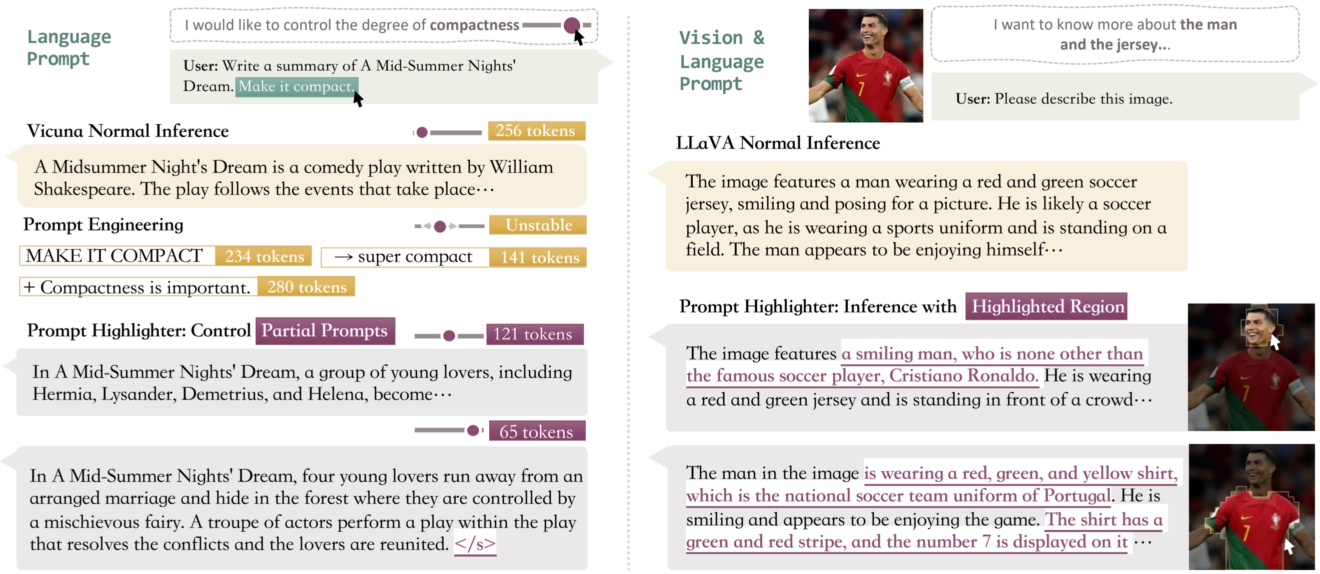
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Prompt Highlighter: Interactive Control for Multi-Modal LLMs

Yuechen Zhang[1] Shengju Qian[1] Bohao Peng[1] Shu Liu[2] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore
https://github.com/dvlab-research/Prompt-Highlighter


Language Prompt
I would like to control the degree of compactness
Vicuna Normal Inference
256 tokens
Prompt Engineering
Unstable
MAKE IT COMPACT
234 tokens
→ super compact
141 tokens
+ Compactness is important.
280 tokens
Prompt Highlighter: Control Partial Prompts
121 tokens
65 tokens
Vision & Language Prompt
I want to know more about the man and the jersey...
LLaVA Normal Inference
Prompt Highlighter: Inference with Highlighted Region

User: Write a summary of A Mid-Summer Nights' Dream. Make it compact.

A Midsummer Night's Dream is a comedy play written by William Shakespeare. The play follows the events that take place···

In A Mid-Summer Nights' Dream, a group of young lovers, including Hermia, Lysander, Demetrius, and Helena, become···

In A Mid-Summer Nights' Dream, four young lovers run away from an arranged marriage and hide in the forest where they are controlled by a mischievous fairy. A troupe of actors perform a play within the play that resolves the conflicts and the lovers are reunited. </s>

User: Please describe this image.

The image features a man wearing a red and green soccer jersey, smiling and posing for a picture. He is likely a soccer player, as he is wearing a sports uniform and is standing on a field. The man appears to be enjoying himself···

The image features a smiling man, who is none other than the famous soccer player, Cristiano Ronaldo. He is wearing a red and green jersey and is standing in front of a crowd···

The man in the image is wearing a red, green, and yellow shirt, which is the national soccer team uniform of Portugal. He is smiling and appears to be enjoying the game. The shirt has a green and red stripe, and the number 7 is displayed on it ···

Figure 1. Prompt Highlighter facilitates token-level user interactions for customized generation, compatible with both LLMs and VLMs. Compared with vanilla inference and prompt engineering, the context-highlighted inference provided by our method offers controllable generations and produces customized results. Outputs correlated with the highlighted parts are underlined.

Abstract

This study targets a critical aspect of multi-modal LLMs' (LLMs&VLMs) inference: explicit controllable text generation. Multi-modal LLMs empower multi-modality understanding with the capability of semantic generation yet bring less explainability and heavier reliance on prompt contents due to their autoregressive generative nature. While manipulating prompt formats could improve outputs, designing specific and precise prompts per task can be challenging and ineffective. To tackle this issue, we introduce a novel inference method, Prompt Highlighter, which enables users to highlight specific prompt spans to interactively control the focus during generation. Motivated by the classifier-free diffusion guidance, we form regular and unconditional context pairs based on highlighted tokens, demonstrating that the autoregressive generation in models can be guided in a classifier-free way. Notably, we find that, during inference, guiding the models with highlighted tokens through the attention weights leads to more desired outputs. Our approach is compatible with current LLMs and VLMs, achieving impressive customized generation results without training. Experiments confirm its effectiveness in focusing on input contexts and generating reliable content. Without tuning on LLaVA-v1.5, our method secured 70.7 in the MMBench test and 1552.5 in MME-perception.

1. Introduction

Large Language Models (LLMs) have driven significant progress in a multitude of natural language processing tasks [1–9]. Further advancements have been achieved by extending these models to handle vision-language tasks [10–15] through visual-language alignment and instruction tuning. These efforts have led to the development of Vision Language Models (VLMs), which can generate text based on multi-modal inputs. Due to its autoregressive nature, the typical generation process in LLMs and VLMs (multi-modal LLMs) is primarily conditioned on input contexts. Prompt engineering [16–19] has emerged as a common interaction mechanism between humans and language models, where diverse formats and content of prompts are employed to steer the generation towards desired outcomes. However, prompt engineering often relies on empirical intuition and requires careful design of the context, making it less accessible for non-experts. As illustrated in the left part of Fig. 1, even the meticulously crafted prompts, which convey the concept of 'compactness' clearly, can lead to unpredictable outputs that fail to meet the requirements.

Instead of manipulating prompt-level contexts (*i.e.*, prompt engineering) to control LMs' generation process, we propose a novel inference approach, Prompt Highlighter, that enables token-level user interactions for personalized generations. Our method allows users to interact with multi-modal LLMs in a manner analogous to applying a highlighter tool on the input context in the text editor, enabling users to emphasize desired parts by highlighting them.

