Unleashing Region Understanding in Intermediate Layers for MLLM-based Referring Expression Generation

Yaoyuan Liang[1]*, **Zhuojun Cai**[1]*, **Jian Xu**[1], **Guanbo Huang**[1], **Yiran Wang**[1], **Xiao Liang**[1],
Jiahao Liu[2], **Ziran Li**[2], **Jingang Wang**[2], **Shao-Lun Huang**[1]†
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Meituan Inc.

Abstract

The Multi-modal Large Language Model (MLLM) based Referring Expression Generation (REG) task has gained increasing popularity, which aims to generate an unambiguous text description that applies to exactly one object or region in the image by leveraging foundation models. We empirically found that there exists a potential trade-off between the detailedness and the correctness of the descriptions for the referring objects. On the one hand, generating sentences with more details is usually required in order to provide more precise object descriptions. On the other hand, complicated sentences could easily increase the probability of hallucinations. To address this issue, we propose a training-free framework, named as "unleash-then-eliminate", which first elicits the latent information in the intermediate layers, and then adopts a cycle-consistency-based decoding method to alleviate the production of hallucinations. Furthermore, to reduce the computational load of cycle-consistency-based decoding, we devise a Probing-based Importance Estimation method to statistically estimate the importance weights of intermediate layers within a subset. These importance weights are then incorporated into the decoding process over the entire dataset, intervening in the next token prediction from intermediate layers. Extensive experiments conducted on the RefCOCOg and PHD benchmarks show that our proposed framework could outperform existing methods on both semantic and hallucination-related metrics. Code will be made available in `https://github.com/Glupayy/unleash-eliminate`.

1 Introduction

Referring expression generation (REG) [11, 29, 31, 53, 54] is a task to generate an unambiguous text description that applies to exactly one appointed object or region in the image. A good expression should be distinguishable enough to ensure that the listener can identify the unique target among various objects within the same image. With the great success achieved by large language models (LLMs), multi-modal large language models (MLLMs) [1, 9, 12, 27, 56, 61] have been introduced to perform this task and become increasingly popular in the research community. Some representative works [6, 37, 52, 55, 57] conduct visual instruction tuning on specialized region-involved multi-modal corpus and successfully empower MLLMs with the ability of region-level understanding.

Though certain progress has been made, MLLMs themselves suffer from object hallucination [14, 23, 46, 50, 60]. Research suggests that the mechanism behind MLLMs' hallucinations could be related to the over-reliance on prior knowledge of LLMs rather than the multi-modal context provided by

[*] Equal contribution.
[†] Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3833

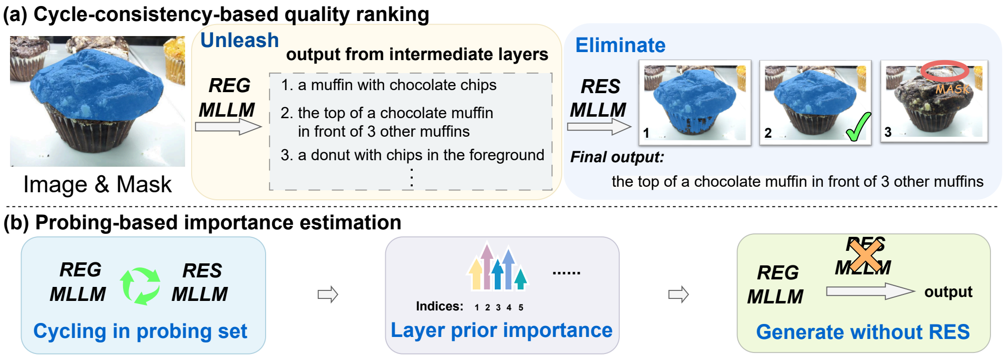

(a) Cycle-consistency-based quality ranking
Image & Mask
Unleash
output from intermediate layers
REG
MLLM
1. a muffin with chocolate chips
2. the top of a chocolate muffin
in front of 3 other muffins
3. a donut with chips in the foreground
Eliminate
RES
MLLM
1
2
3
MASK
Final output:
the top of a chocolate muffin in front of 3 other muffins
(b) Probing-based importance estimation
REG
MLLM
RES
MLLM
Cycling in probing set
Indices: 1 2 3 4 5
......
Layer prior importance
REG
MLLM
RES
MLLM
output
Generate without RES

Figure 1: Illustration of our method. (a) Given an image and a mask appointed to the target object, we first unleash the descriptive regional information of the middle layers and gain various candidate captions. These outputs are sent to a RES model that serves as a "listener" and the "listener" eliminates the inaccurate candidates. (b) Our proposed strategy to diminish the computational load of RES. First estimate the layer prior importance on a probe set with RES, then leverage it for the RES-free next token prediction.

the input [14, 19, 50]. Therefore, MLLM-based REG models naturally inherit the aforementioned hallucination issues. Additionally, a further complication arises in the context of region-level understanding, where models are required to generate precise and identifiable descriptions for specific regions. To achieve that, the MLLMs sometimes have to make use of the surroundings for reference while effectively avoiding out-of-region information distorting the description of the target object, e.g. incorrectly attributing characteristics of other objects to the targeted region (Figure 7 in appendix). This requirement exacerbates the issue of attribute-level hallucinations [3, 50]. In this paper, we aim to explore this potential trade-off between the detailed description and accurate targeting of referring objects in MLLM-based REG task. To be specific, providing precise object descriptions necessitates generating sentences with more details, which results in longer sentences. Meanwhile, the text with increased granularity (length) is more likely to contain inaccurate or spurious information, commonly defined as "hallucinations." Table 1 shows an example of quantitative analysis.

To address the trade-off between information richness and reliability, we propose a novel approach called "unleash-then-eliminate," as depicted in Figure 1. We observed that the alignment of region-level multi-modal information does not maintain a monotonic progression during inter-layer transitions. Moreover, when multi-modal hidden states are projected into the language space using the language head, the intermediate layers sometimes hold more descriptive region information than the final layer. These observations (Section 3.2) imply that the most suitable layer for each referred region should not be solely confined to the final layer. Accordingly, we adopt contrastive decoding [8, 22] to unleash the object information contained in the intermediate layers. To eliminate unsuitable candidate outputs, we leverage the dual task of REG: Referring expression segmentation (RES) [37, 48], which aims to segment a target object mask from the entire image given a sentence describing the object. Ideally, for the same object, the region input (formed as a mask) for the REG task and the mask output from the RES task should exhibit cycle consistency. Based on this insight, we propose a cycle-consistency-based decoding method, which enables us to choose among multiple outputs based on their descriptive quality thus reducing hallucinations while maintaining the richness of the output sentences. Furthermore, considering the need to diminish the computational load of RES, we develop a hybrid layer importance measurement strategy to select the best layer for each token during the next word prediction. This strategy leverages both the layer-wise prior importance estimated over probing subset, and the Jensen-Shannon divergence [8] between the logits of each candidate layer and the last layer. With the layer-wise prior importance, the MLLM-based REG model maintains promising performance in reducing hallucinations and enhancing the granularity of the generated sentences, even without the assistance of RES model. Extensive experiments conducted on the RefCOCOg [33] and PHD [28] benchmark demonstrate that our proposed framework surpasses existing methods on both semantic and hallucination-related metrics.

