18.6	17.2	15.2	15.6	15.0	15.2	15.0
$F_1 \uparrow$	<u>79.4</u>	79.2	79.0	<u>79.4</u>	79.9	78.7	78.5	78.7	78.7

Table 7. CHAIR performance under different models and decoding strategies.

Method	Qwen2.5-VL		Decoding Strategy (LLaVA-1.5)			
			Sample		Beam-3	
	CHAIR _S ↓	F ₁ ↑	CHAIR _S ↓	F ₁ ↑	CHAIR _S ↓	F ₁ ↑
Baseline	16.0	70.1	48.0	74.1	49.6	<u>77.4</u>
PAI	<u>15.4</u>	69.0	<u>36.2</u>	70.3	<u>36.8</u>	76.8
Ours	14.6	<u>69.1</u>	33.6	74.1	29.4	79.2

Table 8. MMHal-Bench performance of LLaVA-1.5. Lower CHAIR_S and higher F₁ indicate better performance. The best results in each column are highlighted in **bold**.

Method	Avg. ↑	Attr. ↑	Adv. Obj. ↑	Comp. ↑	Count. ↑	Spatial ↑	Env. ↑	Hol. ↑	Other ↑
Baseline	2.87	3.42	3.08	3.50	2.58	3.17	2.83	2.00	2.27
Ours	3.15	3.75	3.58	3.92	2.67	3.25	2.92	2.55	2.45

mended default for new models, balancing effectiveness and overhead.

E. Row-Entropy as a Temporal Precursor of Hallucination

Figure 8 visualizes the accumulation of CHAIR_S within caption segments that are demarcated by tokens exhibiting high row-entropy, and provides an illustrative case study. The figure shows that CHAIR_S tends to increase after the occurrence of high row-entropy tokens, suggesting that elevated row-entropy often precedes the amplification of hallucinated content.

This observation indicates a temporal precursor relationship between row-entropy and hallucination. In this perspective, increased row-entropy reflects an early signal of potential semantic drift during generation. The proposed smoothing mechanism therefore functions as a preventive stabilizer, mitigating such drift and guiding the generation process back toward semantically consistent states.

F. Supplementary Experiments

F.1. Decoding Strategy

To evaluate robustness beyond greedy decoding, we follow the decoding protocol of OPERA: we use top- p sampling with $p = 0.9$. Due to GPU memory constraints during rebuttal, we use beam size 3 instead of 5 in OPERA, which remains a standard setting and preserves non-greedy search behavior.

As shown in Table 7, our method achieves the lowest CHAIR_S and highest F₁ under both sampling and beam search on LLaVA-1.5. This confirms that our gains are not an artifact of greedy decoding but generalize to diverse inference settings.

F.2. Qwen2.5-VL

As shown in Table 7, we evaluated KVSmooth on **Qwen2.5-VL-7B** along with the baseline and PAI. Our method achieves a lower CHAIR_S of **14.6**, improving over both PAI (15.4) and the baseline (16.0), while also obtaining a slightly higher F₁ (**69.1**) than PAI (69.0).

F.3. MMHal-Bench

As shown in Table 8, our method improves over the baseline on **MMHal-Bench** across all eight hallucination types, including **Attribute** (3.75 vs. 3.42) and **Spatial Relation** (3.25 vs. 3.17). This confirms effectiveness beyond object-centric errors. Regarding smoothing’s impact on reasoning, our gains on MME’s Cognitive score (+38.2, from 292.9 to 331.1) which includes relational and logical tasks—suggest no degradation, instead, semantic consistency enhances fine-grained reasoning.

G. The Algorithm of KVSmooth

In this section, we present the full algorithmic details of KVSmooth. The algorithm consists of two components: **Algorithm 1: EMA Smoothing on KV-Cache**, corresponding to Section 4.3 (*The Final KVSmooth Method*); and **Algorithm 2: Attention Forward with Adaptive EMA Smoothing**, corresponding to Section 3.1 (*Formulation of MLLMs Generation*) and Section 4.2 (*Entropy-Guided Coefficient Adaptation*).