This highlighting mechanism is achieved by constructing a regular and unconditional input context pair with different textual embeddings in the highlighted tokens. Subsequently, we can adjust the model's focus on the highlighted components by employing the classifier-free guidance [20–22] on predicted token probabilities. Moreover, by probing cross-token attention maps, we discover a robust correlation between attention scores and the semantic significance of tokens. This suggests that, in the autoregressive generation process of language models, the semantic relationship between tokens can be represented to a certain extent by their attention scores. Building on this insight, we introduce an attention activation strategy that adjusts the attention weights associated with a highlighted part. Specifically, Prompt Highlighter employs an adjusted attention mask to reweight corresponding attention scores, enabling a more focused generation on highlighted parts.

As illustrated in Fig. 1, compared to vanilla inference, our highlighted inference can guide the generation process to produce controllable results that align more closely with user needs. Prompt Highlighter is compatible with mainstream transformer-based multi-modal LLMs. This compatibility encompasses VLMs that use precise patch-wise visual token mapping, such as LLaVA [10, 23, 24], as well as methods that employ implicit query-based visual token mapping, like those based on Q-Former [11, 13–15]. This novel interaction paradigm with highlighted sections during the generation process goes beyond what prompt engineering can offer.

We further demonstrate the effectiveness of Prompt Highlighter by evaluating it using comprehensive multi-modal benchmarks. We verify that directly highlighting the full image context in VLMs can significantly improve the quality of generated image captions [25] and question-answering results. Specifically, our method can effectively mitigate the model's propensity to hallucinate by guiding its focus toward reliable contexts, thereby enhancing overall performance. Notably, without additional training, our method improves the performance of the baseline LLaVA-v1.5, securing 2nd place in both MMBench [26] and MME-perception [27] leaderboards.

Our contributions can be summarized as follows: (1) We pioneer the exploration of fine-grained human-model interactions in multi-modal LLMs, proposing a plug-and-play pipeline that enables token-level user interactions for controllable generation. (2) We conduct extensive experiments on comprehensive benchmarks, demonstrating that our method significantly enhances the overall performance.

2. Related Works

2.1. Multi-Modal LLMs

Recent Large Language Models (LLMs) [1, 7–9, 28–30] play a significant role in natural language processing tasks, particularly in language generation and question answering. Building upon these pre-trained language models, Vision-Language Models (VLMs) [10, 11, 13–15, 31] further introduce the alignment between vision and language modalities by leveraging extensive training on image-caption pairs or image-question conversations. There are two prevalent methods for aligning vision and verbiage modalities. The first method, exemplified by LLaVA [10], directly maps image patches to tokens using a projector, establishing a one-to-one correspondence. The second method, represented by models like BLIP2 [13, 32], employs a Query Transformer (Q-Former) after getting image features to establish a non-uniform patch-token mapping. These methods use learnable queries to get compressed image features, yielding visual tokens rich with semantic information.

2.2. Interactions with Multi-Modal LLMs

Prompt engineering and interactions. Based on the autoregressive property of LLMs, users aim to control the generation results by modifying the input contexts. This largely determines the test-time interactions with LLMs, primarily executed through prompt engineering. Representative methods such as CoT [17] introduce demonstrations in the context to enhance reasoning ability. Other multi-branch designs like ToT and GoT [16, 18, 19, 33, 34] have been proposed for rich and reliable context generation and self-checking. Aside from prompt engineering, human-model interactions have not been extensively explored in VLMs. Methods like Kosmos-2 [31], LLaVAInteractive [35], LISA [36], and AlphaCLIP [37] enable grounding perception tasks such as detection, segmentation, caption, and image editing through interaction with LLMs. These task-oriented interactions require additional data collection and task-specific tuning. In contrast, Prompt Highlighter is plug-and-play for general text generation in pre-trained models.

Classifier-free guidance and controllable generation. Classifier-Free Guidance (CFG) [20] enables a control on Diffusion Models' generation process without a conven-

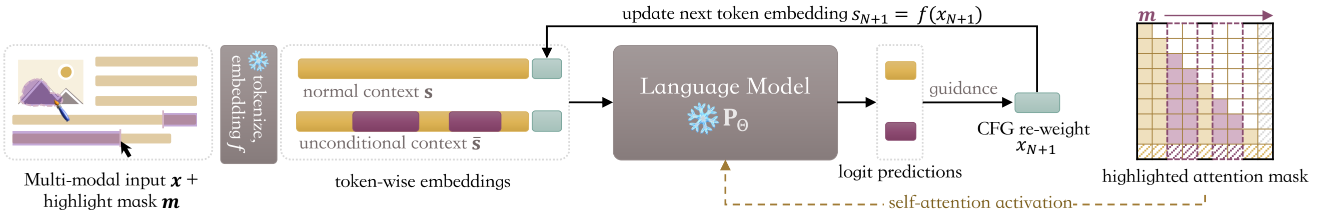

update next token embedding $s_{N+1} = f(x_{N+1})$
tokenize, embedding f
normal context $\mathbf{s}$
unconditional context $\bar{\mathbf{s}}$
Language Model
$\mathbf{P}_\Theta$
guidance
CFG re-weight
x_{N+1}
$\boldsymbol{m}$
Multi-modal input $\boldsymbol{x}$ +
highlight mask $\boldsymbol{m}$
token-wise embeddings
logit predictions
highlighted attention mask
self-attention activation

Figure 2. An abstract pipeline of Prompt Highlighter. Users can control the focus of generation by marking out specific image regions or text spans. Then a token-level mask $\mathbf{m}$ is created to guide the language model's inference.

tional classifier. Specifically, CFG's step-wise sampling allows users to employ a negative prompt within the unconditional branch, effectively guiding the generation away from harmful distributions. This approach has been extended to language models by LLM-CFG [21], allowing a controllable text generation and improved performance. However, LLM-CFG still requires a pair-wise prompt design and does not support partial token-level reweighting within the context, which is vital for controlling VLM's generation. Besides, methods in Diffusion Models [38, 39] achieve fine-grained control over image generation using text prompts by emphasizing areas within cross-attention maps. Fine-grained control over autoregressive generation in LLMs and VLMs is still challenging. Later concurrent works CRG and MARINE [40, 41], adopt CFG in VLMs for grounding and mitigating hallucination, but employ a different design for positive-negative pairs compared to our approach.

3. Prompt Highlighter

An overview of Prompt Highlighter is presented in Fig. 2. Given a pre-trained generative model $\mathbf{P}_\Theta$, we first extract the input tokens from the text and the input image, forming the prompt context $\mathbf{x}$. Subsequently, by marking out specific image regions or text spans, user creates a token-level binary mask $\mathbf{m}$ to highlight specific tokens. Prompt Highlighter then generates the output sequence $\mathbf{y}$ using the two-branch condition based on $\mathbf{m}$ autoregressively. The following section will delve into the specifics of our method.