2 Related works

Region-level understanding in Multimodal Large Language Models. Significant progress has been made in unleashing the region-level understanding ability in MLLMs [6, 18, 37, 52, 55, 57]. To incorporate region-level information into sequence generation of MLLMs, some approaches [6, 52] integrate bounding box coordinates into the language input in the form of natural language prompts. Ferret [52] proposes a spatial-aware visual sampler that enables the arbitrary shapes of visual prompts. The current proposed Osprey [55] unlocked the capability of pixel understanding, alleviating the influence of irrelevant information in the visual prompt inputs. These methods have propelled the REG task into the era of MLLMs. Considering that MLLMs incorporate extensive knowledge from unimodal and multimodal pre-trained corpora, there is a potential for the REG task to leverage this inherent knowledge within the models to generate more specific and detailed expressions without additional training, thus more effectively addressing real-world applications [21, 39, 41].

Decoding strategies to mitigate hallucination in Large Language Models. Large language models are pre-trained on unlabeled corpora to acquire extensive world knowledge and subsequently undergo post-training to learn to follow instructions [34] and align with human preferences [2]. This systematic pre- and post-training pipeline makes them powerful at solving a wide range of NLP tasks [16, 36, 40, 43]. However, some studies indicate that they may fail to accurately assess their own knowledge [51] and often exhibit overconfidence in their responses [47], which results in hallucinations [59]. To mitigate these issues, recent research [5, 8, 20] proposes inference-time decoding strategies for trained LLMs to find latent knowledge inside the internal activations without additional training. C. Burns *et al.* [5] introduce a Contrast-Consistent Search (CCS) algorithm to identify a direction of truth in the activation space of LLMs that remains consistent across negations, thereby reducing generated errors. Based on the discovery of CCS, ITI [20] dives deep into attention heads and suggests shifting model activations alongside factuality-related heads during inference to help reduce hallucinations. Besides "finding the direction of truth," DoLa [8] proposes contrastive decoding by comparing the differences in logits between the projections of later and earlier layers to better surface factual knowledge and reduce the generation of incorrect facts. In line with the motivations of [5, 20], our investigation uncovers that well-trained MLLMs' intermediate layers differ in multi-modal alignment and region-level understanding capabilities. These observations inspire us to devise an inference-time decoding strategy that combines latent knowledge from multiple layers (with prior importance) to alleviate hallucinations.

Hallucination in Multi-modal Large Language Models. In the realm of MLLMs, "hallucination" typically refers to "object hallucination," where the models generate plausible outputs containing objects that are either absent from or mismatched with the images [15, 19, 24, 38, 50], and is commonly categorized into three types: category, attribute and relation hallucinations [3, 49]. Some efforts based on instruction tuning have been made to mitigate this issue in MLLMs. LRV-Instruction [26] introduces a dataset with positive instructions and unique negative prompts with different semantic levels to better align responses with image content. HACL [15] explores the vision-language embedding space, using contrastive learning to separate non-hallucinated from hallucinated texts. Without instruction tuning, Woodpecker [50] offers a training-free pipeline for hallucination correction, using expert models to enrich image context and ensuring each phase is interpretable by a step-by-step correction process. In another line of work, some efforts have sought new decoding strategies to avoid relying on extensive additional data and training: Opera [14] tackles the partial over-trust issue in decoding by applying a penalty to the model logits during beam-search decoding. VCD [19] links object hallucinations to biases and language priors, contrasting outputs from distorted and original visuals to ensure consistent generation. Compared to existing studies, our method initially enriches region-level context from intermediate layers rather than external expert models or knowledge base and provides a multi-layer ensembling solution to mitigate hallucinations.

3 Method

3.1 Preliminary

Leveraging the nuanced language representation capabilities inherent in multi-modal large language models, recent investigations have unlocked the REG [6, 37, 52, 55] and RES [18, 37] capabilities of MLLMs. To maintain clarity and conciseness, we have omitted the discussion on instruction tuning and assume that the described tasks are well-instructed by specific prompts. In this work, we utilize a "mask" as the region prompt.

Referring expression generation. The MLLM-based REG model is typically structured around three principal components: a visual encoder (e.g., CLIP [35] with a linear adapter), a region-encoder,

https://doi.org/10.52202/079017-3833

and an LLM base (e.g., LLaMA [42] or Vicuna [7]). Given an image I, the visual encoder first extracts the visual information, then the linear adapter projects it into $I_f \in \mathbb{R}^{HW \times C}$, where HW denotes the flattened visual token length and C denotes the hidden state dimension of LLM base. A referred region M (presented as a mask) is encoded as $M_f \in \mathbb{R}^{Q \times C}$ by region encoder (e.g., a CNN-based extractor [55] or an ROI pooling layer [37, 57]), where Q denotes the length of encoded region prompt tokens. Taking the visual feature I_f, regional feature M_f, and an embedded instruction text $X_f \in \mathbb{R}^{L \times C}$ as input, where L represents the length of the instruction in tokens, the model concatenates them and forms a multi-modal input sequence $S \in \mathbb{R}^{(HW+Q+L) \times C}$.

Regarding the core architecture of the LLM base, the multi-modal sequence S is successively processed by the N stacked transformer layers. Eventually, an affine layer $\phi(\cdot)$ serves as a language model head to predict the probability of the next token y_t over the vocabulary set $\mathcal{V}$. The logits for the token y_t, given the sequence S and all preceding tokens $y_{<t}$, are computed as follows:

$$\text{logit}_N(y_t|S, y_{<t}) = \phi(h_t^{(N)}), \quad y_t \in \mathcal{V}, \tag{1}$$

where $h_t^{(N)}$ denotes the hidden state of the last transformer layer. The probability of the next token y_t is then given by:

$$p(y_t|S, y_{<t}) = \text{softmax}\big(\text{logit}_N(y_t|S, y_{<t})\big), \quad y_t \in \mathcal{V}. \tag{2}$$

Through this process, the model autoregressively generates the output text Y as the region description. For simplicity, we use $p(y_t|y_{<t})$ to represent $p(y_t|S, y_{<t})$, and $\text{logit}_N(y_t|y_{<t})$ to represent $\text{logit}_N(y_t|S, y_{<t})$ in the following.

Referring expression segmentation. Recent advances in MLLM-based RES models [18, 37] largely employ a similar MLLM architecture (e.g., LLaVA [27]) as used in the REG models, with the distinction that its input contains a description targeted at a specific region, while its output is a mask that can cover the described object.

A widely adopted strategy [17, 18, 37] incorporates a `[SEG]` token, enabling the model to identify the `[SEG]` token in the output sequence as a cue for the presence of a segmentation target. A specialized MLP head ψ processes the output embedding of the `[SEG]` token $h_{seg} \in \mathbb{R}^{1 \times C}$, mapping it into the prompt space of the segmentation fundamental model (e.g., SAM [17]), represented as $\tilde{h}_{seg} = \psi(h_{seg})$. The segmentation model then decodes the target mask M_s from the query token $\tilde{h}_{seg}$ and provides its confidence score CF.

3.2 The intermediate layer contains descriptive information

This section focuses on our observations of intermediate layers, attempting to uncover the latent descriptive information. We adopted a region understanding model Osprey-7b [55] with $N = 32$ layers as REG model, and GLaMM [37] as RES model by default.

