

Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective

Zihao Yue

Renmin University of China
yzihao@ruc.edu.cn

Liang Zhang

Renmin University of China
zhangliang00@ruc.edu.cn

Qin Jin*

Renmin University of China
qjin@ruc.edu.cn

Abstract

Large Vision-Language Models (LVLMs) often suffer from multimodal hallucinations, wherein they may create content that is not present in the visual inputs. In this paper, we explore a new angle of this issue: overly detailed training data hinders the model’s ability to timely terminate generation, leading to continued outputs beyond visual perception limits. By investigating how the model decides to terminate generation with EOS, the special end-of-sentence token, we find that the model assesses the completeness of the entire sequence by comparing the generated text with the image. This observation suggests that the model possesses an inherent potential of making proper EOS decisions based on its visual perception to avoid overly lengthy outputs. To take advantage of such potential, we explore two methods to mitigate multimodal hallucinations: a training objective that enables the model to reduce hallucinations by learning from regular instruction data, and a data filtering strategy to prevent harmful training data from exacerbating model hallucinations. Both methods significantly improve the hallucination performance of LVLMs, without requiring any additional data or knowledge.¹

1 Introduction

Ever since Large Vision-Language Models (LVLMs) (Yin et al., 2023a) were achieved through bridging vision encoders with Large Language Models (LLMs) (Brown et al., 2020; Zhao et al., 2023a), they have been plagued by the problem of multimodal hallucinations, i.e., their text outputs may include unfaithful content to the visual inputs, such as non-existent objects (Rohrbach et al., 2018), which greatly harms the reliability of their applications. Extensive research has shed light on the origins of multimodal hallucinations, including

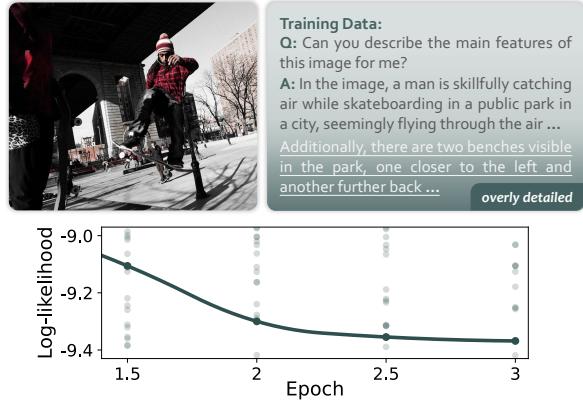


Figure 1: **Top:** An example from the LLaVA instruction data. The training data can be overly detailed to exceed the model’s visual perception limits. **Bottom:** Average log-likelihood of the LLaVA (7b) model predicting EOS at positions labeled as EOS during instruction tuning. Training the model with overly detailed data leads to a decrease in its tendency to stop generation.

the inability of vision encoders to represent fine-grained visual details (Jiang et al., 2023b; Tong et al., 2024), model reliance on inherent parametric knowledge such as language priors and statistical biases (Leng et al., 2023; Zhou et al., 2023), and pervasive hallucinations in the training data itself (Yu et al., 2023a; Liu et al., 2023a). In response to these insights, a variety of strategies have been proposed to mitigate hallucinations in LVLMs (Yin et al., 2023b; Zhai et al., 2023; Zhao et al., 2023b).

Although significant progress has been made, in this paper, we highlight a crucial but often overlooked source of hallucinations: the excessively detailed training data. For example, in the detailed image captioning task, the caption data for an image typically integrates rich visual semantics from multiple human annotations or vision expert models, and is rewritten into lengthy paragraphs by LLMs (Liu et al., 2023c), as shown in Fig. 1. These training data, while high-quality and meeting our expectations for detail, may exceed the

*Corresponding Author.

¹<https://github.com/yuezih/less-is-more>

visual perception capability of LVLMs, especially for subtle image features such as small or easily confusable objects. When trained with such data, in an attempt to fit the detail level and length distribution of ground truth captions, the model may risk expressing details that it cannot discern from the image, and therefore exhibit hallucinations.

Ideally, models should be trained to terminate generation upon reaching their visual perception limits to avoid hallucinations. However, because gauging such limits is not trivial, it is difficult to provide explicit supervision to teach models to stop generation timely, and it is impossible to construct training data that well matches model capabilities. Fortunately, we can draw inspiration from a closer examination of the model’s decisions regarding the generation of the end-of-sentence (EOS) token.

We first employ a saliency-based method to analyze how information flows from the context to the target position where the model predicts the next word. We discover that in predicting the EOS token, the model tends to rely more on all preceding sentences rather than just the current sentence. This leads to a hypothesis that **the model assesses the completeness of the entire sequence when deciding whether to terminate the generation**. Then, by manipulating the context, we observe that the model’s tendency to predict EOS clearly varies depending on the semantic completeness of the generated text relative to the visual input. For instance, reducing visual context (easier to reach textual completeness) makes the model more likely to end the generation, whereas concealing textual context (further away from textual completeness) prompts continued generation. This confirms the hypothesis above and implies that such a completeness assessment is accomplished by comparing the generated text with the perceived visual information. These observations suggest that the model inherently holds the potential to make timely EOS decision to terminate the generation based on its visual perception. When the model decides to end the generation, it indicates that the current generated context sufficiently captures the visual information it can perceive, and any further outputs may exceed the model’s visual perception limits, possibly leading to hallucinations.

To unlock such potential of models, we explore two approaches to enhance the model for better EOS decisions. (1) A learning objective for model training, termed **Selective EOS Supervision**. Simply modified from the Maximum Likelihood Es-

timation (MLE), it enables the model to mitigate hallucinations through learning from regular instruction data. It is applicable both for further training to reduce hallucinations in existing models, and for initial instruction tuning (Ouyang et al., 2022) to alleviate the onset of hallucinations. Specifically, by briefly further training on the original instruction data, the sentence-level and instance-level hallucinations of LLaVA-1.5 (Liu et al., 2023b) are reduced by 26% and 27%, respectively. (2) A data filtering strategy based on **Scoring EOS Supervision**, to eliminate harmful training data that can impair the model’s ability to end sequences. We design two metrics to assess the positive and negative impact of data on the model’s EOS tendency, and combine them to rank and filter the training data. Experimental results show that removing a small portion of the data can significantly reduce the hallucinations of models trained on it. These findings further validate our hypothesis and provide simple yet effective solutions for mitigating multimodal hallucinations in LVLMs.

2 EOS Decision

In autoregressive language models, sequences are completed through continuous next-token prediction (NTP). The termination of the process is achieved by introducing a special end-of-sentence (EOS) token, denoted as v_{EOS} , into the vocabulary. At each NTP step, the model chooses between a regular content token and v_{EOS} , deciding whether to continue the sequence generation or end it, which we refer to as EOS decision (Newman et al., 2020).

In this section, we delve into how LVLMs reach EOS decisions. Specifically, in Section 2.1, we analyze the contextual information that the model relies on to predict v_{EOS} ; in Section 2.2, we explore how the model adjusts its tendency to terminate generation with the multimodal input. Corresponding findings are further discussed in Section 2.3.