3.1. Token-Level Highlight Guidance

In conditioned Diffusion Models [42], given a noisy image x and a class condition c, the model predicts probability likelihood $\hat{\mathbf{P}}$, for the conditioned step-wise sample, $\hat{\mathbf{P}}_\Theta(x|c) \propto \mathbf{P}_\Theta(x) \cdot \mathbf{P}_\Phi(c|x)^\gamma$. Here, $\mathbf{P}_\Phi$ is a classifier, and γ is the guidance strength controlling the weight of likelihood on c. Ho et al. [20] observed that guidance can be offered without a classifier. Applying the Bayes rule, $\mathbf{P}_\Theta(c|x) \propto \mathbf{P}_\Theta(x|c)/\mathbf{P}_\Theta(x)$, the sampling process of the Classifier-Free Guidance (CFG) can be expressed as

$$\hat{\mathbf{P}}_\Theta(x|c) \propto \mathbf{P}_\Theta(x|c)^\gamma / \mathbf{P}_\Theta(x)^{\gamma-1}, \tag{1}$$

$$\begin{aligned}\log \hat{\mathbf{P}}_\Theta(\epsilon_t|x_{t+1}, c) &= \gamma \log \mathbf{P}_\Theta(\epsilon_t|x_{t+1}, c) \\ &\quad - (\gamma - 1)\log \mathbf{P}_\Theta(\epsilon_t|x_{t+1}),\end{aligned} \tag{2}$$

in which ϵ_t is the noise prediction conditioned on the previous output x_{t+1} and the text condition c. LLM-CFG [21] extended this property to autoregressive language models. Given a sequence of N tokens $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood of predicting the entire sequence can be expressed as $\mathbf{P}_\Theta(x) = \prod_i^N \mathbf{P}_\Theta(x_i|x_{j<i})$. The model samples each subsequent token from the conditional probability distribution. Based on Eq. (1), the CFG sampling on the language model can be denoted as

$$\hat{\mathbf{P}}_\Theta(\mathbf{x}|c) \propto \frac{\mathbf{P}_\Theta(\mathbf{x}|c)^\gamma}{\mathbf{P}_\Theta(\mathbf{x})^{\gamma-1}} \propto \prod_{i=1}^{N} \frac{\mathbf{P}_\Theta(x_i|x_{j<i}, c)^\gamma}{\mathbf{P}_\Theta(x_i|x_{j<i})^{\gamma-1}}. \tag{3}$$

Similar to the transaction from Eq. (1) to Eq. (2), the likelihood in LLM is represented as the next-token classification probability. Thus next token's logit prediction $\mathbb{P}x_i = \log \hat{\mathbf{P}}_\Theta(x_i|x_{j<i}, c)$ is

$$\mathbb{P}x_i = \gamma \log \mathbf{P}_\Theta(x_i|x_{j<i}, c) - (\gamma - 1)\log \mathbf{P}_\Theta(x_i|x_{j<i}). \tag{4}$$

The formulation in Eqs. (3) and (4) offers a paradigm for controllable generation in LLMs [21], with the guidance strength γ controls the degree of generation focus. Notably, the effectiveness of this guidance depends on the careful design of the conditional prompt c, which should be naturally formed as a complete phrase or sentence to retain its semantic meaning. Prompt Highlighter extends CFG control in language models in a more generalized manner. The user's selection on the context $\mathbf{x}$ is converted into a token-level binary highlight mask $\mathbf{m} = \{m_1, \ldots, m_N\}$. We define $m_i = 1$ if the i-th token x_i is highlighted, and $m_i = 0$ otherwise. This mask constructs a two-branch condition: the normal and the unconditional contexts. The normal context operates in the same manner as in vanilla inference. Meanwhile, the unconditional context $\bar{\mathbf{s}}$ is derived from the normal conditional context $\mathbf{s} = \{s_1, \ldots, s_N\}$ within the textual embedding space through a token-wise scaling,

$$\bar{s}_i = (\alpha - 1)m_i \cdot f(x_i) + f(x_i), \tag{5}$$

where α is the scaling factor and $f(\cdot)$ is the token-to-embedding function, *i.e.*, $s_i = f(x_i)$. We empirically set a small rescale α (*e.g.*, 0.01) that can ensure a normal inference while ignoring the highlighted part. Then, based on the two-branch condition $(\mathbf{s}, \bar{\mathbf{s}})$, we can define the i-th token sampling process of the token-level highlight guidance as

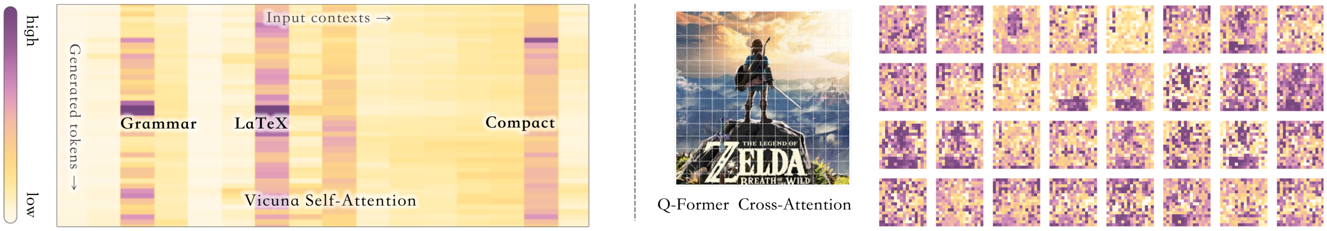


Figure 3. Visualizing attention maps. ***Left***: A segment of the attention map between the generated tokens and the input requirement prompt: "…fix the grammar and keep LaTeX format, make it compact…". Some representative tokens are marked for reference. ***Right***: Query-based token mapping. This shows the attention score on 32 queries in the first cross-attention layer of the Q-Former.

$$\begin{aligned}\log \hat{\mathbf{P}}_\Theta(x_i|s_{j<i}) =& \gamma \log \mathbf{P}_\Theta(x_i|s_{j<i}) \\ & - (\gamma - 1) \log \mathbf{P}_\Theta(x_i|\bar{s}_{j<i}).\end{aligned} \tag{6}$$

Compared with Eq. (4), the additional conditional context c is naturally incorporated as the difference between $\mathbf{s}$ and $\bar{\mathbf{s}}$. This arrangement provides users with the flexibility to control the in-line requirements. After the i-th token is predicted, the highlight mask $\mathbf{m}$ and contexts $\mathbf{s}, \bar{\mathbf{s}}$ are updated by appending $m_i = 0$ and $s_i = \bar{s}_i = f(x_i)$, respectively. The generation process is terminated when the end token </s> is predicted.