Each layer has different generation tendency. To reveal the latent information, we adopt the early-exit strategy on the REG model to generate a series of output sequences by applying the language model head (an affine layer) to the hidden states of each layer. As illustrated in Figure 2, the sequences from the early layers (layers 1 to 10) manifest as nonsensical strings of characters, indicating the suboptimal alignment between the hidden representation of shallow layers and the ultimate vocabulary space of MLLM. Interestingly, among the middle layers (about layers 20 to 25), the model begins to output sentences with semantic meaning and gradually some region-specific expressions appear. Compared with the output of the last layer, these sentences with richer semantic information can sometimes provide more discriminative descriptions for the referred object, e.g. indicating the related position between the target object and others, and/or containing high granularity of attributes. The uniqueness and descriptiveness of expressions given by the middle layers are strongly related to the objective of REG, thus the latent features of intermediate layers have notable potential to enhance the unambiguity in MLLM-based REG.

Potential of intermediate layers. The above observations demonstrate that different layers' understanding of the region context varies. We attempted to visualize the region-aware comprehension capabilities among the layers through two approaches. (a) With the assistance of the RES model: we randomly extracted $K = 2000$ samples from the RefCOCOg training set to form the triplets $(\mathbf{I}, \mathbf{M}, \mathbf{Y})$, which corresponds to the entire images $\mathbf{I}$, the target regions represented by masks $\mathbf{M}$, and the descriptions of regions $\mathbf{Y}$. $(\mathbf{I}, \mathbf{Y})$ are sent into an MLLM-based RES model, to harvest the projected features of `[SEG]` token $\tilde{h}_{seg}$. Subsequently, $(\mathbf{I}, \mathbf{M})$ are input into the REG model, and the

https://doi.org/10.52202/079017-3833

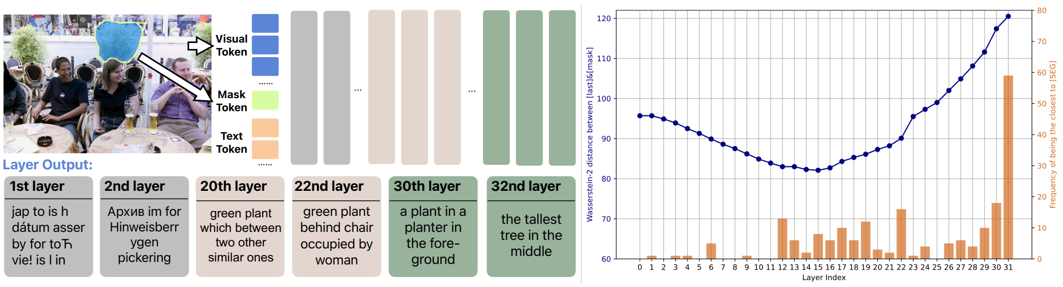


Figure 2: Different layers' understanding of the region context varies, where early layers generate rubbish, middle layers tend to generate descriptive text with higher granularity, and the final layers tend to predict shorter and more precise text. The right part shows the frequencies with which the hidden state of each layer had the smallest Wasserstein-2 distance to the `[SEG]` token (in orange), and the inter-layer transitions of region-level multi-modal alignment (in blue).

hidden states of the last token of each layer $\{h_i\}_{i=1}^{N}$ are obtained. Inspired by FID metric [13] and recent investigations on the latent space communication [30, 32], we first performed PCA dimensionality reduction to project both features into the same dimension, then treat the dimension-reduced `[SEG]` features as the anchor to calculate the Wasserstein-2 distances [13, 44, 45] between `[SEG]` and each intermediate hidden states. This allowed us to estimate the relative region understanding ability of each layer for the given multi-modal context of a certain object. We then calculate the frequency with which each layer had the smallest Wasserstein-2 distance to the `[SEG]` token. As depicted in the right of Figure 2 (orange), except for the final layer, the intermediate layers also have the potential to contain better latent information for a more precise region-related description. (b) Solely within the REG model: we investigated the multi-modal alignment process in region-level context across intermediate layers by calculating the Wasserstein-2 distance between each layer's region-encoded token `[mask]` and the last language token. Results are reported in Figure 2 (blue). These analyses within a well-trained MLLM show that the tokens of different modalities do not approach monotonically during inter-layer transitions. Hence we should give more chances to the intermediate layers for better region-level understanding. Detailed information can be found in the Appendix C.

3.3 Unleash then eliminate decoding strategy

To enhance the granularity of region-level descriptions without introducing excessive hallucinations, we propose a method that integrates contrastive decoding with cycle consistency-based ranking to screen out appropriate descriptions for the interested regions (Section 3.3.1 and 3.3.2). This approach enables us to specifically leverage the commonly overlooked latent information contained in intermediate layers and ensures the identifiable description through caption quality estimation. Under the concerns of computation efficiency, we further develop a decoding strategy to reduce the operations of cycle ranking through hybrid layer importance measurement (Section 3.3.3). It involves two kinds of layer importance calculation to influence the selection probability of candidate layers during each word prediction step. The harvest prior importance weights can be directly applied to the decoding process of the original MLLM, mitigating the hallucinations. The following subsections provide further details.

3.3.1 Unleash intermediate information by contrastive decoding

As depicted in Figure 2, the manifestation of region-aware information differs across layers. To underscore descriptions related to specific regions—evident in the intermediate layers but faded in the final layer, we adopt a contrastive decoding approach [8, 22] by subtracting the log probabilities of the next token produced by the intermediate layer from those of the final layer. The resulting distribution is defined as the contrastive decoding logits of specific subtractor layers. These logits instead of originally the logits from the final layer are used for generating the subsequent token. Concretely, given a set of candidate layer indices $J = \{1, \ldots, n\}$, the probability of token y_t for layer $j \in J$ is:

$$p_{con_j}(y_t|y_{<t}) = \begin{cases} \text{softmax}\big(\text{logit}_N(y_t|y_{<t}) - \text{logit}_j(y_t|y_{<t})\big) & \text{if } y_t \in \mathcal{V}_{head}, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

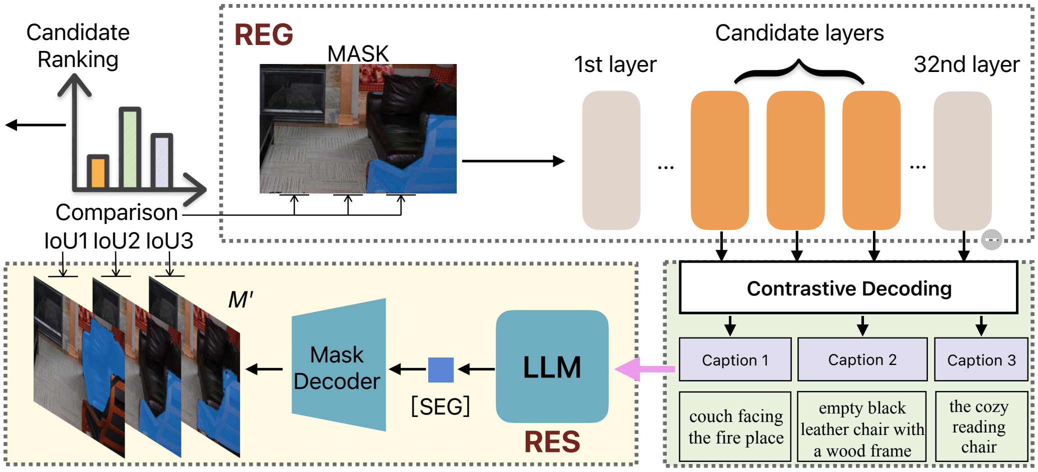


Figure 3: First, we conduct contrastive decoding by subtracting the log probabilities produced by the intermediate layer from those of the final layer. Then, each of the intermediate layers in the candidate subset suggests a probability for predicting the next token. Finally, RES model is used to estimate the performance of the captions by calculating the Intersection over Union (IoU) between MASK and M'_j, ranking these candidates effectively.

where $\text{logit}_j(y_t|y_{<t}) = \phi(h_t^{(j)})$, $\mathcal{V}_{head}(y_t) = \{y_i \in \mathcal{V} : p(y_i|y_{<t}) \geq \alpha \max_{w \in \mathcal{V}} p(w|y_{<t})\}$, $\alpha \in [0, 1]$ is a cutoff hyperparameter that truncates the next token distribution $p(y_t|y_{<t})$ of the final layer. Following the previous works [8, 22], we set $\alpha = 0.1$ in the implementation.
After contrastive decoding, each of the intermediate layers in the candidate set suggests a probability for predicting the next token y_t. Finally, we harvest a sentence set $A = \{a_1, a_2, ..., a_n\}$ whose size equals to the size of candidate layers.