2.1 Information Basis of EOS Decision

We first investigate where the information for EOS decisions comes from in the context. Given that the context usually contains a long paragraph with multiple sentences, we group the context tokens into three parts: image tokens, preceding sentences, and the current sentence, to observe their respective contributions to the model’s prediction decision. For comparison, we also examine the information for non-EOS target predictions occurring in the

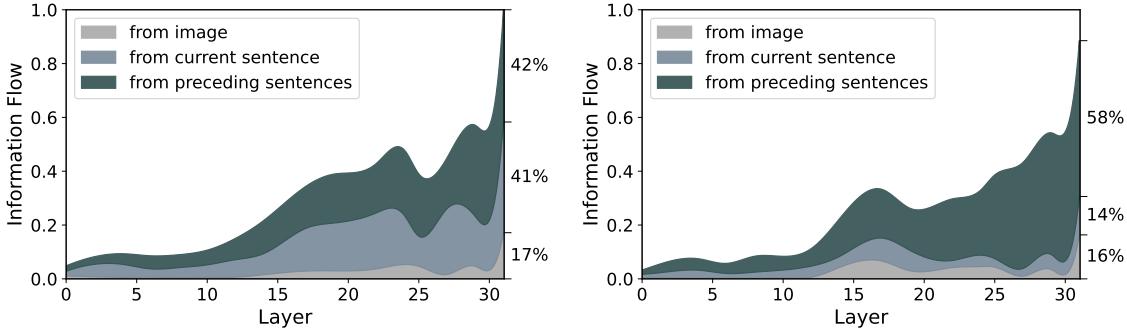


Figure 2: Significance of the information flows from different parts of the input context to the target position during the prediction of a random token (left) and the EOS token (right). The significance refers to the proportion of these information flows out of a layer’s total flows.

middle of a sequence, where the model needs to predict a regular content token. Since the context exposed for v_{EOS} prediction is the entire sequence, for a fair comparison, non-EOS targets are randomly selected from the last 10 tokens of the last sentence in the sequence. This ensures that they have access to all previous sentences while perceiving a sufficient portion of the current sentence.

We adopt the saliency score (Simonyan et al., 2013; Michel et al., 2019) as the metric for investigation. The saliency score of a token represents the sensitivity of the model to this token, i.e., how much its change affects the model prediction. As suggested by Wang et al. (2023b), we use the saliency score to quantify the information flow between tokens. Concretely, we feed the model with the first $n - 1$ tokens to predict the n -th target token through a forward pass, and obtain the cross-entropy loss $\mathcal{L}(x)$ at the n -th target position. The saliency matrix I is given by

$$I = \left| A \odot \frac{\partial \mathcal{L}(x)}{\partial A} \right|,$$

where A denotes the self-attention score matrix of the language model, \odot means element-wise product, and $I(i, j)$ reflects the significance of the information flow from the j -th token to the i -th token. We compare the information flow patterns of EOS predictions and non-EOS predictions to elucidate the information basis of EOS decisions.

Implementation details. We choose the 7b version of LLaVA-1.5 as the model, containing a language decoder with 32 layers and 32 attention heads. The saliency matrix per layer is derived by averaging across all heads. The data used for investigation comes from Detail23K, a subset of the LLaVA Instruction dataset (Liu et al., 2023c), containing

23K detailed image descriptions for instruction tuning. For each run, we calculate the expectation over a random sample of 500 data entries.

Results. As illustrated in Fig. 2, we first observe a pronounced information flow from contextual tokens to the target position, especially at higher layers (near the output). This implies a clear information aggregation pattern for model prediction. Then, we want to figure out where the information used for prediction comes from. As shown in Fig. 2 (left), for non-EOS predictions, the significance of information flows from the current sentence is comparable to that from previous sentences, despite the latter being significantly longer. This indicates that the current sentence is of great importance to model predictions. However, when the model is tasked with predicting v_{EOS} , as depicted in Fig. 2 (right), the significance of information flows from previous sentences significantly increases and dominates. This suggests that the model, when predicting v_{EOS} , places more emphasis on integrating information from all already generated content. This distinctive behavior indicates that the model’s EOS decision is related to the current state of the entire sequence. Thus, we speculate that the model might be actively assessing the completeness of its text generation relative to its visual input, i.e., whether the current text is sufficient to describe its perceived visual information.

2.2 Semantic Comparison for EOS Decision

To validate the hypothesis from Section 2.1, we intervene in the multimodal input context and analyze the model’s tendency for EOS predictions. Please note that the EOS decision does not solely occur at the final position of a sequence but at every position. However, for a well-trained language model, EOS predictions typically occur at the end of each sen-

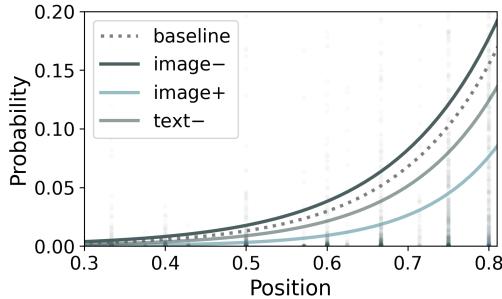


Figure 3: The predictive probability of v_{EOS} at various target positions, fitted by exponential functions. Position denotes the relative location i/N of the i -th target token among all N target tokens in the sequence.

tence, i.e., the position right after the period. Hence, we focus on these target positions. We employ the same data and model mentioned in Section 2.1 for analysis, and obtain the conditional probabilities of v_{EOS} at various target positions through a forward pass. Fig. 3 (dotted line) illustrates the model’s expected EOS tendency at each target position. A clear trend is that such a tendency increases as the sequence lengthens, implying the correlation between the textual richness and the EOS tendency. However, this correlation could stem from a variety of factors, for example, the length bias in training data can prompt the model to rely on positional embeddings for v_{EOS} predictions. To ablate potential disturbances, we additionally design three context manipulation methods:

- **Visual reduction (*image-*):** Applying a Gaussian noise mask to the input image, to reduce recognizable semantics in the image.
- **Visual augmentation (*image+*):** Concatenating the image with a random one, to introduce visual information not described in the current text.²
- **Textual reduction (*text-*):** Using an attention mask to hide a portion of the exposed text. Here, we mask the first 30 tokens to ensure the coherence of the adjacent context for v_{EOS} predictions in the end part of the sequence.

These manipulation methods enable the augmentation or reduction of the multimodal contextual semantics without altering the sequence length.

Results. As illustrated in Fig. 3, the reduction of image information through noise notably increases the model’s tendency to predict v_{EOS} . Conversely, introducing new image information or concealing text information, both implying a reduction in the

²We also implement a variant that replaces the input image with a random new one instead of concatenation, to avoid increasing the absolute information richness (see Appendix B.2).

relative textual completeness, lead to a decreased tendency of v_{EOS} prediction. These observations further support our conjecture that the model tends to assess the completeness of the current text to make an EOS decision, particularly, by comparing the generated text to the input image. Specifically, the more completely the image is described, the more likely the model is to terminate generation.

2.3 Discussion

Our investigation on the information basis and the model’s intrinsic criteria for EOS decisions reveal that models consider the current state of the entire sequence (Section 2.1) and assess the completeness of generated text relative to the image (Section 2.2). These findings suggest that while models may fit the training data length distribution and generate text beyond their capability limits, they still retain the inherent potential to adjust generation length according to visual perception. When the model tends to terminate generation, it can imply that the currently generated text adequately describes the visual information that the model can perceive. In Section 3, we explore how this potential can be harnessed to mitigate multimodal hallucinations.