3.2. Attention Activation

The token-level highlight guidance anchors the generative process with a token-wise logit reweighting. However, its effectiveness may diminish when facing long and complex input contexts with few highlighted tokens, making it difficult to distinguish $\mathbf{s}$ and $\bar{\mathbf{s}}$. To further investigate token-wise correlations and their impact on generation results, we exclude sink tokens that dominate the attention score [43] and visualize cross-token self-attention score maps during inference. For instance, in the left of Fig. 3, pivotal tokens form a band-like pattern on the attention map, drawing attention from nearly all following tokens. This pattern endures with changes in the model's layer number or attention heads, suggesting the attention mechanism's consistency and robustness in well pre-trained LLMs [7, 8]. Meanwhile, it implies that attention scores within the model can represent the semantic correlation between tokens.

When addressing diverse requirements, LLMs need to balance attention among multiple tokens. For instance, as seen in Fig. 3 (left), as one of the requirements in the prompt, 'compactness' might not get enough attention during the generation process, resulting in an output that is less compact than expected. Given the direct correlation observed between attention and tokens, we propose an attention activation strategy to activate the attention scores on highlighted tokens within the attention mechanism. This strategy can effectively steer the model's focus towards or away from specific tokens, allowing for more nuanced and precise control over the output. We reformulate the model inference function in Eq. (6) to a mask-conditioned one $\tilde{\mathbf{P}}_\Theta(x_i|s_{j<i}, \mathbf{m})$. In each of its self-attention layers, let $\mathbf{k}_i$ represent the i-th column vector of the query-key multiplied attention score matrix in one attention head. The activated attention score h_i ***in the normal context branch*** is defined as

$$h_i = \log(\beta) \cdot m_i + k_i. \tag{7}$$

Then, the attention probability p_i is calculated as

$$p_i = \frac{\exp(h_i)}{\sum_{j=1}^{N} \exp(h_j)} = \frac{\beta^{m_i} \cdot \exp(k_i)}{\sum_{j=1}^{N} \beta^{m_j} \cdot \exp(k_j)}. \tag{8}$$

This mechanism defines the activation scaling factor as β. ***For the unconditional branch***, the attention score is deactivated by using a scaled negative mask in the inference $\tilde{\mathbf{P}}_\Theta(x_i|\bar{s}_{j<i}, -\delta\mathbf{m})$. Eq. (8) presents a SoftMax probability $p_i = \mathrm{softmax}(h_i)$ on the activated scores, with a consistent β^{m_i}-times probability augmentation on highlighted tokens. The attention activation operates under the assumption that users cannot highlight dominant 'sink' tokens as explored in [43]. Consequently, the attention activation will not catastrophically impair the model's fundamental generative capabilities during the inference.

3.3. Highlighting Visual Tokens

Methods for highlighting visual tokens can be classified into two categories based on the type of token mapping involved, as discussed in Sec. 2.1. In direct token mapping, such as in LLaVA [10], the highlighting of visual tokens is straightforward. Image patch-level feature forms the sequential visual contexts $\mathbf{s}^{\mathrm{im}}$. This enables a natural patch-level scaling on embeddings in Sec. 3.1 and attention activation introduced in Sec. 3.2.

In contrast, the scenario becomes more complex with query-based token mapping. For example, in works like BLIP-2 [11, 13, 14], the image feature is transformed into a unified set of few learnable queries $\mathbf{q}$ via Q-Former, which are then input into LLMs as textual embeddings. This process obscures the direct correlation between image patches and input tokens, as demonstrated in the example on the right side of Fig. 3. To address this challenge, we leverage the fact that the Q-Former itself is a Transformer model to perform the token highlighting directly inside it. First, we

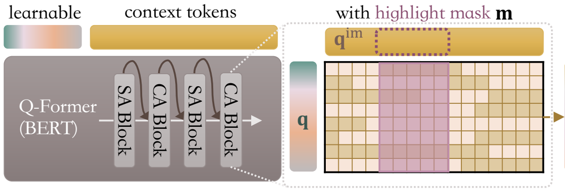
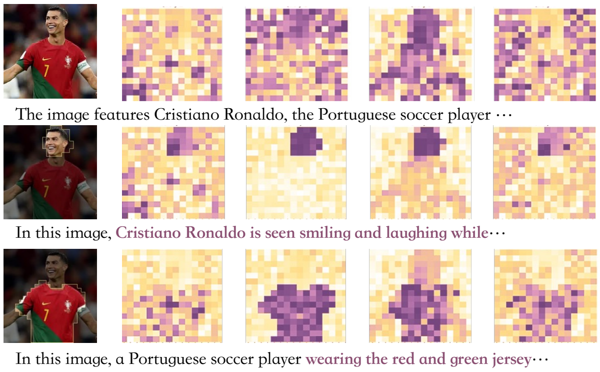


Figure 4. Highlighting visual tokens with Q-Former-based methods. In comparison with vanilla inference, we augment the learnable queries $\mathbf{q}$ by activating corresponding attention weights in the Cross-Attention (CA) blocks.



The image features Cristiano Ronaldo, the Portuguese soccer player ⋯

In this image, Cristiano Ronaldo is seen smiling and laughing while⋯

In this image, a Portuguese soccer player wearing the red and green jersey⋯

Figure 5. Attention scores in the first four queries of the Q-Former. Each row shows a different user selection and text output.

adopt the embedding rescale in Eq. (5) on the patch-wise image feature and use the output queries from the Q-Former to form the image context pair $(\mathbf{s}^{\text{im}}, \bar{\mathbf{s}}^{\text{im}})$. Then, by activating attention scores within the corresponding patch-wise user-selection mask $\mathbf{m}$ in Q-Former's cross-attention layers, we can effectively steer the model to adjust its focus. This process is depicted in Fig. 4. In the Q-Former model, cross attention is calculated across the learnable query $\mathbf{q} = \{q_1, \dots, q_M\}$ and the image feature $\mathbf{q}^{\text{im}} = \{q_1^{\text{im}}, \dots, q_N^{\text{im}}\}$. We then activate the attention score corresponding to the mask $\mathbf{m}$ within the cross-attention map. It can be expressed using a formulation similar to Eq. (7),

$$\hat{\mathbf{q}} = \text{softmax}\left((\log(\beta) \cdot \mathbf{w} + QK^\top)/\sqrt{d_k}\right) V, \tag{9}$$

where $Q = f_q(\mathbf{q})$, $K = f_k(\mathbf{q}^{\text{im}})$, and $V = f_v(\mathbf{q}^{\text{im}})$ represent the transformed QKV vectors in the cross attention with projections $f_{(\cdot)}$, and d_k denotes the dimension of K. As a variation of the activation process in Eqs. (7) and (8), $\mathbf{w}$ is a resized version of $\mathbf{m}$, expanding dimension from N to (H, M, N), where H denotes the number of attention heads. We illustrate the variations in attention scores with changing input masks in Fig. 5. Attention score maps show attention concentrations in masked areas. This guides the model's focus towards specific visual elements, leading to various generation outputs.