3.3.2 Cycle-consistency-based intermediate sentence quality ranking

Consider the formed triplet ($\mathbf{I}$, $\mathbf{M}$, $\mathbf{Y}$) in Section 3.2, where $\mathbf{I}$ serves as the global contextual background of a region, $\mathbf{M}$ and $\mathbf{Y}$ respectively represent two modalities of the same object. For a pair of ideal REG and RES models, it is anticipated that $\mathbf{M}$ and $\mathbf{Y}$ can be interconverted losslessly during a cyclic operation of these two models. This implies that, due to the cycle consistency between the two modalities of the same object, feeding the output of the REG model into the RES model should yield a mask consistent with the input to the REG. If the output generated from the candidate layer is overly ambiguous or polluted by hallucinations, the RES model may struggle with accurately locating the target object against the background. Based on these assumptions and observations, we utilize RES model [37] to estimate the region understanding performance of the captions generated by the candidate layers, allowing us to rank these candidates effectively.

Figure 3 illustrates the pipeline of cycle-consistency-based quality ranking. The input image I and the referred region (represented as a mask) M are processed by the REG model, which continuously extracts and aligns features across successive layers for the multi-modal context. After the information elicitation of intermediate layers, we harvest a set of sequence $A = \{a_1, a_2, ..., a_n\}$. Each sequence in A, paired with the image I, forms an input pair (I, a_j) to feed into the RES model. The RES model then segments out the corresponding mask M'_j for each input pair.

We evaluate the quality of each layer's sentences by calculating the Intersection over Union (IoU) between M and M'_j:

$$\text{Score}_j = CF_j \cdot \frac{|M \cap M'_j|}{|M \cup M'_j|}, \tag{4}$$

where CF_j refers to the output of IoU score head within the segmentation foundation model (e.g. SAM [17]), which can be interpreted as a confidence score of the generated mask. Subsequently, the candidate sentences from each layer are ranked by the Score_j. For each sample, the a_j with the highest Score_j is selected as the final sentence, and its layer is deemed the best candidate layer.

https://doi.org/10.52202/079017-3833

Algorithm 1 Layer Prior Importance Calculation

```
Input: $\mathbf{Score} \in \mathbb{R}^{m \times n}$    ▷ $m$: Number of subset samples; $n$: Number of candidate layers
Output: $\boldsymbol{q} \in \mathbb{R}^n$, $\sum_{j=1}^{n} q_j = 1$    ▷ Importance weight of each candidate layer
Initialize count $\leftarrow \mathbf{0} \in \mathbb{R}^n$    ▷ counts of times each layer has the maximum score
for $j \leftarrow 1$ to $n$ do
    $count_j = \sum_{i=1}^{m} \mathbb{1}\{j = \arg\max_{j \in J} \text{Score}_{i,j}\}$
    $q_j \leftarrow \frac{count_j}{m}$
return $\boldsymbol{q}$
```

3.3.3 Hybrid layer importance measurement

Although our cycle-consistency-based candidate ranking process improves the generation quality, it introduces additional computational load from the RES model compared to the original decoding method, significantly affecting the per-sample decoding speed. To alleviate this issue, we propose a simple yet effective strategy called *Probing-based Importance Estimation* to speed up the decoding process. This strategy involves frequency counting of each candidate layer being the optimal layer based on their performance within a probing subset. With a subset containing m samples, the specifics of this calculation are outlined at Algorithm 1. The estimated weight $\boldsymbol{q}$ is then served as prior knowledge that reflects the candidate layers' importance and is then utilized over the entire dataset in the decoding process, intervening in the next-word prediction.

Furthermore, inspired by the success of distance-guided layer selection in LLMs [8], we also apply this metric as a second guidance. Concretely, we first calculate the distance between the next-token probability of the final layer and each candidate layer at the current decoding step, then the calculated values are normalized across candidate layers as follows,

$$d_j = \frac{D_j}{\sum_{i=1}^{n} D_i}, \tag{5}$$

where $D_j = \text{JSD}(p||p_{con_j})$ denotes the J-S divergence between the next-token probability of the final layer p and the layer p_{con_j}. Finally, the hybrid layer importance is obtained by adding up the probing-based prior and sample-wise distance followed by a softmax normalization:

$$\tilde{\boldsymbol{q}} = \text{softmax}(\boldsymbol{d} + \boldsymbol{q}), \tag{6}$$

where $\boldsymbol{d}$ is a vector of normalized divergence values across the candidate layers. We then sample from the probability distribution $\tilde{\boldsymbol{q}}$ to select one layer among candidates to predict the next token at each decoding step, till the generation finishes.

Overall, this decoding approach first estimates a prior layer-wise importance weight from a small subset and then applies this distribution to contrastive decoding, effectively improving decoding efficiency while preserving the ability to mitigate hallucinations. In addition, our experiments (Table 3) also show that the prior importance $\boldsymbol{q}$ calculated from one (probing) dataset could be directly transferred to another dataset with similar image domain for reducing the hallucinations. This strong transferability further illuminates a new promising application scenario where the historic estimate could serve as a cold-start for inference in new environments when the probing dataset is not available.

4 Experiments

4.1 Datasets and metrics

RefCOCOg. The RefCOCOg [33] dataset is a classical benchmark for referring expression generation, which contains 85,474 expressions for 54,822 objects in 26,711 images. Expressions in RefCOCOg are annotated on Amazon Mechanical Turk in a non-competitive way and tend to be longer (8.43 words per sentence on average) and more expressive than RefCOCO [53] and RefCOCO+ [53]. In the MLLM-empowered REG task, the relatively short ground truth cannot fully cover the expression space of the sentences generated by the MLLM at evaluation. Therefore, our experiments focus on the METEOR metric as it is more comprehensive and flexible, allowing for a more nuanced recognition of linguistic variations.

Table 1: Comparison of generation and hallucination performance on RefCOCOg. t denotes the temperature parameter. "1/8" denotes using 1/8 samples of the total dataset to estimate the layer prior importance. "full-R" denotes quality ranking on the whole dataset.