3 Mitigating Multimodal Hallucinations

Inspired by the preceding analysis, we propose two approaches to mitigate multimodal hallucinations: (1) a learning objective, namely Selective EOS Supervision (Section 3.1), which unlocks the model’s capability to make EOS decisions at proper positions, thereby mitigating hallucinations; (2) a data filtering strategy, namely Scoring EOS Supervision (Section 3.2), which eliminates training data that may hinder the model’s capability to terminate generation in a timely manner.

3.1 Selective EOS Supervision for Training

The instruction tuning of LVLMs typically utilizes Maximum Likelihood Estimation (MLE) as the training objective. Given the visual content v and previous tokens $w_{<}$, the model predicts a probability distribution $P^V = \{p_1, p_2, \dots, p_{|\mathcal{V}|}\}$ over the vocabulary \mathcal{V} to determine the next word, where p_j represents the probability of the j -th word in \mathcal{V} . The model parameter θ is optimized to maximize the likelihood of the label word indexed y , with the loss function defined as:

$$\mathcal{L}_{\text{MLE}} = -\log(p_y | v, w_{<} ; \theta).$$

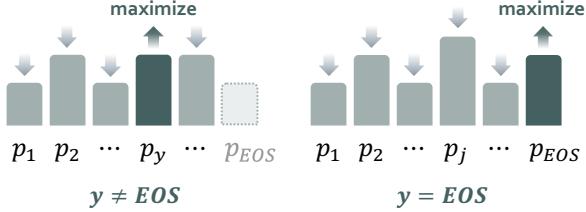


Figure 4: Illustration of the probability distribution derived from our proposed Selective EOS Supervision. Arrows indicate the maximizing and minimizing effects of the training objective on the probability of each word. When the label is not EOS, the EOS token is excluded from the probability distribution.

With such an objective, two optimization situations would happen regarding the v_{EOS} prediction: first, when the label is v_{EOS} , the model’s tendency to predict v_{EOS} will be enhanced; second, when the label is not v_{EOS} , and if the model assigns some probability to v_{EOS} , it will be penalized, becoming less likely to predict v_{EOS} . Recalling our analysis in Section 2, the model’s tendency for v_{EOS} prediction implies that the current text adequately represents its perceived visual information. Thus, in the second situation, stopping generation is the right choice. However, as the corresponding label is not v_{EOS} but a regular content token, the model will be discouraged from stopping and encouraged to continue generating content that may exceed its visual perception limits. Therefore, we aim to selectively preserve the first optimization situation, allowing the model to learn when to end generation, while minimizing the second optimization situation, to prevent compromising the model’s EOS decision ability due to overly detailed training data.

To achieve the aforementioned goal, we implement a minor modification to MLE. Concretely, at positions where the label is not v_{EOS} , we exclude v_{EOS} from the calculation of probability distribution. This means that the label’s probability is determined using a modified softmax operation:

$$p_y = \text{softmax}^*(\mathbf{z}_y) = \frac{\exp(\mathbf{z}_y)}{\sum_{j \in \mathcal{V} \setminus \{v_{EOS}\}} \exp(\mathbf{z}_j)},$$

where \mathbf{z} denotes logits. Since v_{EOS} does not participate in the probability distribution, it will not be suppressed by maximizing the label’s probability, as depicted in Fig. 4. This modification prevents MLE from undermining the model’s inherent tendency to predict v_{EOS} . For positions where the label is v_{EOS} , we retain vanilla MLE as objective, allowing the model to learn when to end sequences.

3.1.1 Experimental Settings

Models and datasets. Our training objective can be applied to any LVLMs with an EOS token and optimized by MLE. As a representative, we select two widely used open-source LVLMs, LLaVA (Liu et al., 2023c) and MiniGPT (Zhu et al., 2023) series. Among them, LLaVA, LLaVA-1.5 (Liu et al., 2023b), and MiniGPT-v2 (Chen et al., 2023a) are trained with data recipes that include the LLaVA-Instruction dataset. Thus, we validate our method with these models by fine-tuning them on LLaVA-Instruction. Additionally, our experiment results show that a smaller subset of LLaVA-Instruction-150K which contains detailed captions, Detail23K, has a similar effect but brings significant computational efficiency. Thus, most of our experiments are conducted with Detail23K. If not specified, models undergo one epoch of training with LoRA (Hu et al., 2022). Other training details remain consistent with the models’ official documentation.

Evaluation. Following previous works (Huang et al., 2023; Yu et al., 2023a), we evaluate model hallucination with Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018), a framework that quantifies object hallucination in image captions by comparing generated objects to the ground truth objects. The sentence-level score, CHAIR_S , represents the proportion of captions that contain hallucinations, and the instance-level score, CHAIR_I , denotes the frequency of hallucinated objects relative to all mentioned objects by the model. In addition, we measure an object recall to evaluate the semantic richness of generated captions. Our CHAIR tests involve 500 images randomly chosen from the MSCOCO validation set (Lin et al., 2014). We also adopt another metric FaithScore (Jing et al., 2023) to evaluate caption hallucination. It verifies the consistency between atomic facts in the caption and the input image with LLMs and visual expert models, for which we employ ChatGPT (OpenAI, 2023) and OFA (Wang et al., 2022). It also provides a sentence-level score, FaithScore_S .

Baselines. Since our method facilitates models to timely terminate generation, which often results in shorter responses, we incorporate baselines that simply reduce the generation length, including sequence truncating and decoding with a length penalty. The truncating method keeps only the initial $R\%$ of words in each caption, and the decoding method adopts an exponential length penalty

Table 1: Hallucination performance of different models. **w/ Cap.** and **w/ Inst.** denote fine-tuning models with the detailed caption subset, Detail23K, and with the full LLaVA-Instruction-150K, respectively. **Faith**: FaithScore.

Row	Model	Method	Length	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall \uparrow	Faith \uparrow	Faith _S \uparrow
1	LLaVA-1.5 (7b)	-	100.6	50.0	15.4	77.1	87.0	68.8
2		VCD	100.4	48.6	14.9	77.3	87.1	70.2
3		OPERA	98.6	47.8	14.6	76.8	88.0	72.6
4		OPERA (fast)	85.3	48.6	14.5	76.7	87.7	71.3
5		Ours (w/ Inst.)	76.2	36.8	11.3	74.3	88.4	73.0
6		Ours (w/ Cap.)	79.7	40.2	12.3	75.7	89.3	72.3
7	LLaVA-1.5 (13b)	-	100.9	47.2	13.0	77.3	87.6	73.1
8		Ours (w/ Cap.)	85.1	36.8	11.4	75.3	88.8	72.8
9	LLaVA (7b)	-	57.8	35.4	13.8	64.8	86.9	67.4
10		Ours (w/ Cap.)	39.9	27.0	13.2	57.1	88.9	71.6
11	MiniGPTv2 (7b)	-	87.2	38.0	11.1	66.3	85.6	67.8
12		Ours (w/ Cap.)	62.2	27.0	9.8	66.6	89.9	76.0

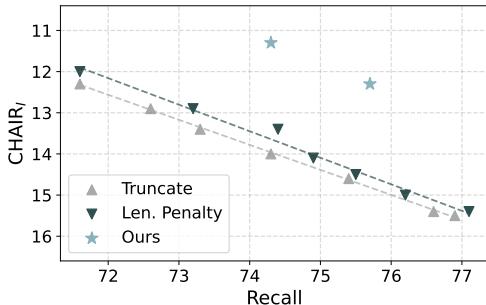


Figure 5: Hallucination vs. Recall performance of LLaVA-1.5 (7b). **Ours**: the models fine-tuned on **Inst.** and **Cap.** respectively with our training objective.

to adjust the score of the EOS token during generation. Varying the truncating proportion or the length penalty factor leads to different generation length, and affects both hallucination and recall performance. Additionally, we include two recently proposed plug-in methods: (1) Visual Contrastive Decoding (VCD) (Leng et al., 2023), which contrasts the output distributions derived from the original and noisy visual inputs, to reduce the influence of the model’s parametric knowledge. (2) Over-Trust Penalty and Retrospection-Allocation (OPERA) (Huang et al., 2023), a decoding strategy that penalizes the model’s over-reliance on certain tokens and allows roll-back when needed. We test VCD at different noise steps of 200, 500, 700, and 999, and report the optimal results. For OPERA, our implementation follows the suggestions in their released code, including two hyperparameter configurations for standard and fast inference.