4. Experiments

4.1. Implementation Details

Pre-trained models. Prompt Highlighter can be applied to a variety of general frameworks. For LLMs, we employ the LLaMA model architecture [7] and utilize Vicuna-13B v1.1 [8] as the test model. For VLMs, experiments are done on one direct token mapping model LLaVA-13B [10, 23] and one query-based mapping model InstructBLIP-Vicuna-13B [15]. We adopt the LLaVA-v1.5 13B model [23] in quantitative evaluations. All experiments are conducted on a single NVIDIA-A800 GPU.

Hyper-parameters. The following parameter configurations are constant throughout almost all examples. In the highlight guidance, we set the guidance strength $\gamma = 1.3$ in Eq. (6), and the scaling parameter α in Eq. (5) is set to 0.01. In attention activation, we accommodate the diverse feature domains across different models by setting $\beta = 2.0$ in Eq. (7) and $\beta = 20.0$ in Eq. (9). The scale factor δ in Eq. (7) is set to satisfy $\delta\mathbf{m} = (\log(\beta) + 2)\mathbf{m}$.

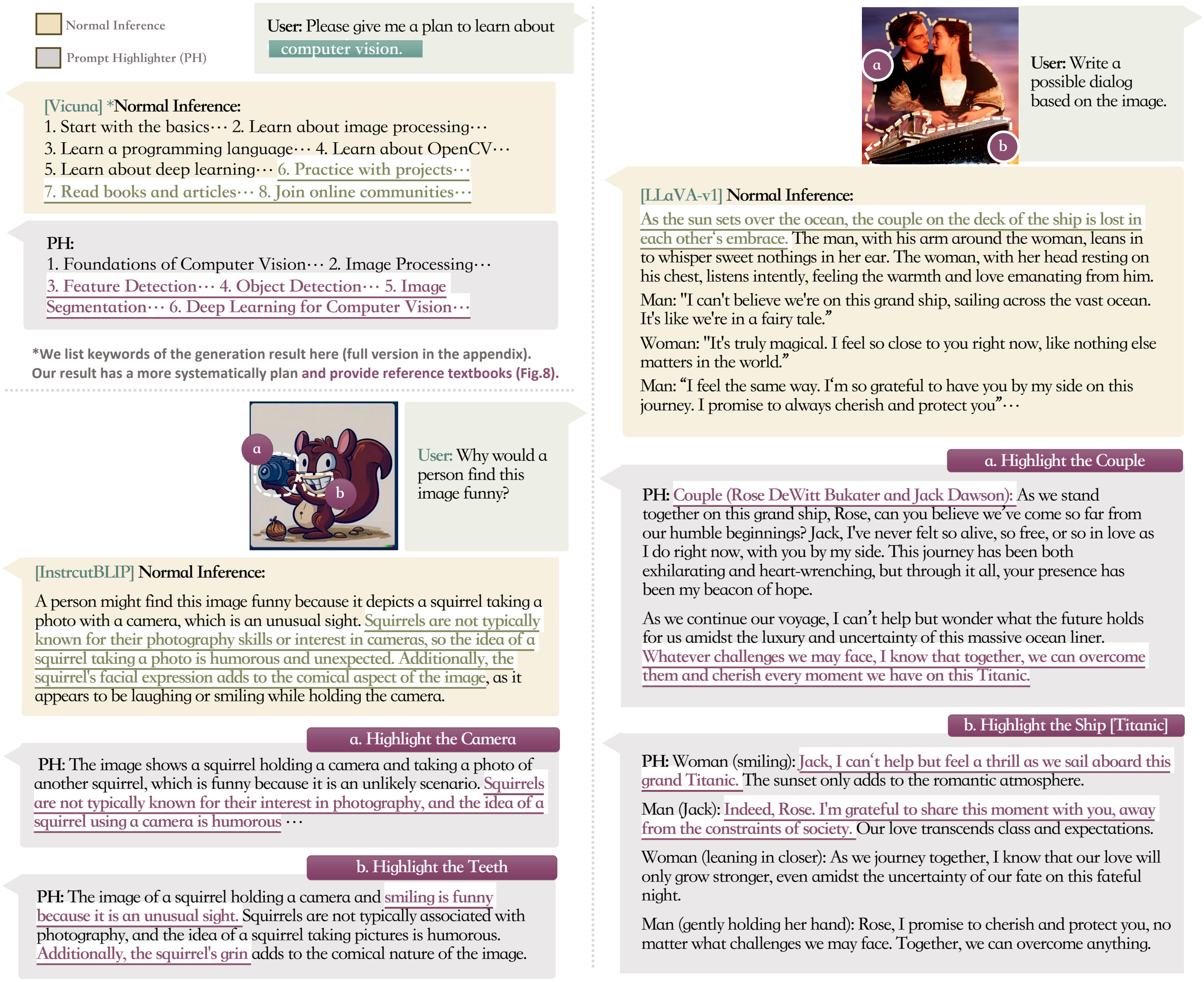
Inference. During user interactions, if the text range selected by the user does not align perfectly with the tokens from the tokenizer, we adjust the start or end selection position to ensure that the selected range is fully encompassed. For images, the input mask is downsampled to a patch-wise binary mask based on the input image. The size depends on the visual encoder's patch size (e.g., CLIP). In the autoregressive generation process, we employ a greedy search and cache prior KV values across layers.

4.2. Applications and Comparisons

Partial context highlighting is a fundamental application of Prompt Highlighter, utilized in scenarios where the emphasis is required on specific spans of the context. As shown in Figs. 1 and 6, our method enables users to guide the model's attention toward relevant parts of the input by highlighting them. It enhances the focus and relevance of the generated output across a range of diverse frameworks [8, 10, 15], making results align more precisely with user-selected tokens. It's worth noting that the examples in Fig. 6 can not be achieved by LLM-CFG [21], as it requires a prompt design that has complete sentences or images. This approach proves particularly useful in tasks such as content summarization and interactive conversations.

Generation control. As demonstrated in Fig. 1, we can further control the degree of correlation to the highlighted part in text generation. Users can manipulate the model's output dynamically by adjusting the highlight guidance strength γ in Eq. (6). This capability can prove advantageous in a variety of tasks, ranging from generating descriptive captions for images to creating customized responses in conversational agents.

Reliable description. When carrying out long text genera-



User: Please give me a plan to learn about computer vision.