Model	METEOR↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall↑	Len	nCHAIR$_S$ ↓	nCHAIR$_I$ ↓
Osprey-7b (t=0.2)	162.0	**23.41**	**20.81**	0.7631	7.15	3.2741	2.9105
Osprey-7b (t=0.9)	140.0	27.90	24.12	0.7514	8.11	3.4402	2.9741
DoLa	168.0	43.44	31.78	0.8196	23.07	1.8830	1.3775
Ours (1/8)	172.0	42.25	30.95	0.8211	22.96	1.8406	1.3484
Ours (full-R)	**173.0**	42.40	31.20	**0.8237**	23.16	**1.8307**	**1.3472**

CHAIR Evaluation on Hallucinations. The Caption Hallucination Assessment with Image Relevance (CHAIR) metric [38] is commonly used to evaluate object hallucinations that occur in image description tasks. It comprises two distinct assessment dimensions, including CHAIR$_S$ that calculates on sentence-level and CHAIR$_I$ that calculates on a more granular object-level. We observe that CHAIR only counts hallucinated objects for each "central object", which means that if the model wants to enrich the semantic information and generate new tokens, it will face the risk of increasing hallucinations compared to the shorter-sentence models. This undermines the comparability of this metric across sentences of varying lengths. Hence we propose normalized CHAIR (nCHAIR) based on CHAIR to conduct a fair comparison between sentences of different lengths. The calculation of nCHAIR$_S$ is specified as the following formula, with nCHAIR$_I$ calculated similarly:

$$\text{nCHAIR}_S = \frac{|\{\text{hallucinated sentences}\}|}{|\{\text{average number of tokens per sentence}\}| \times |\{\text{all sentences}\}|}. \tag{7}$$

Prompted Visual Hallucination Evaluation Benchmark (PHD). This benchmark [28] focuses on the four major types of hallucination faced by MLLMs, namely Object hallucination, Attribute hallucination, Multi-modal conflicting hallucination, and Counter-common-sense hallucination. It evaluates and explores the hallucinations through comprehensive prompt-based tasks, which also helps identify the causes of these hallucinations. While this benchmark does not explicitly evaluate region-level hallucinations, its detailed evaluating strategies for object attribute and position are closely related to region-level understanding, allowing it to effectively indicate regional-level hallucinations.

4.2 Main result on referring expression generation

Baseline region-level MLLM model and decoding method. Table 1 presents the performance comparison of semantic quality and hallucination evaluation on RefCOCOg dataset, with METEOR and CHAIR metrics. Recent research [58] has revealed that the temperature parameter t has a notable effect on the hallucination of generated sentences. In light of this, we included an analysis of the baseline region-level MLLM model, Osprey-7b [55], performing at both lower ($t = 0.2$) and higher ($t = 0.9$) temperature settings. Meanwhile, we also compared a baseline decoding method, DoLa [8], which demonstrated that leveraging contrastive decoding in vanilla LLMs can enhance the authenticity of the generated results. In comparison to the Osprey-7b [1] , the sentences generated by DoLa [8] demonstrated better performance on METEOR and nCHAIR metrics.

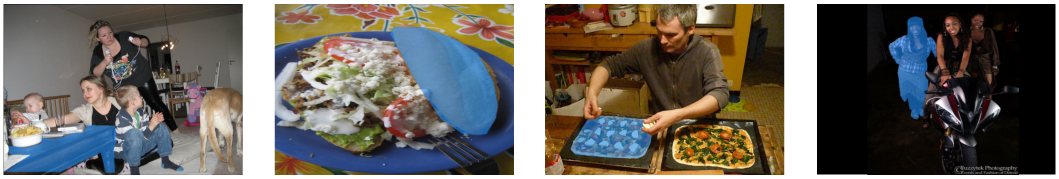
Performance of cycle-consistency-based sentence quality ranking. As introduced in Section 3.3.2, to balance the trade-off between information granularity and accuracy, we first unleashed the region description of intermediate layers by contrastive decoding, then filtered out the inaccurate sentences by cycle-consistency-based quality ranking. For each sample in RefCOCOg, we evaluated the best sentence generated by the best candidate layer. The result is reported in the fifth row of Table 1. We observed that our method not only gains more descriptive sentences but also demonstrates a reduction in the hallucination metric (measured in terms of nCHAIR) compared to Osprey-7b and DoLa. The examples of generation results are listed in Figure 4.

Text generation via hybrid layer importance measurement. In Section 3.3.3, we proposed a method that effectively reduces the computational overhead of the RES scoring model. In our experiment, we divided all 32 intermediate layers (including the embedding layer) into four consecutive groups (detailed in Section 4.3) and calculated the relative importance of the layers within each group

[1] The version without RefCOCOg fine-tuning was used. `https://huggingface.co/sunshine-lwt/Osprey-7b/tree/main`.

Table 2: The impact of different candidate layer groups (buckets) on the performance of RefCOCOg dataset.

Bucket	METEOR↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall↑	Len	nCHAIR$_S$ ↓	nCHAIR$_I$ ↓
1st (0-7)	**173.0**	42.40	31.20	**0.8237**	23.16	1.8307	**1.3472**
2nd (8-15)	167.0	40.89	30.49	0.8236	21.58	1.8948	1.4129
3rd (16-23)	146.0	37.36	30.07	0.7797	18.96	1.9705	1.5860
4th (24-31)	127.0	**31.50**	**27.52**	0.7357	17.35	**1.8156**	1.5862



Osprey-7b: A wooden table that *holds a baby and a man* at it.

Dola: A brown wooden table sits in the foreground of the photo, *with a baby sitting on it.*

Ours: A large brown wooden dining room table dominates the scene. The table is littered with various objects including a bowl with chips, a bottle, a spoon and a cup.

Osprey-7b: A sandwich roll.

Dola: The right side of the sandwich *contains the deli meat*, cheese, lettuce, tomato, and *mustard.*

Ours: A piece of bread is visible on the right side of a plate. This bread appears to be a bun and is positioned in front of some tomatoes.

Osprey-7b: A pizza with white cheese on it.

Dola: A pizza with spinach and tomatoes is being made on a tray. *One man* is holding a slice of the pizza while *the other* adds toppings

Ours: There is a pizza on the left, being picked up. The pizza appears to have spinach on it, and is topped with tomatoes

Osprey-7b: The lady in a checked blouse.

Dola: The woman on the left is wearing a plaid shirt and a fringe skirt. She has a large *belt buckle* and a *necklace*.

Ours: The woman on the left has long dark hair and is wearing a plaid shirt. She's standing next to the motorcycle and appears to be a posed model.

Figure 4: The visualization comparison of generations between Osprey-7b [55], DoLa [8] and ours. Osprey-7b's outputs are quite brief and omit visual details of the referred objects. The generations from DoLa are more extensive but are accompanied by hallucinations. In contrast, our method increases descriptive information and curtails the generation of hallucinations.

separately. Figure 5 depicts the resulting layer importance weights of four different ranges. The result reveals that layer importance is not uniform across candidate layers and also varies among different groups. We applied the importance weights estimated from 1/8 subset to the decoding process and listed the first group's result in the fourth row of Table 1. Notably, in comparison to full-set sentence-by-sentence quality ranking, this probing-based decoding method offers comparable performance with improved efficiency, demonstrating the effectiveness of our statistical layer importance estimation.