Algorithm 1: EMA Smoothing on KV-Cache

Input: KV-cache $C_t^l = (K_{1:t}^{(l,h)}, V_{1:t}^{(l,h)})_{h=1}^H$, smoothing coefficient $\tilde{\lambda}_t^l$

Output: Smoothed KV-cache $C_t^l = (K_{1:t}^{(l,h)}, V_{1:t}^{(l,h)})_{h=1}^H$

1. EMA update for each head h :

$$\hat{K}_t^{(l,h)} \leftarrow (1 - \tilde{\lambda}_t^l) K_t^{(l,h)} + \tilde{\lambda}_t^l K_{t-1}^{(l,h)},$$

$$\hat{V}_t^{(l,h)} \leftarrow (1 - \tilde{\lambda}_t^l) V_t^{(l,h)} + \tilde{\lambda}_t^l V_{t-1}^{(l,h)}.$$

2. Update smoothed cache:

$$K_t^{(l,h)} \leftarrow \hat{K}_t^{(l,h)}, \quad V_t^{(l,h)} \leftarrow \hat{V}_t^{(l,h)}.$$

return C_t^l

Algorithm 2: Attention Forward with Adaptive EMA Smoothing

Input: Hidden state h_t , mask M , KV-cache C_{t-1}^l , entropy queue S^l , reference coefficient λ_{ref}

Output: Updated hidden state h_t^l , attention weights $\alpha_t^{(l,h)}$, KV-cache C_t^l , entropy queue S^l

1. **Compute Q, K, V for each head h :**

$$Q_t^{(l,h)} = h_t W_Q^{(l,h)}, \quad K_t^{(l,h)} = h_t W_K^{(l,h)}, \quad V_t^{(l,h)} = h_t W_V^{(l,h)}.$$

2. **Update KV-cache:** Append $(K_t^{(l,h)}, V_t^{(l,h)})$ to C_{t-1}^l to obtain C_t^l .

3. **Compute attention scores:**

$$A_{t,j}^{(l,h)} = \frac{(Q_t^{(l,h)})^\top K_j^{(l,h)}}{\sqrt{d}} + M_{t,j}.$$

4. **Compute attention weights and row-entropy:**

$$\alpha_{t,j}^{(l,h)} = \text{Softmax}_j(A_{t,j}^{(l,h)}),$$
$$z_{t,j}^l = -\frac{1}{H} \sum_{h=0}^{H-1} \sum_{i=0}^{L-1} \alpha_{i,j}^{(l,h)} \log(\alpha_{i,j}^{(l,h)} + \varepsilon).$$

Insert $z_{t,j}^l$ into the entropy queue S^l .

5. **Determine adaptive smoothing coefficient:** Compute the percentile rank k of $z_{t,j}^l$ in S^l .

$$\hat{\lambda}_t^l = k / |S^l|,$$
$$\tilde{\lambda}_t^l = \text{clip}(\hat{\lambda}_t^l, \lambda_{\text{ref}} - 0.2, \lambda_{\text{ref}} + 0.2).$$

6. **Compute attention output:**

$$o_t^{(l,h)} = \sum_{j=0}^{L-1} \alpha_{t,j}^{(l,h)} V_j^{(l,h)},$$
$$h_t^l = \text{Concat}_h(o_t^{(l,h)}) W_O^l.$$

7. **Apply EMA smoothing for each head:**

$$(K_{1:t}^{(l,h)}, V_{1:t}^{(l,h)})_{h=1}^H \leftarrow \text{EMASmoothingon}((K_{1:t}^{(l,h)}, V_{1:t}^{(l,h)})_{h=1}^H, \tilde{\lambda}_t^l).$$

return $h_t^l, \alpha_t^{(l,h)}, C_t^l, S^l$

H. Qualitative Results

To illustrate the effectiveness of KVSmooth, we show representative examples where **orange** denotes baseline hallucinations and **cyan** denotes the corrections by KVSmooth, demonstrating its ability to fix factual errors while preserving coherence.