3.1.2 Results

Versus the original model. As shown in Table 1, after a single training epoch on the detailed caption

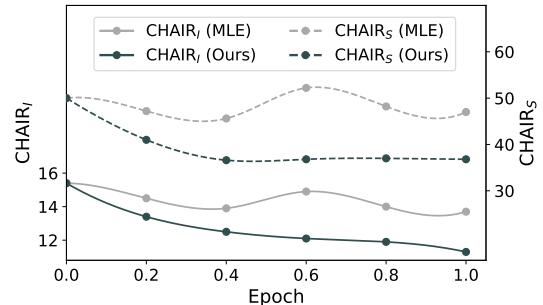


Figure 6: CHAIR performance trends of LLaVA-1.5 (7b) throughout training on LLaVA-Instruction-150K.

subset, Detail23K, using our learning objective, all models tend to produce shorter captions and notably reduce hallucinations at both the sentence and instance levels (r_1 vs r_6 , r_7 vs r_8 , etc.). This improvement is even more significant when using the full 150K instruction data (r_1 vs r_5), resulting in a 26.4% and 26.6% decrease in CHAIR_S and CHAIR_I of LLaVA-1.5 (7b), respectively. While our method leads to some decrease in recall (e.g., -1.8% of LLaVA-1.5 (7b) and -2.6% of LLaVA-1.5 (13b)), we view this as a beneficial compromise since the models become more conservative and less likely to “guess” uncertain visual content. More analysis can be found in Appendix B.3.

Versus baselines. As demonstrated in Fig. 5, the truncating and length-penalty decoding baselines, with varying length-controlling configurations, effectively reduce hallucinations at the cost of Recall. However, these methods fall short of ours for either being less effective in alleviating hallucinations or resulting in more significant recall loss. Our method also outperforms existing methods, i.e., VCD and OPERA, as shown in Table 1. Our proposed method does not require additional data

Table 2: CHAIR evaluation results of the LLaVA (7b) models instruction-tuned from scratch.

Loss	Length	CHAIR _S	CHAIR _I	Recall
MLE	57.8	35.4	13.8	64.8
Ours	36.1	24.2	<u>11.6</u>	55.9
Combined	42.7	<u>26.6</u>	11.0	57.5

construction or other expert models. Furthermore, unlike decoding methods, it does not slow down inference and remains technically compatible with various decoding strategies. Therefore, it presents a viable, practical supplement or alternative to current methods.

Versus MLE. To confirm that the improvement in model hallucination performance results from our modification to MLE, we also conduct a comparison by further training the model using the vanilla MLE. As shown in Fig. 6, the performance of the model optimized by MLE varies throughout the training and remains at the original level. This variation suggests that different training samples impact the model differently; some may enhance the model’s EOS tendency while others do the opposite, collectively preserving the model’s initial generation habits. In contrast, with our modified learning objective, the degree of model hallucination steadily decreases throughout one epoch of training, with the model eventually significantly outperforming its MLE counterpart. This indicates that our selective supervision consistently enhances the model’s EOS decision with varying data inputs.

Instruction tuning from scratch. Beyond further training existing models, our method is also compatible with instruction-tuning new models, starting from a vision-language aligned yet not instruction-tuned state. Table 2 presents the results of fine-tuning LLaVA for 3 epochs with LLaVA-Instruction-150K. With our learning objective, the model’s sentence-level and instance-level hallucinations are reduced by 31.6% and 15.9%, respectively. Combining our learning objective with the vanilla MLE at a 1:1 ratio achieves a more balanced performance between hallucinations and recall.

3.2 Scoring EOS Supervision for Training Data Filtering

As the preceding analysis shows, learning from overly detailed data can impair a model’s ability to predict v_{EOS} , so an intuitive solution is to filter out such “harmful” training data.

As described in Section 3.1, there exist two op-

timization situations regarding v_{EOS} prediction, corresponding to a positive effect that enhances the model’s EOS tendency when the label is v_{EOS} and a negative effect otherwise. We thus design two metrics to quantitatively evaluate the two effects on models when trained with a certain data sample:

$$S_{pos} = -\sum_{i=1}^N [y_i = v_{EOS}] \log(p_{v_{EOS}}|v, w_{<}; \theta^*);$$

$$S_{neg} = -\sum_{i=1}^N [y_i \neq v_{EOS}] \log(1 - p_{v_{EOS}}|v, w_{<}; \theta^*).$$

Here, θ^* is a reference model used for evaluating the data. For positions where the label is v_{EOS} , we define S_{pos} as the cross-entropy loss of the reference model predicting the label. A large cross-entropy loss indicates that, on this particular training data, the model fails to predict v_{EOS} to end the sequence, and the feedback from the training loss will enhance the model to learn this capability. Thus, S_{pos} quantifies the positive effect of the data on the model’s EOS prediction. Conversely, for positions where the label is not v_{EOS} , if the model tends to predict v_{EOS} , this tendency will be undesirably suppressed. Particularly, a larger $p_{v_{EOS}}$ leads to a more significant negative effect, especially when $p_{v_{EOS}}$ approaches 1. Therefore, as defined above, S_{neg} serves to estimate the negative effect of the data on the model’s EOS decision.

Intuitively, our goal is for S_{pos} to be as high as possible, indicating strong penalties for the model’s inability to predict v_{EOS} where it should, and for S_{neg} to be as low as possible, reflecting minimal suppression on the model’s EOS tendency. Therefore, we calculate a composite score $S_{final} = S_{neg} - S_{pos}$ to estimate the “harmfulness” of the data. It is recommended to remove the highest-scoring data parts from training to achieve a more desired outcome for appropriate EOS decisions. With the shared goal of preserving the model’s EOS decision capability, our data filtering strategy can serve as an alternative to the Selective EOS Supervision described in Section 3.1.

3.2.1 Experimental Settings

Data filtering. We apply the proposed data filtering strategy to the LLaVA-Instruction-150K dataset. The model used for scoring S_{final} is the instruction-tuned version of LLaVA-1.5 (7b). We test three data filtering ratios, ranging from 10% to 30%, to remove data with the highest S_{final} . Additionally, we evaluate a random filtering strategy where 20%

Table 3: CHAIR evaluation results of models trained with different data. $C_{S/I}$ denotes $\text{CHAIR}_{S/I}$.