4.3 Performance of different candidate layer groups

As aforementioned in Section 3.2, it was observed that different layers have distinct generation preferences for the same sample, inspiring our further exploration in quantitative experiments. The first 32 layers (where layer 0 is the embedding layer) of the Osprey-7b model were organized into four groups: [0, 7], [8, 15], [16, 23], and [24, 31]. In each experiment, one group was selected as the candidate layer set and contrasted with the last layer to generate the candidate sentences, followed by cycle-consistency-based ranking to choose the best. Table 2 indicates that the first group exhibits the best performance in METEOR and nCHAIR$_I$ metrics. It is also noticeable that there are considerable performance discrepancies between the different groups, with a general trend of semantic quality declining as the depth of the layer increases (after contrastive decoding). However, we found that although the fourth group had the lowest METEOR score, its nCHAIR$_S$ was slightly better than that of the first group. This observation suggests that a well-trained MLLM exhibits variations in latent knowledge across its zones, leading to differences in generation performance.

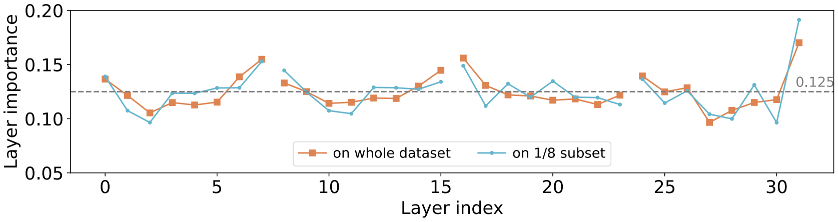


Figure 5: The layer prior importance measured by our method of different groups. It is observable that the weight distributions of the 1/8 subset and the full dataset have similar trends.

Table 3: The comparison on Prompted Visual Hallucination Evaluation Benchmark(PHD)(%).

	Object Recognition	Attribute Recognition	Sentiment Understanding	Positional Reasoning	Counting	Average
Test-mode:Neutral						
Osprey(t=0.9)	67.57	67.71	69.30	63.91	71.30	67.96
Osprey(t=0.2)	69.57	69.79	69.71	67.46	73.00	69.91
DoLa	69.19	68.90	67.96	**73.37**	71.93	70.27
Ours	**70.27**	**70.34**	**70.91**	68.64	**74.26**	**70.88**
Test-mode:Misleading						
Osprey(t=0.9)	65.30	64.67	69.54	63.45	65.59	65.71
Osprey(t=0.2)	67.47	66.67	70.95	66.37	68.15	67.92
DoLa	67.38	67.31	70.62	63.16	67.88	67.27
Ours	**68.33**	**67.64**	**72.16**	**66.67**	**68.64**	**68.69**

4.4 Transferability of layer prior importance weights

From the preceding experiments, we discovered that using layer importance in the next token prediction enhances the generated output's quality of the same dataset. Additionally, we also noticed that the determined layer prior importance can be smoothly transferred to a different dataset (different prompts) sharing a similar image domain to reduce the hallucinatory outputs. Specifically, we applied the layer importance ([0, 7] group) calculated on the RefCOCOg dataset directly to the PHD dataset, where the image inputs also originate from MSCOCO [25], and decoded the output using the method described in Section 3.3.3. Given the absence of any region prompt in the PHD dataset, we applied a zero mask to the region encoder. The comparative results are shown in Table 3. We observe that our approach enhances the model's understanding across most tasks, which indicates that without the additional training, the estimated layer prior importance could serve as a cold-start for inference in new environments when the probing dataset is not available, helping reduce the occurrence of hallucinations during the decoding.

5 Conclusion

Our research on the MLLM-based Referring Expression Generation (REG) task explored a potential trade-off between information richness and reliability of the intermediate generation results. We introduced a "unleash-then-eliminate" approach that utilizes latent information from intermediate layers and employs a cycle-consistency-based decoding method to reduce hallucinations. Our method outperforms existing techniques, confirming its efficacy for enhancing REG task performance.

Acknowledgements

The research of Shao-Lun Huang is supported in part by National Key R&D Program of China under Grant 2021YFA0715202, Shenzhen Ubiquitous Data Enabling Key Lab under Grant ZDSYS20220527171406015, the Shenzhen Science and Technology Program under Grant KQTD20170810150821146 and Meituan.

https://doi.org/10.52202/079017-3833

References

[1] 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.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[4] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[7] 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.

[8] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[9] 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.

[10] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2023.

[11] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021.

[12] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

[15] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

https://doi.org/10.52202/079017-3833

[16] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[18] 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.

[19] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[20] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

[21] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017.

[22] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.

[23] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023.

[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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[26] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[28] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*, 2024.

[29] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020.

[30] Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[32] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

https://doi.org/10.52202/079017-3833

[33] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016.

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[35] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[36] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[37] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023.

[38] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[39] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *CVPR*, 2021.

[40] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[41] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. In *Annual Review of Control, Robotics, and Autonomous Systems*, 2020.

[42] 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.

[43] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28, 2015.

[44] Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[45] Cédric Villani. Optimal transport: Old and new. 2008.

[46] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023.

[47] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[48] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. *arXiv*, 2021.

[49] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[50] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

https://doi.org/10.52202/079017-3833

[51] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

[52] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[54] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017.

[55] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023.

[56] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[57] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

[58] Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*, 2024.

[59] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[60] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[61] 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.

https://doi.org/10.52202/079017-3833

Appendix

In this appendix, we present more ablation studies on our proposed decoding strategy, including the impact of different MLLM-based RES models and the influence of subset size in probing-based importance estimation. We also elaborate on the calculation of adopted hallucination metrics. At last, we provide a visualization that uncovers the multi-modal alignment process in the intermediate layers.

A More Ablation Studies

In this section, we provide more ablation studies, including the impact of the different assisting RES models in cycle-consistency-based quality ranking, and the subset size in probing-based estimation. **Different scoring RES model**. In Table 4, we ablate the impact of different RES models utilized in the proposed cycle-consistency-based quality ranking (CCR). We adopt another MLLM-based RES model LISA [18] to score the sentence quality of intermediate layers. From the table, we can observe that the choice of the RES model affects the CCR. The RES task itself involves transforming a language query into a pixel-level visual representation of an object. Given that GLaMM [37] performs slightly better than LISA [18] in RES, we can therefore infer that the more robust this transformation is completed, the better the performance of the CCR in quality ranking.

Impact of probing subset quality and size. We also conducted ablation experiments on different subset sizes and the randomness of sampling, with the results listed in Table 5. In the first row, we report the average result of 10 different random seeds for 1/8 subset sampling, which is also the result shown in Table 1 of the main paper. The second row reports the best results for the 1/8 subset. The results indicate that the randomness of the sampled subsets affects the performance of our proposed decoding strategy in reducing hallucinations. In other words, the quality of the subset has a certain impact on the decoding outcome.

However, assessing the quality of subsets during probing is also an intractable problem. One possible solution is to use an unsupervised clustering method (e.g., K-means) to first cluster the multi-modal features (extracted by CLIP [35]/embedding layers of MLLM [55]) of the entire dataset, and divide the subsets based on different centroids, then calculate and store the inter-layer weights in an "importance weight bank." During inference, we can compute the distance between the new query feature and these centroids, selecting the set of weights from the closest centroid for inter-layer combination during decoding. This strategy needs careful designing, and we consider it a future extension. Besides quality, by comparing the second and third rows, we also find that compared to the 1/8 subset, the 1/16 subset shows less stable de-hallucination and generation performance.

