Multi-Modal Hallucination Control by Visual Information Grounding

Alessandro Favero* Luca Zancato Matthew Trager Siddharth Choudhary
Pramuditha Perera Alessandro Achille Ashwin Swaminathan Stefano Soatto

AWS AI Labs

alessandro.favero@epfl.ch

{zancato,mttrager,sidchoud,pramudi,aachille,swashwin,soattos}@amazon.com

Abstract

Generative Vision-Language Models (VLMs) are prone to generate plausible-sounding textual answers that, however, are not always grounded in the input image. We investigate this phenomenon, usually referred to as "hallucination" and show that it stems from an excessive reliance on the language prior. In particular, we show that as more tokens are generated, the reliance on the visual prompt decreases, and this behavior strongly correlates with the emergence of hallucinations. To reduce hallucinations, we introduce Multi-Modal Mutual-Information Decoding (M3ID), a new sampling method for prompt amplification. M3ID amplifies the influence of the reference image over the language prior, hence favoring the generation of tokens with higher mutual information with the visual prompt. M3ID can be applied to any pre-trained autoregressive VLM at inference time without necessitating further training and with minimal computational overhead. If training is an option, we show that M3ID can be paired with Direct Preference Optimization (DPO) to improve the model's reliance on the prompt image without requiring any labels. Our empirical findings show that our algorithms maintain the fluency and linguistic capabilities of pre-trained VLMs while reducing hallucinations by mitigating visually ungrounded answers. Specifically, for the LLaVA 13B model, M3ID and M3ID+DPO reduce the percentage of hallucinated objects in captioning tasks by 25% and 28%, respectively, and improve the accuracy on VQA benchmarks such as POPE by 21% and 24%.

1. Introduction

Recent autoregressive Vision-Language Models (VLMs) have shown remarkable multimodal capabilities [4, 13]. However, VLMs, similarly to large language models (LLMs), are prone to "hallucinations" – generating plausible-sounding answers without factual basis, leading to potentially *ungrounded* or fabricated information [7]. Consequently, the community has been developing ever so complex and expensive alignment algorithms involving direct human supervision [17] and often brittle prompt engineering methods ("model begging") [28].

In this work, we propose to investigate hallucinations in VLMs through a quantifiable measure of visual *prompt dependency*. We assess whether a model output is ungrounded with respect to a visual prompt by comparing its likelihood with the likelihood of generating the same output

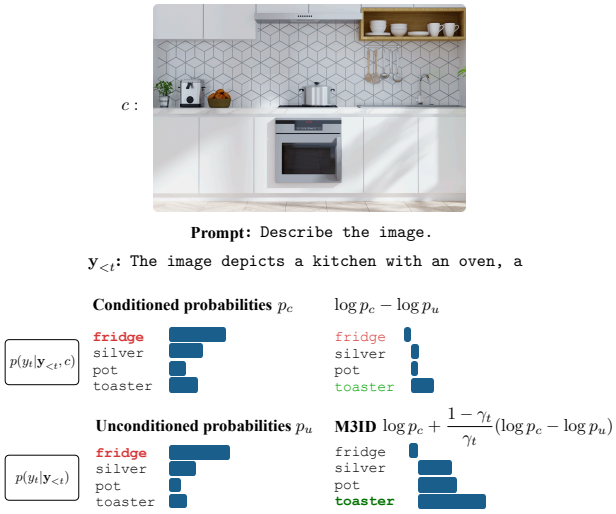

Figure 1. **Multi-Modal Mutual Information Decoding (M3ID).** Given a VLM p, an image c, and a text prompt, M3ID intervenes in the generative distribution by finding tokens that "surprise" the unconditioned VLM (i.e., the VLM without the image prompt). M3ID amplifies the conditioned directions that are not already predicted by the unconditioned model more as new tokens are generated by leveraging a progressively smaller γ_t. In the example, the VLM assigns a high likelihood to the hallucinated object *fridge*, over-relying on its unconditioned language prior. Instead, M3ID assigns a high likelihood to *toaster*, which is present in the image.

*Work done during an internship at AWS AI Labs.

without the visual information. In general, our measure of visual prompt dependency quantifies the extent to which a token is generic or specific to a visual signal. Importantly, low visual prompt dependency does not necessarily imply that a token is hallucinated, as it may occur in linguistically essential elements like conjunctions and prepositions. Yet, unlike the notion of hallucinations, which requires human judgment, our prompt dependency measure is always well-defined even without any ground truth annotations.