Row	Train. Data		Model Performance			
	Filter	Len.	Len.	C_S	C_I	Recall
1	Original	178.3	57.8	35.4	13.8	64.8
2	10%	171.7	63.7	35.4	14.0	64.5
3	20%	168.2	45.5	27.0	10.6	58.9
4	30%	166.7	49.2	29.4	<u>11.7</u>	58.0
5	Random	178.2	68.9	35.5	11.8	61.9
6	Reversed	176.8	100.6	46.6	18.9	68.6

of the data is removed randomly, as well as a reversed filtering strategy, with 20% of data with the lowest S_{final} being removed.

Models. We fine-tune the LLaVA (7b) model from scratch with the filtered data to validate their effectiveness. Following common practice, all models are trained for 3 epochs with a batch size of 128. We adopt QLoRA (Dettmers et al., 2023) to reduce computational load. Note that in this subsection, models are trained with the vanilla MLE.

3.2.2 Results

As shown in Table 3, by removing a small proportion, i.e., 20%, of “harmful” data from the original training set, the model significantly reduces learning hallucinations during instruction tuning, resulting in the sentence-level and instance-level hallucinations reduced by 23.7% and 23.2%, respectively. In contrast, the reversed filtering which removes the least “harmful” data leads to opposite effects, greatly exacerbating the model hallucination, while the random removal brings no improvements in sentence-level hallucination performance. This shows that our criteria used for data filtering well reflect the impact of the data on the model’s ability to end generation.

Another interesting finding is that filtering the data does not bring a big change in the length of the training data, but it does significantly affect the length of the model generation. For instance, the reversed filtering leaves the average length of the training data almost unchanged, but the average length of model generation nearly doubles. This implies that the impact of our data filtering strategy does not come from changing the length distribution of the training data. Instead, it affects the model through manipulating the EOS supervision, further validating our motivation.

3.3 Discussion

In this section, to fully exploit the potential of the model to properly end generation according to its visual perception, we propose to mitigate hallucinations by not suppressing its inherent EOS tendency, through a training objective (Section 3.1) and a data filtering strategy (Section 3.2). From a practical perspective, the former approach has the merits of broader applicability and easier deployment, especially when used to further train existing models. The latter has greater compatibility since filtered data can be paired with various training methods.

4 Related Works

Hallucination Origins. Investigations on the causes of multimodal hallucination in LVLMs identify three main factors: (1) Limited visual representations. For example, the visual encoders commonly employed in LVLMs depict abstract features while struggling to capture fine-grained visual details (Jiang et al., 2023b; Tong et al., 2024; Wang et al., 2023a). Jiang et al. (2023a) observe a modality gap persists between visual and textual features, despite vision-language alignment. (2) Models’ over-reliance on parametric knowledge, such as statistical biases and language priors, rather than on visual evidence (Zhou et al., 2023; Leng et al., 2023; Liu et al., 2023a; Zhai et al., 2023; Guan et al., 2023). (3) Inferior data for instruction tuning. This includes insufficient visual supervision (Chen et al., 2023b), a lack of positive/negative human feedback (Yu et al., 2023b), and the presence of hallucinations within the training data (Yu et al., 2023a; Liu et al., 2023a). Our paper identifies a new source of hallucinations: overly detailed training data hinders the model’s inherent EOS decision ability, further enriching existing explanations.

Mitigation Solutions. An effective way to reduce hallucinations is to construct high-quality **data**, including employing automatic data cleaning pipelines (Yu et al., 2023a), generating (Liu et al., 2023a) or rewriting (Wang et al., 2024) training data with LLMs, and integrating human feedback into annotations (Gunjal et al., 2023). **Training** approaches view hallucinatory data as negative examples, and adopt preference optimization (Zhai et al., 2023; Zhao et al., 2023b; Sun et al., 2023; Yu et al., 2023b; Li et al., 2023a) or contrastive learning (Jiang et al., 2023a) to enhance models’ resistance to hallucinations. **Inference** strategies focus on the decoding process, suppressing models’

reliance on parametric biases (Leng et al., 2023) or penalizing inferior attention patterns (Huang et al., 2023). Other works explore **posthoc-fixing** ways to rectify hallucinations in model outputs, by training a revisor model (Zhou et al., 2023), employing expert models (Yin et al., 2023b), and prompting the original model for self-correction (Lee et al., 2023). In this paper, we propose a new learning objective and a data filtering strategy, belonging to the training and data perspectives, respectively.

5 Conclusion

This paper investigates the multimodal hallucination issue in large multimodal models. We suggest that overly detailed training data can prevent the model from stopping generation at the appropriate time, thus leading to hallucinated outputs. By examining the model’s inner behavior of EOS prediction, we discover that the model inherently holds the potential to terminate generation based on its visual perception limits. To enhance such potential, we develop two approaches, a learning objective for training models and a data filtering strategy for selecting training data, both of which facilitate the model learning to timely terminate generation and significantly reduce hallucinations.

Limitations

This work presents a novel perspective on the origins of multimodal hallucinations in large multimodal models with corresponding solutions. However, it faces several limitations. First, it focuses solely on generative tasks, i.e., detailed image description, without covering hallucinations in broader tasks like classification-oriented Visual Question Answering (VQA). Second, our solutions are examined only on multimodal models, though technically they could also be applied to unimodal large language models. We leave this possibility for future exploration. Third, our solutions mitigate hallucinations by enhancing the model’s ability to timely conclude sequences. While effective, they address only the simplest source among various causes of hallucinations. Fully solving the problem of hallucination remains a substantial challenge.

Ethics Statement

This work focuses on reducing hallucinations in large multimodal models to enhance their reliability and trustworthiness. We have carefully considered the ethical implications of our work and an-

ticipate no significant ethical concerns. This work was carried out using publicly available and commonly used data and models, and our findings may inherit the biases and limitations carried in these resources.

Acknowledgements

We thank all reviewers for their insightful comments and suggestions. This work was partially supported by the Beijing Natural Science Foundation (No. L233008), the National Natural Science Foundation of China (No. 62072462), and the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. [Minigpt-v2: large language model as a unified interface for vision-language multi-task learning](#). *ArXiv preprint*, abs/2310.09478.
- Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023b. [Mitigating hallucination in visual language models with visual supervision](#). *ArXiv preprint*, abs/2311.16479.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *ArXiv preprint*, abs/2305.14314.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). *ArXiv preprint*, abs/2306.13394.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pages arXiv–2310.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. [Detecting and preventing hallucinations in large vision language models](#). *ArXiv preprint*, abs/2308.06394.

- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [De-noising diffusion probabilistic models](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. [Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation](#). *ArXiv preprint*, abs/2311.17911.
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023a. [Hallucination augmented contrastive learning for multimodal large language model](#). *ArXiv preprint*, abs/2312.06968.
- Dongsheng Jiang, Yuchen Liu, Songlin Liu, Xiaopeng Zhang, Jin Li, Hongkai Xiong, and Qi Tian. 2023b. [From clip to dino: Visual encoders shout in multi-modal large language models](#). *ArXiv preprint*, abs/2310.08825.
- Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. [Faithscore: Evaluating hallucinations in large vision-language models](#). *ArXiv preprint*, abs/2311.01477.
- Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. [Volcano: mitigating multimodal hallucination through self-feedback guided revision](#). *ArXiv preprint*, abs/2311.07362.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. [Mitigating object hallucinations in large vision-language models through visual contrastive decoding](#). *ArXiv preprint*, abs/2311.16922.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. [Silkie: Preference distillation for large visual language models](#). *ArXiv preprint*, abs/2312.10665.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). *ArXiv preprint*, abs/2305.10355.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. [Mitigating hallucination in large multi-modal models via robust instruction tuning](#). *ArXiv preprint*, abs/2306.14565.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. [Improved baselines with visual instruction tuning](#). *ArXiv preprint*, abs/2310.03744.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. [Visual instruction tuning](#). *ArXiv preprint*, abs/2304.08485.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. [Are sixteen heads really better than one?](#) In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.
- Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. 2020. [The EOS decision and length extrapolation](#). In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 276–291, Online. Association for Computational Linguistics.
- OpenAI. 2023. Introducing chatgpt. <https://openai.com/blog/chatgpt>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. [Object hallucination in image captioning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. [Aligning large multimodal models with factually augmented rlfh](#). *ArXiv preprint*, abs/2309.14525.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. [Eyes wide shut? exploring the visual shortcomings of multimodal llms](#). *ArXiv preprint*, abs/2401.06209.
- Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. [Evaluation and analysis of hallucination in large vision-language models](#). *ArXiv preprint*, abs/2308.15126.

- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. [Label words are anchors: An information flow perspective for understanding in-context learning](#). *ArXiv preprint*, abs/2305.14160.
- Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. [OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023a. [A survey on multimodal large language models](#). *ArXiv preprint*, abs/2306.13549.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023b. [Woodpecker: Hallucination correction for multimodal large language models](#). *ArXiv preprint*, abs/2310.16045.
- Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueteng Zhuang. 2023a. [Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data](#). *ArXiv preprint*, abs/2311.13614.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023b. [Rlfh-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback](#). *ArXiv preprint*, abs/2312.00849.
- Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. [Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption](#). *ArXiv preprint*, abs/2310.01779.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. [A survey of large language models](#). *ArXiv preprint*, abs/2303.18223.
- Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023b. [Beyond hallucinations: Enhancing lmlms through hallucination-aware direct preference optimization](#). *ArXiv preprint*, abs/2311.16839.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. [Analyzing and mitigating object hallucination in large vision-language models](#). *ArXiv preprint*, abs/2310.00754.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [Minigpt-4: Enhancing vision-language understanding with advanced large language models](#). *ArXiv preprint*, abs/2304.10592.

A Additional Implementation Details

A.1 Experiment in Figure 1

For the trend depicted in Fig. 1, the model undergoes fine-tuning on the LLaVA-Instruction-150K dataset over 3 epochs and is evaluated with the same data used in Section 2. At the initial stages of training, the model shows notable fluctuations in performance due to the lack of prior fitting to the instruction data. To better demonstrate how data affects the model’s EOS tendency, we focus on the latter half of the training period, where performance begins to stabilize.

A.2 Context Manipulation

In Section 2.2, we reduce the semantics of images by overlaying them with a Gaussian noise mask. This involves gradually introducing minor amounts of Gaussian noise over T steps, mirroring the forward diffusion process used in image generation tasks (Ho et al., 2020). Our implementation follows that of Leng et al. (2023), and we set T to 500 for analysis.

A.3 Computation

Table 4 presents the computational cost of model training on a setup with 8 NVIDIA RTX A6000 GPUs. Our proposed training objective and data filtering strategy do not introduce a noticeable increase in training costs. In this work, all experimental results are derived from single runs, with greedy search as the decoding strategy.

B Additional Results

B.1 Information Aggregation Pattern

In Section 2.1, we observe a significant information aggregation from the context to the target position during token prediction. This section further clarifies the details of this information aggregation pattern. At lower layers (near the input), the information from regular content tokens within a sentence converges at the sentence’s end, typically a period,

Table 4: The computational burden of model training. **Train. Strategy** means either further training models that have already been fine-tuned for instruction following or instruction tuning from scratch. **PEFT** stands for the Parameter-Efficient Fine-Tuning (PEFT) strategies we adopt.

Train. Strategy	PEFT	Model	Data	Epoch(s)	Total GPU Time
Further	LoRA	LLaVA-1.5 (7b)	Cap.	1	~0.6 h
		LLaVA-1.5 (7b)	Inst.	1	~9.0 h
		LLaVA-1.5 (13b)	Cap.	1	~3.0 h
		LLaVA (7b)	Cap.	1	~0.6 h
		MiniGPTv2 (7b)	Cap.	1	~2.5 h
From-scratch	QLoRA	LLaVA (7b)	Inst.	3	~11.0 h

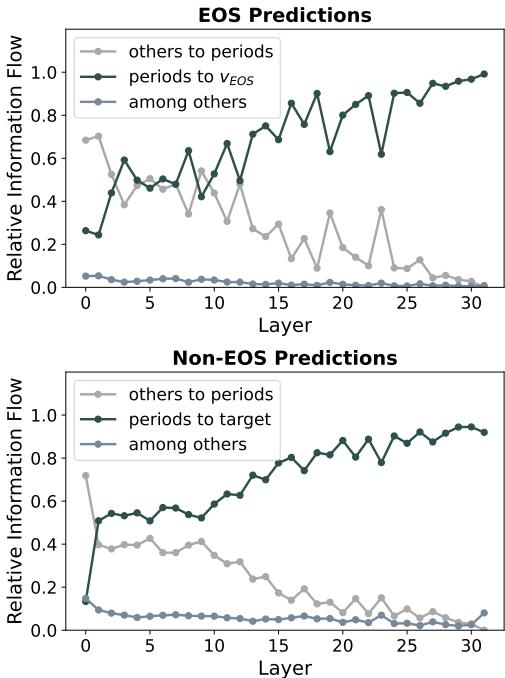


Figure 7: Relative significance of the information flow from regular content tokens (others) to periods, from periods to the target position for prediction, and among others. The significance is averaged over information flow targets and normalized across these three aspects for clearer comparison.

seemingly summarizing the entire sentence. At higher layers (near the output), this “summarized” information then aggregates to the target position for the next token prediction. Following Wang et al. (2023b), we illustrate these effects in Fig. 7, where such effects occur for both EOS and non-EOS predictions. This observation closely aligns with the findings by Wang et al. (2023b) in in-context learning (ICL), where the labels of in-context demonstrations act as “anchors” that aggregate information at lower layers and provide it for the final prediction at higher layers. This hierarchical information aggregation pattern elucidates how information moves within contexts and underpins our analysis in Sec-

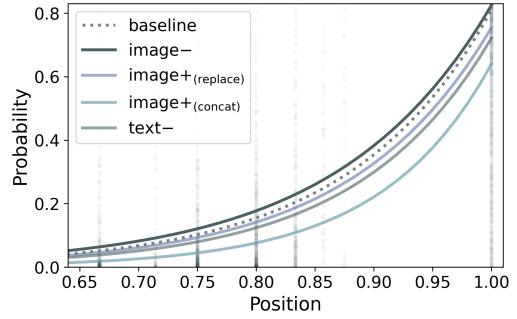


Figure 8: The predictive probability of the EOS token at different target positions within a sequence.

tion 2.1. We hope these observations can shed some light on future research.

B.2 Context Manipulation

In Section 2.2, we design three context manipulation methods to analyze how the model adjusts its EOS tendency according to these interventions. In addition to these methods, we also implement a variant of visual augmentation (*image+*), where we replace the input image with a random new one instead of concatenating a random image with the input image. This method can also decrease the relative completeness of the text, while not necessarily increasing the absolute information richness. The results in Fig. 8 demonstrate a similar impact from both variants, suggesting that the model does not merely compare the absolute semantic richness of the text and the image, but assesses the relative semantic completeness of the text to the image, i.e., whether the existing text encompasses the perceived visual information. This observation further supports our conjecture.