Baseline

The image features a large, shiny orange motorcycle on display at a show. The motorcycle is prominently positioned in the center of the scene, drawing attention to its unique design. There are several people surrounding the motorcycle, admiring and observing it. Some of them are standing closer to the motorcycle, while others are positioned further away. In addition to the motorcycle, there are a few bottles scattered around the scene, possibly indicating that the event is a casual gathering or a social event.

KVSmooth

The image features a large, shiny orange motorcycle on display at a show, likely an exhibition or a convention. The motorcycle is prominently displayed, occupying a significant portion of the scene. There are several people surrounding the motorcycle, with some standing closer to the front and others further back. In addition to the main motorcycle, there are a few other motorcycles visible in the background, indicating that this event showcases various motorcycles. The scene captures the attention of the visitors, as they gather around to admire the impressive display.



Baseline

The image features a man wearing a coat and holding a sandwich in his hand. He appears to be eating the sandwich while standing outside. There are several other people in the scene, some of them carrying backpacks and handbags. In the background, there are a few cars parked, and a bench can be seen nearby. The scene seems to be a busy outdoor area with people going about their daily activities.

KVSmooth

The image features a man wearing a tan jacket, standing outside in the snow. He is holding a sandwich in his hand, likely enjoying a bite of his meal. The man appears to be focused on his food, possibly taking a break from the cold weather. There are several other people in the scene, some of them standing in the background, possibly walking by or engaging in their own activities. A backpack can be seen in the background, possibly belonging to one of the people in the scene.



Baseline

The image features two women in a room, playing a video game together. One woman is standing and holding a Wii remote, while the other woman is sitting on a bed, also holding a Wii remote. They appear to be enjoying their time together, possibly playing a game like Wii Sports. The room has a couch in the background, and a laptop is placed on the bed. There are also two potted plants in the room, one near the couch and the other closer to the standing woman. The scene captures a fun and interactive moment between the two women.

KVSmooth

The image features two women in a room, with one woman standing and another person sitting on a bed. The woman standing is holding a Wii remote, likely playing a video game. The woman on the bed is sitting on a bed, watching and possibly observing the gameplay. The room appears to be a cozy living space, with a couch in the background and a laptop placed on a surface nearby. The scene suggests that the two women are enjoying a fun and interactive gaming session together.

Figure 9. Qualitative results on LLaVA-1.5: baseline (w/o KVSmooth) vs. KVSmooth.



Baseline

The image shows a large clock hanging from a pole on the side of a building. The clock has a black face and white numbers, and is surrounded by green leaves. The building behind the clock is tall and made of stone. There are several signs on the pole, including a street sign pointing to the right and a sign for a business on the left. The sky is clear and blue, and there are a few clouds in the distance.

KVSmooth

The image shows a large clock hanging from a pole on the side of a building. The clock has a black face and roman numerals. The clock is surrounded by street signs pointing in different directions. The building behind the clock is a tall, modern building with many windows and a large, ornate facade. The sky is clear and blue with a few clouds in the sky.



Baseline

The image shows a long train traveling down a railroad track. The train is made up of several cars, each carrying different types of cargo. The train appears to be carrying a mix of goods, including grains, minerals, and other materials. The track is lined with trees and other vegetation, and there is a clear blue sky in the background. The image is taken from a high angle, looking down on the train as it moves along the track.

KVSmooth

The image shows a long train traveling down a railroad track. The train is yellow and red, with yellow cargo cars. There are trees in the background. The image is taken from above, looking down on the train as it goes along the track.



Baseline

The image shows a group of people in a living room. There is a cake on a table in front of them, and they are all wearing party clothes. One person is holding a knife and is about to cut the cake. Another person is holding a glass of champagne. The room is dimly lit, and there are several pieces of furniture in the background.