[Vicuna] *Normal Inference:
1. Start with the basics··· 2. Learn about image processing···
3. Learn a programming language··· 4. Learn about OpenCV···
5. Learn about deep learning··· 6. Practice with projects···
7. Read books and articles··· 8. Join online communities···

PH:
1. Foundations of Computer Vision··· 2. Image Processing···
3. Feature Detection··· 4. Object Detection··· 5. Image Segmentation··· 6. Deep Learning for Computer Vision···

*We list keywords of the generation result here (full version in the appendix).
Our result has a more systematically plan and provide reference textbooks (Fig.8).

User: Why would a person find this image funny?

[InstrcutBLIP] Normal Inference:

A person might find this image funny because it depicts a squirrel taking a photo with a camera, which is an unusual sight. Squirrels are not typically known for their photography skills or interest in cameras, so the idea of a squirrel taking a photo is humorous and unexpected. Additionally, the squirrel's facial expression adds to the comical aspect of the image, as it appears to be laughing or smiling while holding the camera.

a. Highlight the Camera

PH: The image shows a squirrel holding a camera and taking a photo of another squirrel, which is funny because it is an unlikely scenario. Squirrels are not typically known for their interest in photography, and the idea of a squirrel using a camera is humorous ···

b. Highlight the Teeth

PH: The image of a squirrel holding a camera and smiling is funny because it is an unusual sight. Squirrels are not typically associated with photography, and the idea of a squirrel taking pictures is humorous. Additionally, the squirrel's grin adds to the comical nature of the image.

User: Write a possible dialog based on the image.

[LLaVA-v1] Normal Inference:

As the sun sets over the ocean, the couple on the deck of the ship is lost in each other's embrace. The man, with his arm around the woman, leans in to whisper sweet nothings in her ear. The woman, with her head resting on his chest, listens intently, feeling the warmth and love emanating from him.

Man: "I can't believe we're on this grand ship, sailing across the vast ocean. It's like we're in a fairy tale."

Woman: "It's truly magical. I feel so close to you right now, like nothing else matters in the world."

Man: "I feel the same way. I'm so grateful to have you by my side on this journey. I promise to always cherish and protect you"···

a. Highlight the Couple

PH: Couple (Rose DeWitt Bukater and Jack Dawson): As we stand together on this grand ship, Rose, can you believe we've come so far from our humble beginnings? Jack, I've never felt so alive, so free, or so in love as I do right now, with you by my side. This journey has been both exhilarating and heart-wrenching, but through it all, your presence has been my beacon of hope.

As we continue our voyage, I can't help but wonder what the future holds for us amidst the luxury and uncertainty of this massive ocean liner. Whatever challenges we may face, I know that together, we can overcome them and cherish every moment we have on this Titanic.

b. Highlight the Ship [Titanic]

PH: Woman (smiling): Jack, I can't help but feel a thrill as we sail aboard this grand Titanic. The sunset only adds to the romantic atmosphere.

Man (Jack): Indeed, Rose. I'm grateful to share this moment with you, away from the constraints of society. Our love transcends class and expectations.

Woman (leaning in closer): As we journey together, I know that our love will only grow stronger, even amidst the uncertainty of our fate on this fateful night.

Man (gently holding her hand): Rose, I promise to cherish and protect you, no matter what challenges we may face. Together, we can overcome anything.

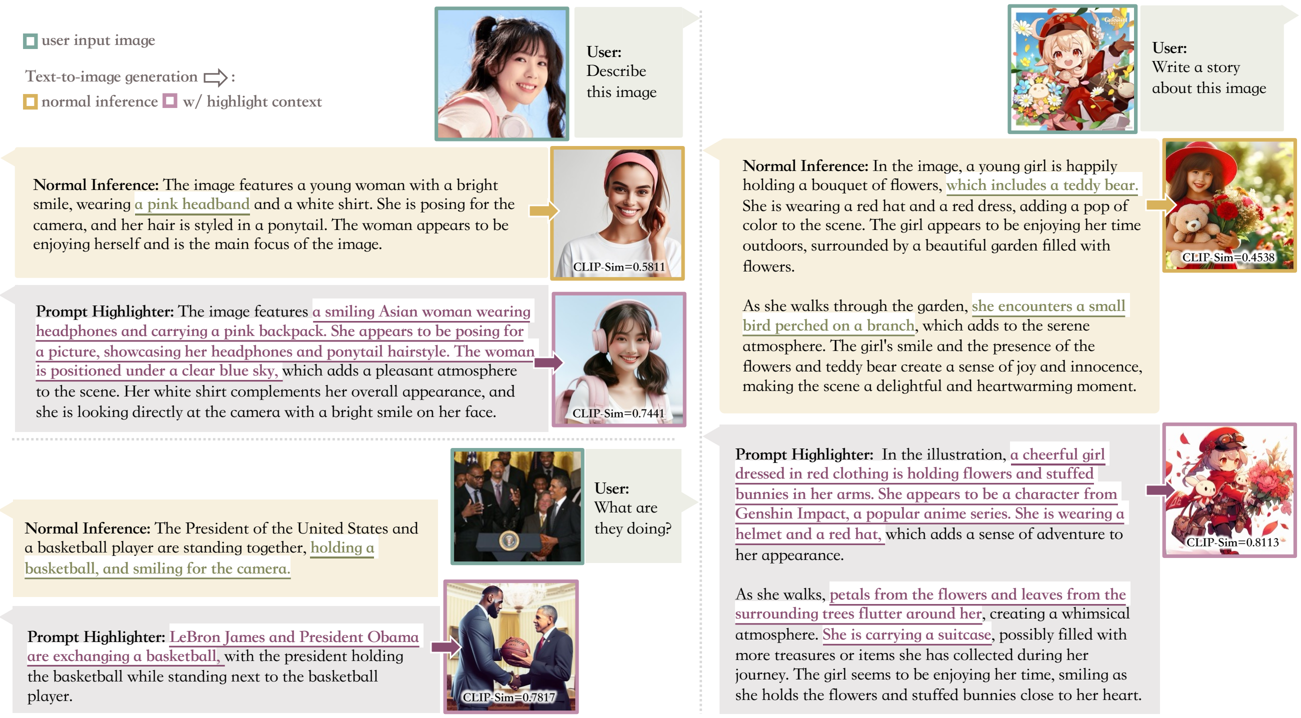
Figure 6. Partial context highlighter. When given user input , vanilla inference might lead to unfocused generations. Our context-highlighted inference can faithfully capture the content of the highlighted part. The highlighted sections in each case are marked with circled indices in the image. Outputs correlated with the highlighted parts are underlined.

tion tasks, every predicted token interacts with all previous content, including the newly generated contexts, to gather information. This may lead to a gradual divergence and loss of attention between the generated tokens and the input context, causing the model to hallucinate. We address this issue by highlighting all user input condition tokens, thereby guiding the generated content to align more closely with the input context. This approach is particularly vital for descriptive tasks, such as generating image captions. We demonstrate this in Fig. 7. Our method can generate more accurate and detailed descriptions. Moreover, when the generated description is fed into an image generation model like DALLE-3 [44], our method demonstrates its advantages in facilitating image-text alignment.