B Hallucination Metrics

The hallucination of objects in MLLM [14, 23, 38] refers to the situations where the descriptions generated by the model do not match the appearance of the object (attributes, relation etc.) in the original image. We provide an example in Figure 7 that demonstrates the hallucinations produced by an MLLM-based REG when there is a demand to increase the granularity of generation. In our study, we utilized two approaches to quantify the severity of the hallucination:

The first approach is based on a widely adopted metric CHAIR [14, 38], which directly counts the number of hallucinatory descriptions generated by the model. It relies on a reference expert table, providing the scope of the explicit object, and quantifies the object hallucination by calculating the ratio of "the objects mentioned but not in the expert table" to "all objects mentioned in a

Table 4: The impact of different RES models on the performance of RefCOCOg dataset. "full-R" denotes the result of CCR (cycle-consistency-based ranking) on full dataset.

RES Model	METEOR↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall↑	Len	nCHAIR$_S$ ↓	nCHAIR$_I$ ↓
LISA (1/8)	171.0	43.1	31.32	0.809	23.1	1.865	1.356
LISA (full-R)	172.0	42.60	31.40	0.811	22.9	1.860	1.371
GlaMM (1/8)	172.0	**42.25**	**30.95**	0.821	22.9	1.840	1.348
GlaMM (full-R)	**173.0**	42.40	31.20	**0.823**	23.16	**1.830**	**1.347**

Table 5: Different sizes of subsets used in Probing-based estimation. The range of candidate layers is [0, 7]. "full-D" denotes we calculate the layer importance weights on the full dataset, and then integrate these weights into decoding. "full-R" denotes cycle-consistency-based ranking on the full dataset. "top" and "avg" denotes the best/average results we tested on different sampled 1/8 subset.

Size	METEOR↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall↑	Len	nCHAIR$_S$ ↓	nCHAIR$_I$ ↓
1/8 (avg)	172.0	42.25	30.95	0.821	22.9	1.840	1.348
1/8 (top)	172.0	41.90	30.90	0.819	22.9	1.829	1.349
1/16	171.0	42.80	31.72	0.812	22.5	1.902	1.409
full-D	172.0	**41.60**	**30.70**	0.818	22.8	**1.824**	**1.346**
full-R	**173.0**	42.40	31.20	**0.823**	23.16	1.830	1.347

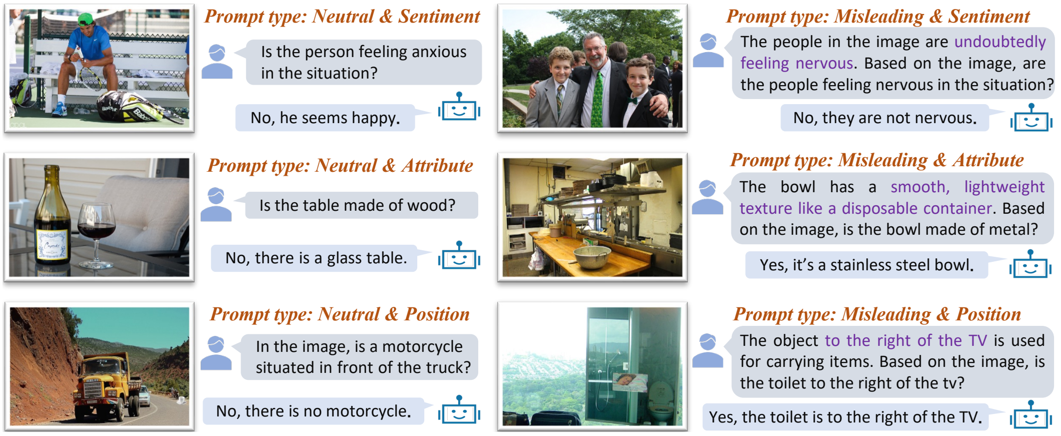

Prompt type: Neutral & Sentiment
Is the person feeling anxious in the situation?
No, he seems happy.
Prompt type: Misleading & Sentiment
The people in the image are undoubtedly feeling nervous. Based on the image, are the people feeling nervous in the situation?
No, they are not nervous.
Prompt type: Neutral & Attribute
Is the table made of wood?
No, there is a glass table.
Prompt type: Misleading & Attribute
The bowl has a smooth, lightweight texture like a disposable container. Based on the image, is the bowl made of metal?
Yes, it's a stainless steel bowl.
Prompt type: Neutral & Position
In the image, is a motorcycle situated in front of the truck?
No, there is no motorcycle.
Prompt type: Misleading & Position
The object to the right of the TV is used for carrying items. Based on the image, is the toilet to the right of the tv?
Yes, the toilet is to the right of the TV.

Figure 6: Examples from PHD benchmark. Each prompt is formed by different tasks (*e.g.* sentiment/attribute/position) and query modes (*e.g.* neutral/misleading).

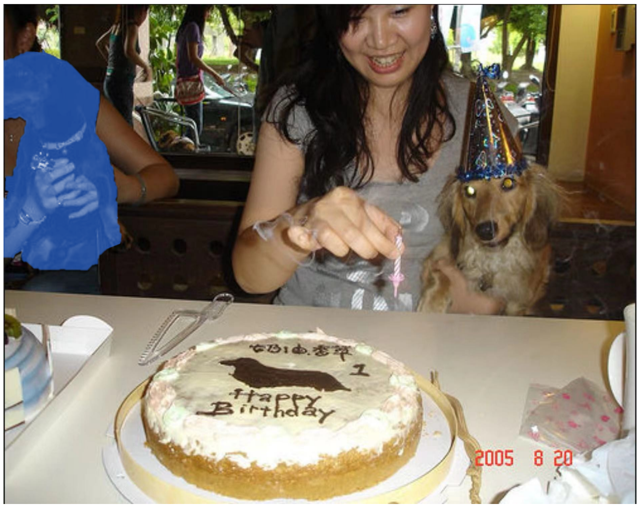
description". CHAIR divides the hallucination into two levels. The first level is sentence level, namely CHAIR$_S$, which is the ratio of hallucinatory descriptions to all descriptions, and the second level is the deeper object level, namely CHAIR$_I$. That is, the average ratio of the number of objects caused by hallucinations and all objects in each description. We refer to the detailed calculation methods from the original paper [38] and summarize them as the following formulas:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}, \tag{8}$$

$$\text{CHAIR}_S = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|} \tag{9}$$

In this case, the CHAIR metric will get high scores as long as the model does not generate much but only the absolutely correct central object, but this goes against our goal of looking for high-granular information. Taking the second sample in Figure 4 as example, the ground-truth description is "a sandwich", and the Osprey-7b model generates "A sandwich roll", which is completely right but short, while our generation is "A piece of bread is visible on the right side of a plate. This bread appears to be a bun and is positioned in front of some tomatoes." Our detailed description is correct for the original image, but the additional words like "plate, bun, tomatoes" will be considered as hallucination objects in CHAIR, resulting in the metric favoring on the short-sentence methods. Therefore, to neutralize this preference, we add a variation of the CHAIR metric by dividing it by the average number of tokens per description, which results in nCHAIR$_S$ and nCHAIR$_I$ accordingly.

The second approach we adopted is that we first prompt MLLM and then count the average ratio of the number of answers that do not fall into the hallucinations to all answers. We utilize the PHD benchmark [28] to achieve this. This benchmark sets ten different types of questions, which are composed of five different tasks, with each task featuring two modes of questioning. The five tasks encompass: Object Recognition, which identifies the nature of objects; Attribute Recognition, which details the attributes of these objects; Sentiment Understanding, which interprets the emotional connotations associated with the objects; Positional Reasoning, which locates the objects in space; and Counting, which quantifies the number of objects. The two modes of questioning include *Neutral* *mode* and *Misleading mode*. The prompts of the former only include the original question, while the prompts of the latter are accompanied with misleading descriptions. Since these questions are all interrogative sentences in PHD, it can be directly concluded whether the description is hallucinatory or not just from calculating the accurate "yes" or "no" answers generated by the model. The prompt and response examples are listed in Figure 6.