B.3 Selective EOS Supervision

EOS prediction tendency. In Fig. 9, we illustrate the EOS prediction tendency (average probability) of the LLaVA-1.5 (7b) model during further training on Detail23K. With Selective EOS Supervision

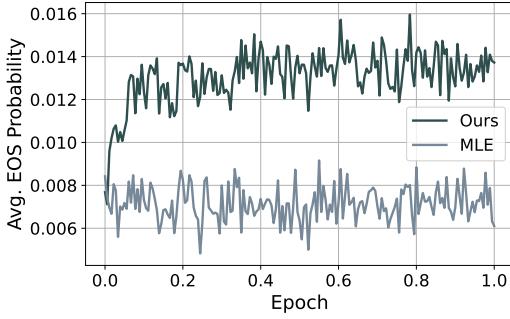


Figure 9: The average probability of the LLaVA-1.5 (7b) model predicting the EOS token at each position within the minibatch during further training.

Table 5: Hallucinated and correct objects “omitted” from the original model outputs by our methods.

Method	#Halluc.	#Correct	Halluc. Rate↑
Ours (w/ Inst.)	263	104	71.7%
Ours (w/ Cap.)	244	93	72.4%

Table 6: Average correct and hallucinated object counts of generated captions. **Original model:** LLaVA-1.5 (7b).

Model	#Correct↑	#Hallucinated↓
Original model	2.45	0.90
Ours (w/ Cap.)	2.40	0.63
Ours (w/ Inst.)	2.36	0.55

proposed in Section 3.1, the model’s tendency to predict EOS rises and stabilizes, while the model optimized by MLE shows no change in this behavior. This suggests that the proposed training objective effectively helps the model regain its capability to timely conclude sequences.

Dissecting omitted content. As our method reduces the generated content to alleviate hallucinations, it is interesting to investigate what is “omitted” by our method from the originally generated captions, specifically, how many “omitted” objects are correct and how many are hallucinations. We extract the generated objects from the outputs of both the original model and our further trained models, using the same technique as in the CHAIR evaluation. Then, we focus on these objects that are mentioned by the original model but not by our models, which are “omitted” from the original captions. As the results of the *Halluc. Rate* (hallucinated object rate of omission) in Table 5 shows, nearly 3/4 of the “omitted” objects are hallucinations, implying that such an omission is beneficial.

Furthermore, we analyze the average counts of correct and hallucinated objects in the model generation, as a supplement to the CHAIR metrics,

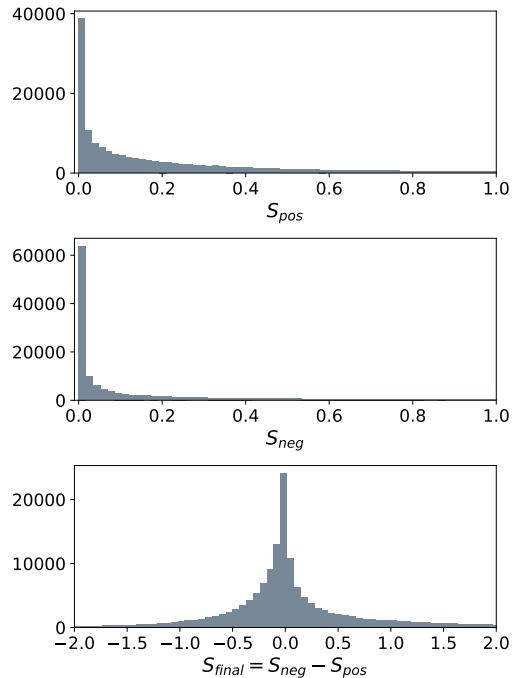


Figure 10: The score distributions of S_{pos} , S_{neg} , and S_{final} in the LLaVA-Instruction-150K dataset.

to demonstrate more comprehensively how our method impacts the quality of model generation. As shown in Table 6, our method reduces hallucinations while largely preserving the correct content.

B.4 Scoring EOS Supervision

Data Score Distributions. In Section 3.2, we discuss two metrics, S_{pos} and S_{neg} , which are summed over positions labeled as EOS and non-EOS, respectively. A natural concern is that the number of non-EOS positions far exceeds that of EOS positions, raising the question of whether the combined S_{final} might be dominated by S_{neg} . To clarify this, we examine the score distributions within the LLaVA-Instruction-150K dataset. As illustrated in Fig. 10, the value distributions of S_{pos} and S_{neg} are comparable in magnitude, and the S_{final} distribution is approximately normal with zero mean. This indicates a balance between S_{pos} and S_{neg} with neither metric dominating, and the top S_{final} scores necessitate both high S_{neg} and low S_{pos} . Thus, by maintaining a straightforward formulation of $S_{final} = S_{neg} - S_{pos}$ without introducing a balancing hyperparameter, the contributions of both metrics are reflected. Initial experiments also reveal that relying solely on S_{neg} for data filtering increases hallucinations, as it can lead to mistakenly removing data with high S_{pos} ; whereas using S_{pos} alone does reduce hallucina-

Table 7: MME and POPE evaluation results of baselines and models trained with our proposed two methods. For LLaVA-1.5 (7b), we compare the original model (**Baseline**), the model trained with **MLE**, and the one with Selective EOS Supervision (**Ours**). For LLaVA (7b), **Baseline** and **Ours** refer to the models trained with the original data and the data filtered by our Scoring EOS Supervision, respectively.

Model	Method	MME		POPE			
		Perception	Cognition	F1	Accuracy	Precision	Recall
LLaVA-1.5 (7b)	Baseline	1,516.1	348.2	85.9	86.9	94.0	79.1
	MLE	1,470.9	372.5	86.1	87.0	93.6	79.7
	Ours	1,490.4	367.9	86.0	86.8	93.5	79.5
LLaVA (7b)	Baseline	883.1	263.6	73.3	63.8	58.8	97.5
	Ours	910.9	260.0	71.2	59.7	56.0	98.1

tions, but is not as effective as S_{final} and will bring greater recall loss. Balancing both metrics yields the most desirable outcomes.

B.5 MME and POPE Evaluation

As mentioned in [Limitations](#), our proposed techniques focus on mitigating hallucinations in generative tasks by adjusting the models’ propensity for appropriately concluding outputs. However, these methods are not directly transferable to addressing hallucination problems in broader Visual Question Answering (VQA) tasks, such as those evaluated in the MME ([Fu et al., 2023](#)) and POPE ([Li et al., 2023b](#)) benchmarks. The MME benchmark assesses the model’s capabilities in terms of perception and cognition, whereas POPE concentrates on object hallucinations. Both benchmarks challenge models with *Yes-or-No* questions. As shown in Table 7, our methods do not yield performance gains on these benchmarks. The effectiveness of our approaches in generative tasks suggests that a model’s failure to timely stop generation is an

important hallucination source. However, addressing this issue alone does not fundamentally solve all hallucination problems as the origins of multimodal hallucinations are multifaceted. This area remains open for further investigation.