KVSmooth

The image shows a group of people in a living room. There is a young woman standing in front of a cake on a table, holding a knife in her hand. She is wearing a green shirt and jeans. There is a man standing behind her, watching her. The room is dimly lit and there is a television in the background.

Figure 10. Qualitative results on MiniGPT-4: baseline (w/o KVSmooth) vs. KVSmooth.




Baseline

The image features a young woman wearing a pink sweater and a skirt standing in a kitchen. She is posing for the camera, and her hair is styled in a ponytail. The kitchen is equipped with various appliances, including a sink, an oven, and a refrigerator. There are also several bottles and bowls placed around the kitchen, suggesting that the woman is preparing food or drinks. In addition to the kitchen appliances and food items, there are several other objects scattered throughout the scene, including a chair, a table, and a book. The overall atmosphere of the image is lively and welcoming. The woman's ponytail and the appliances in the kitchen give the impression of a cozy and functional space. The presence of the book suggests that the woman may be engaging in some form of learning or entertainment. Overall, the scene captures a young woman in a cheerful and productive kitchen setting.

KVSmooth

The image features a young woman standing in a kitchen, wearing a pink sweater and a skirt. She is holding her hand on her hip and posing for the camera. In the background, there are several cabinets visible, including one on the left side of the room and another on the right side. Various bottles can be seen scattered around the kitchen, with some located near the cabinets and others placed on the countertops. There are also two bowls present in the scene, one on the countertop to the left and another on the countertop to the right. The overall atmosphere of the kitchen appears to be cozy and inviting.




Baseline

The image shows a luxurious bedroom with a large bed in the center. The bed is adorned with two folded towels that resemble elephants, sitting on top of the bed as if they were guarding the room. There are several lamps placed around the bed, providing ample lighting for the space. A chair can be seen on the right side of the bed, and a second chair is positioned on the left side. In addition to the towels and lamps, there are several other items in the room, including a book placed on the bed, a lamp on the left side, and a second book placed on the right side. The room appears to be well-appointed and inviting, with a comfortable atmosphere for guests to relax and enjoy their stay.

KVSmooth

The image shows a luxurious bedroom with a large, neatly made bed. Two folded towels are placed on the bed in the shape of two elephants, creating an eye-catching display. There are several lamps positioned around the room, including one on each side of the bed and multiple cushions can be seen scattered across the surface of the bed. A chair is also present in the room, placed close to the bed. The arrangement of the bedding and decorations creates a cozy and inviting atmosphere for guests or visitors to enjoy. The towel elephant display adds a playful touch to the otherwise elegant bed setup, making it a unique and memorable feature of the room. Overall, the bedroom appears to be well-appointed and welcoming.



Baseline

The image features a table set up with a white plate containing a sandwich, which appears to be a grilled cheese sandwich. The sandwich is placed on top of the plate and is surrounded by various ingredients, such as tomatoes and lettuce. There are two bottles of juice on the table, one on the left side and the other on the right side. A glass of punch is also visible, placed in front of the sandwich. A knife is placed near the plate, likely for cutting the sandwich. A bowl is situated next to the plate, possibly containing additional ingredients or condiments. In the background, there is a refrigerator and a sink, suggesting that this is a kitchen setting. Overall, the scene showcases a well-prepared and appetizing meal.

KVSmooth

The image features a table set up with a white plate containing a sandwich, a glass of juice, and a bottle of water. The plate is placed on a countertop, and there are two bottles visible in the background. One bottle is located on the left side of the image, while the other is on the right side. The plate is adorned with various vegetables, including tomatoes and lettuce. The sandwich appears to be a hearty and flavorful meal. The table also includes a bottle of juice, which is likely to complement the sandwich. The setup suggests that the person enjoying the meal is taking a break from their daily routine. Overall, the scene showcases a well-prepared and satisfying meal.

Figure 11. Qualitative results on InstructBLIP: baseline (w/o KVSmooth) vs. KVSmooth.