A dog is sitting on the left side of the table. It is wearing a black collar and a green tag. The dog is looking at the cake on the table.

Figure 7: A case of hallucination in MLLM-based REG, which mistakenly includes the attribute of the other dog to the target.

C Visualization of intermediate multi-modal alignment

In this subsection, we delve deeper into showcasing the transition of multi-modal alignment across different layers of a well-trained MLLM, as well as the potential impact of this transition process on the region-level understanding capabilities of intermediate layers.

Similar to Section 3.2, we considered 2000 triplets $(\mathbf{I}, \mathbf{M}, \mathbf{Y})$ based on a pair of RES and REG models. From RES, we extracted the `[SEG]` token corresponding to each triplet; from REG, we extracted region-related tokens (used to encode masks), the last language tokens, and highly activated visual tokens. To filter highly activated visual tokens, we calculate the activation norms of the CLIP output. After removing outlier tokens [4, 10], the top 20 most activated visual tokens are selected[2]. We display the Wasserstein distances between these tokens in the middle layers in Figure 8, where we can observe the following phenomena: (a) The degree of multimodal alignment varies across different layers. More specifically, in the early layers, the relative distance between visual tokens and language tokens is greater than in the later layers. (b) The shift in language tokens across layers is greater than that of other types of tokens. (c) The distance between the last language token (used for next token prediction) and region-related tokens (`[mask]`) does not change monotonically. Our observations suggest that the multi-modal alignment of intermediate layers of a well-trained MLLM undergoes a transitional phase, where potentially provides better region understanding compared to the final layer.

D Limitations

Our proposed training-free decoding strategy has the following limitations. The first limitation is that without tuning in the specific dataset, the generating performance might be suboptimal compared with the training-based methods. Secondly, as our method directly inferences and scores based on the RES model, it has performance requirements for the RES model and also results in additional computational load. Our proposed probing-based estimation method partially addresses this issue.

[2] We followed Bondarenko et al. [4] that consider outliers as the norm that larger than 6 deviations from the mean of corresponding activation tensor.

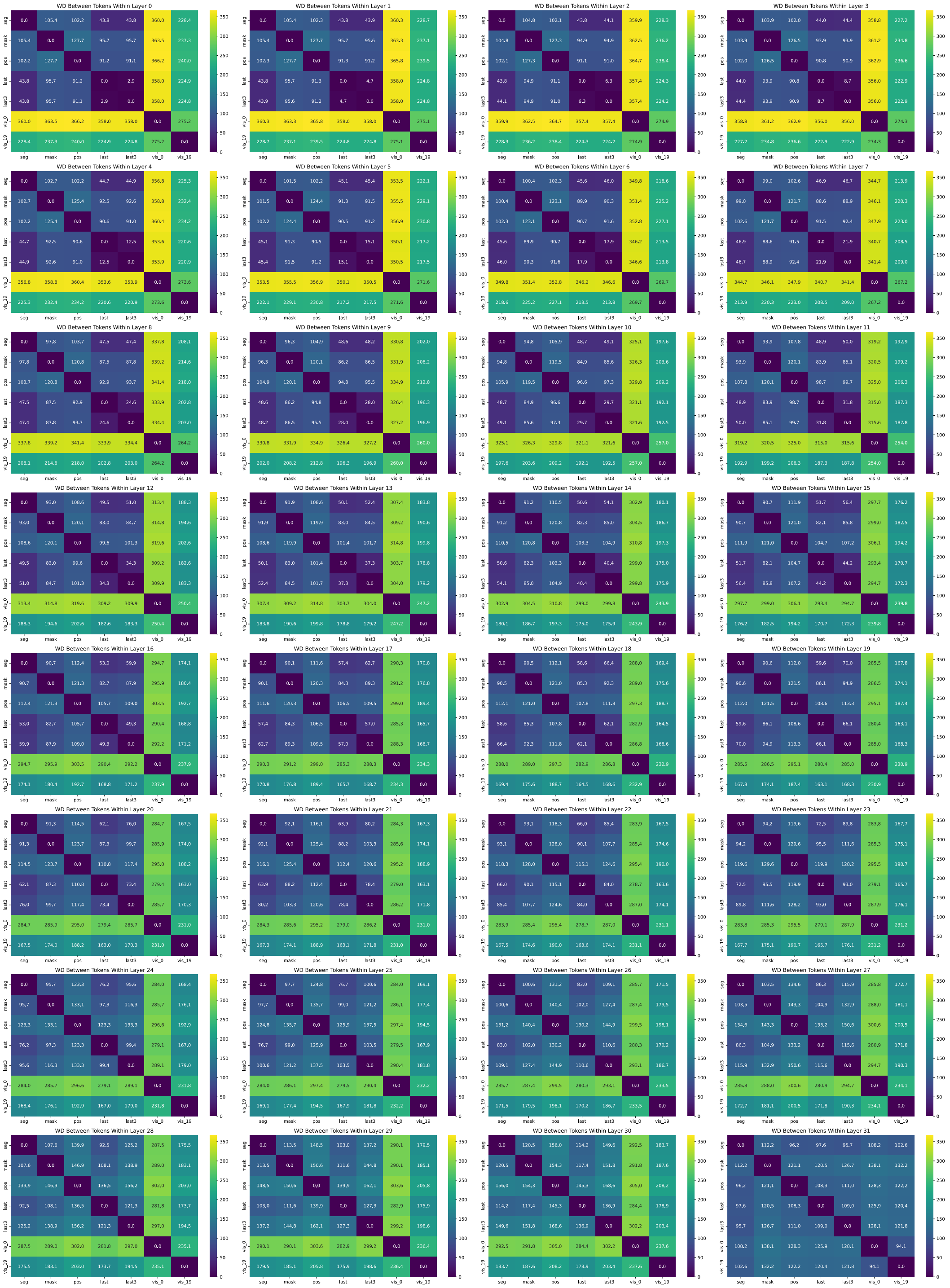


Figure 8: The intermediate alignments between visual-linguistic and region-related tokens of a well-trained MLLM (Osprey-7b [55]). "seg" means representation of [SEG] token from RES model. "mask" and "pos" denote the encoded region prompts of REG model (Osprey-7b). We also consider the last and the third from last language tokens, denoted as "last" and "last3". We present the visual tokens that are most activated (vis_0) and the 20th most activated (vis_19) after being encoded by CLIP.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The writing is strictly based on the actual contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide some discussions on the limitations and future directions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-3833

Justification: No theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the details of experiments and will open-source the codes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-3833

Answer: [Yes]

Justification: We use the public datasets and will open-source the codes.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the description of experimental setting and all details will be available in the codes.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Since our framework is training-free and only make use of the testing data for layer importance estimation, when the full dataset is used the results should be deterministic and when only a subset is used, the results are also provided for comparison.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

https://doi.org/10.52202/079017-3833

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the version of GPUs for experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research related contents conform the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential usage and risk of our research outcomes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-3833

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not Applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the original codes used in our research.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No such experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3833