B.6 Qualitative Results

We present qualitative examples of our methods, Selective EOS Supervision in Fig. 11 and Scoring EOS Supervision in Fig. 12. The baseline models often produce hallucinations towards the end of their outputs, as they try to include too many details from the image, sometimes beyond their visual perception limits. This also explains why simply truncating sequences can reduce hallucinations. However, with our methods, the models better retain the innate ability to stop generation right after covering what they can visually perceive. This prevents the generation of overly lengthy, inaccurate, or irrelevant outputs that lower the overall quality and information density of the generated content, echoing the principle that “less is more.”

**Baseline:**

The image features a man sitting at a desk, working on a laptop computer. He is surrounded by numerous books, which are scattered around the desk and **on the floor**. Some books are placed close to the laptop, while others are located further away.

In addition to the laptop and books, there are a few other items on the desk, such as a mouse and a **cell phone**. A **clock can also be seen on the wall, indicating the time**. The scene suggests a focused and studious environment, with the man likely engaged in work or research.

Ours:

The image features a man sitting at a desk, working on a laptop computer. He is surrounded by numerous books, which are scattered across the desk and on the shelves behind him. The books vary in size and are placed in different orientations, creating a busy and studious atmosphere.

**Baseline:**

The image features a person wearing a wetsuit, standing in the water and holding a sailboard. The person is positioned in the middle of the scene, with the sailboard prominently visible in front of them. The water appears to be choppy, adding to the dynamic nature of the scene.

In the background, there are several boats of varying sizes scattered across the water, creating a lively atmosphere. Some of the boats are closer to the person with the sailboard, while others are further away, providing a sense of depth to the image.

Ours:

The image features a person wearing a wetsuit, standing in the water and holding a sail. The person appears to be preparing to windsurf or engage in a water sport that involves a sail. The sail is large and positioned in front of the person, with the water surrounding them.

**Baseline:**

The image captures a young man skillfully performing a trick on a skateboard, jumping into the air while riding it. He is wearing a plaid shirt and appears to be in a park-like setting. The skateboard is positioned underneath him, with the wheels visible as he jumps.

In the background, there are a few other people present, possibly watching the skateboarder or engaging in their own activities. The scene also features a bench, a chair, and a potted plant, adding to the park-like atmosphere.

Ours:

The image captures a young man skillfully performing a trick on his skateboard, jumping into the air with the skateboard beneath his feet. He is wearing a plaid shirt and appears to be in a park-like setting. The skateboarder is the main focus of the scene, with his impressive stunt drawing attention.

**Baseline:**

The image features a train station with a train on the tracks. A man is standing next to the train, holding a green trash can, likely preparing to clean the train. Another man is standing in the train, possibly a passenger or a worker. **There are several other people in the scene, some of them standing near the train, while others are scattered around the station. A handbag can be seen placed on the ground, possibly belonging to one of the people in the scene.**

Ours:

The image features a train station with a train on the tracks. A man in a blue shirt is standing next to the train, holding a green trash can. He appears to be in the process of unloading the trash can from the train. Another man is standing in the open door of the train, possibly waiting to get off or assisting with the unloading process.

**Baseline:**

The image features a wooden dining table with a **tray** holding three delicious donuts. The donuts are arranged in a visually appealing manner, with one donut placed in the center and the other two positioned on either side. The table appears to be in a restaurant setting, **as there is a bench situated nearby**.

In addition to the donuts, there is a cup placed on the table, possibly containing a beverage to accompany the sweet treats. The overall scene conveys a pleasant and inviting atmosphere, perfect for enjoying a snack or dessert.

Ours:

The image features a wooden dining table with three donuts placed on a napkin. The donuts are of different sizes and are arranged in a way that they are stacked on top of each other. The table is positioned near a window, allowing natural light to illuminate the scene.

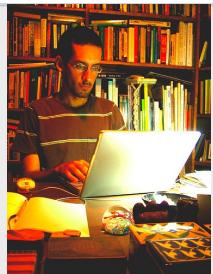
**Baseline:**

The image captures a man skillfully riding a wave on a surfboard in the ocean. He is wearing a wetsuit and appears to be enjoying the thrill of the sport. The surfer is positioned in the center of the scene, with the surfboard beneath him. **In the background, there are several cars parked along the shore, likely belonging to beachgoers or spectators. The cars are scattered around the area, with some closer to the water and others further back.** The scene is a lively representation of a day at the beach, **with people enjoying various activities.**

Ours:

The image captures a man in a black wetsuit skillfully riding a wave on a surfboard. He is positioned in the center of the scene, with the surfboard beneath him as he expertly navigates the water. The wave he is riding is quite large, providing an exciting challenge for the surfer.

Figure 11: Qualitative results of the LLaVA-1.5 (7b) model (**Baseline**) and its counterpart further trained on LLaVA-Instruction-150K with Selective EOS Supervision (**Ours**).

**Baseline:**

A man is sitting at a desk with a laptop computer in front of him. He is wearing glasses and appears to be working on the computer.

There are several books on the desk, **and a cup is placed on the desk as well. The man is sitting in a chair**, and the desk is surrounded by various items, including a lamp and a **clock**.

Ours:

In the image, a man is sitting at a desk with a laptop computer in front of him. He is wearing glasses and appears to be working on the computer. There are several books and papers on the desk, and a lamp is also present.

**Baseline:**

The image shows a snow-covered street with a one-way sign and a traffic light. The street is empty, and the snow is piled up on the sidewalk.

There are two cars parked on the street, one on the left side and the other on the right side. The cars are parked close to each other, and they are both facing the same direction.

Ours:

The image shows a snow-covered street with a one-way sign and a traffic light. The traffic light is green, indicating that it is safe to proceed.

**Baseline:**

The image shows a bathroom with a large mirror and a sink. The sink is made of white marble and has a faucet on it. The mirror is large and has a frame around it. The bathroom is clean and well-maintained, **with a towel rack and a toilet in the background**.

Ours:

The image shows a bathroom with a sink and a mirror. The sink is made of white marble and has a faucet on it. The mirror is large and has a frame around it.

**Baseline:**

A person is holding a sandwich in their hand, and they are eating it. The sandwich is a large, juicy steak sandwich with a bun.

There are two people in the image, one of them is eating the sandwich, and the other person is standing nearby. The sandwich is placed on a plate, and there are two bottles on the table.

Ours:

A man is holding a sandwich in his hand, and he is eating it. The sandwich is a large, juicy steak sandwich with a bun and meat.

**Baseline:**

In the image, a group of people are gathered around a large dining table, enjoying a meal together. There are several chairs around the table, and a few people are sitting in them. The table is set with various items, including cups, plates, and utensils.

There are also a few bottles on the table, and a vase is placed in the center of the table. The scene appears to be a social gathering or a business meeting, as people are engaged in conversation and enjoying each other's company.

Ours:

In the image, there is a large dining table with a variety of dishes and utensils, including forks, knives, and spoons. The table is set for a formal dinner, with a white tablecloth and a centerpiece. There are multiple people sitting around the table, enjoying their meal.

**Baseline:**

The image shows a bathroom with a white toilet **and a white sink**. The toilet is located on the left side of the bathroom, **and the sink is situated on the right side**.

Ours:

The image shows a white toilet with a white seat and a white tank. The toilet is located in a bathroom, and it has a window above it.

Figure 12: Qualitative results of the LLaVA (7b) model trained with original LLaVA-Instruction-150K data (**Baseline**) and with the data filtered by Scoring EOS Supervision (**Ours**).