4.3. Quantitative Evaluation

General VLM benchmarks. In Tab. 1, we evaluate our method on common comprehensive Vision-Language benchmarks, MME [27] and MMBench [26]. Prompt Highlighter on LLaVA-v1.5 demonstrates a consistent performance improvement compared to well-trained models by designating the entire image as the highlighted part in the input context. Notably, though these benchmarks primarily assess overall performance with single-token generation and are not designed for user interactions, we still get a competitive place in both MMBench and MME perception. Additionally, further benchmarking with different hyper-parameter selection validates our performance enhancements with LLaVA-v1.5 13B, as evidenced by gains on POPE (85.9 - **87.8**, +1.9), MMB^{CN} (63.4 - **64.0**, +0.6), and SQA^{I} (71.6 - **72.4**, +0.8). With LLaVA-NeXT-34B [45], we achieve a SOTA performance on MMB^{test} with 81.3.

Reliable description. We then evaluate the reliability of image captions generated by our method. In this case, we utilized the reference-free metric CLIP Score [48], ex-



User: Describe this image

Normal Inference: The image features a young woman with a bright smile, wearing a pink headband and a white shirt. She is posing for the camera, and her hair is styled in a ponytail. The woman appears to be enjoying herself and is the main focus of the image.

Prompt Highlighter: The image features a smiling Asian woman wearing headphones and carrying a pink backpack. She appears to be posing for a picture, showcasing her headphones and ponytail hairstyle. The woman is positioned under a clear blue sky, which adds a pleasant atmosphere to the scene. Her white shirt complements her overall appearance, and she is looking directly at the camera with a bright smile on her face.

User: What are they doing?

Normal Inference: The President of the United States and a basketball player are standing together, holding a basketball, and smiling for the camera.

Prompt Highlighter: LeBron James and President Obama are exchanging a basketball, with the president holding the basketball while standing next to the basketball player.

User: Write a story about this image

Normal Inference: In the image, a young girl is happily holding a bouquet of flowers, which includes a teddy bear. She is wearing a red hat and a red dress, adding a pop of color to the scene. The girl appears to be enjoying her time outdoors, surrounded by a beautiful garden filled with flowers.

As she walks through the garden, she encounters a small bird perched on a branch, which adds to the serene atmosphere. The girl's smile and the presence of the flowers and teddy bear create a sense of joy and innocence, making the scene a delightful and heartwarming moment.

Prompt Highlighter: In the illustration, a cheerful girl dressed in red clothing is holding flowers and stuffed bunnies in her arms. She appears to be a character from Genshin Impact, a popular anime series. She is wearing a helmet and a red hat, which adds a sense of adventure to her appearance.

As she walks, petals from the flowers and leaves from the surrounding trees flutter around her, creating a whimsical atmosphere. She is carrying a suitcase, possibly filled with more treasures or items she has collected during her journey. The girl seems to be enjoying her time, smiling as she holds the flowers and stuffed bunnies close to her heart.

Figure 7. Results when highlighting all input contexts. Given user input , vanilla inference might lead to hallucinations. In contrast, context-highlighted inference can accurately capture the content of the image. We further feed descriptions into DALLE-3 [44] (shown on the right) to provide a visually apparent difference. The CLIP-Similarity [30] with the input image is reported for each generated image.

method	MME	MMB$_{\texttt{dev}}$	MMB$_{\texttt{test}}$	PoPE
QWen-VL-Chat [46]	1487.5	60.6	61.8	-
mPLUG-Owl-2 [14]	1540.2	64.5	66.0	-
LLaVA-v1.5 [23]	1531.3	67.7	67.0	85.9
Prompt Highlighter	**1552.5** +21	**69.7** +2.0	**69.5** +2.5	**87.8** +1.9

Table 1. Evaluations on comprehensive VLM benchmarks, including MME Perception [27] and MMBench (MMB) [26].

		text → image		image → text	
method	S-CLIP	R@1	R@5	R@1	R@5
FuseCap [47]	0.785	37.2	62.3	47.7	72.3
CoCa-CFG [22]	0.808	44.6	**71.7**	-	-
LLaVA-v1.5 [23]	0.809	36.6	62.2	50.7	76.4
Prompt Highlighter	**0.829**	**45.2**	**71.7**	**62.2**	**85.0**

Table 2. CLIP-based reference-free image caption evaluation conducted on the MSCOCO [25].

pressed as S-CLIP $= 2.5 \cdot \max(\cos(\boldsymbol{v}, \boldsymbol{c}), 0)$, to evaluate the embedding similarity between the image $\boldsymbol{v}$ and the generated caption $\boldsymbol{c}$. We also report the text $\leftrightarrow$ image retrieval recall R@1, R@5 in the CLIP embedding space. As demonstrated in Tab. 2, on the MSCOCO Karpathy test set [25], our method shows a state-of-the-art CLIP score when compared to recent competitive caption-focused methods [22, 47]. Our results exhibit significant advantages across all metrics compared to the baseline.

User study. We conduct a user study to assess the usability and effectiveness of our method. This study asks participants to rank generation results across five tasks (image captioning, image/text partial highlight generation, text for image generation, and image understanding). Compared with the original inference baselines, the collected 255 valid preference results indicated that **77.3%** of users found Prompt Highlighter to generate more correlated results and be beneficial in accomplishing the task objectives.

More details about the quantitative evaluation can be found in Appendix A.2.

4.4. Ablation Study

Module-wise ablation. In Tab. 3, we systematically conduct an ablation study by removing each module of Prompt Highlighter and noting the performance change in MME and MMBench-`dev`. The results revealed that removing the attention activation module led to the most considerable performance reduction, and combining the highlight guidance and attention activation significantly improves the overall performance.

Hyper-parameters. In Tab. 4, we explore the impact of three scaling parameters (α, β, γ) in Eqs. (5) to (7) for highlight control on the performance of our method. We observed a trade-off between higher concentration (higher β and γ) and general vision-language understanding ability.

settings	Guidance	Attention	MME	MMB$_{\text{dev}}$
baseline (our impl.)			1528.7	67.7
w/ Guidance	✓		1531.1	68.5
w/ Attention		✓	1537.2	69.5
Full pipeline	✓	✓	**1552.5**	**69.7**

Table 3. A module-wise ablation study.

α (in Eq. (5))	0.0	0.01	0.1	1.0
(α, 2.0, 1.3)	1517.6	**1552.5**	1522.3	1524.2
β (in Eq. (7))	1.0	2.0	3.0	4.0
(0.01, β, 1.3)	1527.2	**1552.5**	1537.3	1535.7
γ (in Eq. (6))	1.0	1.3	1.5	2.0
(0.01, 2.0, γ)	1537.2	**1552.5**	1532.6	1524.0

Table 4. Hyper-parameter ablation on MME. We probe for the most suitable value for one of the combinations in (α, β, γ).

method	token/s ↑	memory (MB) ↓	S-CLIP ↑
Baseline	6.67	16231	0.809
Baseline beam=2	5.97	18537	0.807
Prompt Highlighter	5.95	17373	0.829

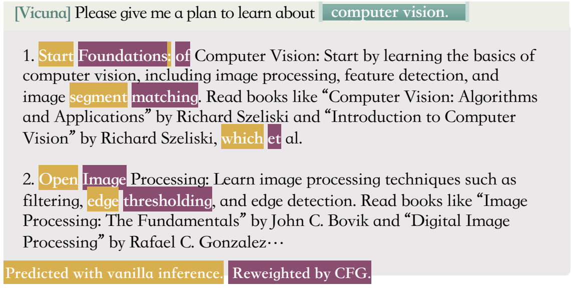
Table 5. Evaluation on inference speed and GPU memory. The memory is dominated by the model weight, consuming 13971 MB.

4.5. Discussions

Prediction control by CFG. Given that our highlight guidance operates on the predicted token probability, similar to LLM-CFG [21], we further investigate the semantic-level distinctions in normal and unconditional branches. This is visualized through an example of token prediction in Fig. 8. When the embeddings associated with the highlighted tokens are perturbed, the unconditional-context branch tends to predict unrelated to the highlighted parts. This, in turn, enables the rectification of responses in text-generation tasks.

Attention activation. To confirm that the attention activation operates as anticipated, we not only visualize the activated attention scores corresponding to different regions in Fig. 5, but we also demonstrate its effectiveness using 500 attention maps from the caption experiment in Fig. 9. Firstly, we validate the band-like pattern property discussed in Sec. 3.2, by comparing the vertical and horizontal gradients. We observed a significant gradient gap with $G_x > G_y$ in all 500 cases. Given this property, the attention activation can capture a higher attention probability p, as defined in Eq. (8), from the given contexts. This attention contribution plotted in Fig. 9 is denoted as $\sum_{m_i=1}(p_i)/\sum_{j<i}(p_j)$. Consequently, attention activation leads to results that adhere more closely to the input context.

Limitations and future work. While our approach introduces a novel method for controlling generation in multi-modal LLMs, it has certain limitations: *(a). Additional computations:* Our method requires an extra decoding branch, which brings additional computational overhead and GPU memory requirements. However, these additional loads are marginal and acceptable with the batched inference. We validate this by the caption experiment in Tab. 5. *(b). Dependence on base model:* Content generation quality is tied to the base model's capabilities, which may result in over-emphasis or miss-emphasis on highlighted parts when using poorly-trained base models.


[Vicuna] Please give me a plan to learn about computer vision.

1. Start Foundations: of Computer Vision: Start by learning the basics of computer vision, including image processing, feature detection, and image segment matching. Read books like "Computer Vision: Algorithms and Applications" by Richard Szeliski and "Introduction to Computer Vision" by Richard Szeliski, which et al.

2. Open Image Processing: Learn image processing techniques such as filtering, edge thresholding, and edge detection. Read books like "Image Processing: The Fundamentals" by John C. Bovik and "Digital Image Processing" by Rafael C. Gonzalez···

Predicted with vanilla inference. Reweighted by CFG.

Figure 8. An example of token changed with CFG.

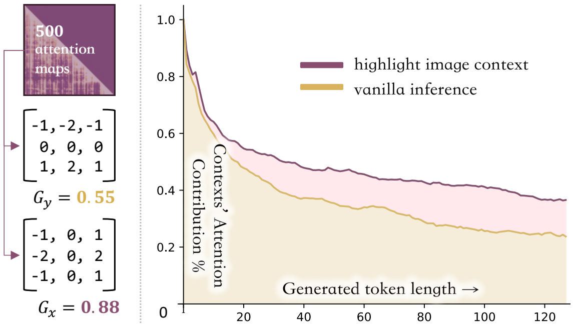


Figure 9. ***Left***: A simple verification of the vertical band-like pattern in the attention map, with a report on the gradient summation. ***Right***: Following [49], we present a visualization displaying the average contribution of context's attention during generation.

One direction for future work will be to create a more intuitive highlighting scheme. We also aim to extend our method to support a greater variety of interactions.

5. Conclusion

We introduce Prompt Highlighter, a novel paradigm for user-model interactions in multi-modal LLMs, offering output control through a token-level highlighting mechanism. This approach, requiring no extra training, competes well on standard benchmarks and provides reliable generation outputs by merely highlighting input context. Further, diverse applications demonstrate its intuitive usability and effectiveness in enhancing control over the generation process. This work represents a promising direction for enhancing user control in multi-modal LLMs, and we anticipate it will inspire further research.

Acknowledgement This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R and the Shenzhen Science and Technology Program under No. KQTD20210811090149095.

References

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1, 2

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 4, 5

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 4, 5, 10, 14

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 1, 2

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 4, 5

[11] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 4

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 4

[14] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 4, 7, 10

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2, 5, 12, 15

[16] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 1, 2

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 2

[18] Significant-Gravitas. Significant-gravitas/autogpt: An experimental open-source attempt to make gpt-4 fully autonomous. 2

[19] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 1, 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[21] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023. 3, 5, 8, 12

[22] Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. Guiding image captioning models toward more specific captions. In *ICCV*, pages 15259–15269, 2023. 2, 7, 10

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 5, 7, 9, 10, 15

[24] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 7

[26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2, 6, 7, 10

[27] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 6, 7, 10

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov,

Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[29] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 7, 10

[31] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[32] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023. 2, 9

[33] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023. 2

[34] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Decomposition enhances reasoning via self-evaluation guided decoding. *arXiv preprint arXiv:2305.00633*, 2023. 2

[35] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing, 2023. 2

[36] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2

[37] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want, 2023. 2

[38] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *ICCV*, 2023. 3

[39] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023. 3

[40] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. 2024. 3

[41] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance, 2024. 3

[42] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 3

[43] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv*, 2023. 4

[44] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. 6, 7, 10, 16

[45] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 6

[46] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7

[47] Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models to fuse visual data into enriched image captions. *arXiv preprint arXiv:2305.17718*, 2023. 7

[48] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6, 10

[49] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. In *NeurIPS*, 2023. 8

[50] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning, 2023. 10

[51] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. `https://github.com/Alpha-VLLM/LLaMA2-Accessory/blob/main/SPHINX/SPHINX_paper.pdf`, 2023. 10

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 10