Our first contribution is to empirically demonstrate that the visual prompt dependency measure decreases as more tokens are generated. In Fig. 3, we show that, as more tokens are generated, the conditioning information gets diluted and "forgotten" or ignored by the model, possibly leading to more hallucinations. We refer to this effect as *conditioning dilution* or *fading memory effect*.

To counteract this, we propose an intervention on the generative distribution of a VLM that maximizes visual prompt dependency at inference time. We show that our method for *prompt amplification* maximizes the mutual information between the text output tokens and the visual prompt, effectively rescaling the image-conditioned component against the unconditioned distribution. We name our inference-time intervention on the generative distribution *Multi-Modal Mutual Information Decoding* (M3ID). M3ID is applicable to any off-the-shelf model without additional training or access to model weights, offering a low computational overhead alternative to standard decoding algorithms [13, 24]. Our results show that M3ID enhances the dependence on the visual prompt and reduces the number of hallucinations across various benchmarks while preserving the linguistic fluency of the original model.

Additionally, for users with access to model weights, we propose a training objective to further ground the model outputs on visual prompts. This approach involves recasting our prompt amplification objective as a preference optimization problem, using Direct Preference Optimization (DPO) [19] to align the pre-trained VLM. This method aims to increase the likelihood of continuations with a higher visual prompt dependency measure and allows to learn a better generation policy, further reducing the conditioning dilution effect and hallucinations.

In summary, our main contributions are as follows:

1. We propose a visual prompt dependency measure (PDM) to assess whether a model output is ungrounded with respect to the visual input. Empirically, we demonstrate that PDM decreases as more tokens are generated making the generations more likely to be hallucinated.
2. We introduce M3ID, a training-free intervention on the generative distribution of autoregressive VLMs which improves visual grounding and reduces hallucinations by amplifying the importance of the visual prompt over the language prior. In addition, we extend M3ID with DPO for multi-modal preference optimization to further improve visual grounding.
3. We show that applying M3ID or DPO reduces the percentage of hallucinated objects in *captioning tasks* [20] by 25% and by 28%, respectively and improves accuracy on the POPE [9] VQA hallucination benchmark by 21% and 24% over the base model.

2. Related work

Hallucinations in VLMs. A recent line of work introduced VLMs obtained by *grafting* visual encoders into pre-trained Large Language Models (LLMs) [4, 13, 32]. Forcing a pre-trained LLM to learn to "see" by reading vision tokens from a pre-trained vision backbone has proven successful and the resulting models show remarkable vision-language understanding capabilities on many multi-modal tasks [13, 23]. However, while inheriting strong linguistic capabilities and fluency from their base LLM, grafted VLMs also inherit the tendency to produce ungrounded or fabricated information [1, 6]. This is commonly referred to as "hallucination" in the recent machine learning literature [1, 6, 25]. In particular, several recent works empirically reported a sharp tendency of grafted VLMs to report objects not grounded on the visual information when probed with questions on the image content [1, 6, 9, 27]. The authors of [9] observed that this phenomenon is especially pronounced with objects that are either common or frequently appear together in the datasets used at training time. Furthermore, the authors of [11] suggest that VLMs often fail to correctly follow instructions involving absent objects and propose a new instruction-following training objective to enhance model alignment and robustness in the face of uncertainty. Parallel to our work, [31] delves into the factors contributing to object hallucinations in VLMs, exploring aspects like object co-occurrence, model uncertainty, and spatial position of hallucinations in the sentence. They propose a post-hoc algorithm to identify and correct hallucinations in VLM-generated content.

Context-dependent decodings. Decoding algorithms can be classified as either search or sampling algorithms [8]. Current search methods (e.g., greedy and beam search) can produce factually accurate generations but usually suffer from tedious and repetitive continuations. Sampling methods, on the other hand, (like nucleus [5], or typical decoding [15]) produce more diverse text, but suffer from topic drift. To solve these shortcomings, various decoding methods have been proposed to enhance reasoning and factual accuracy in generative autoregressive language models. Approaches such as context-aware decoding [21] and Pointwise Mutual Information (PMI) decoding [16, 25] focus on aligning the outputs of language models with the intended context or factual data. These strategies aim to reduce hal-

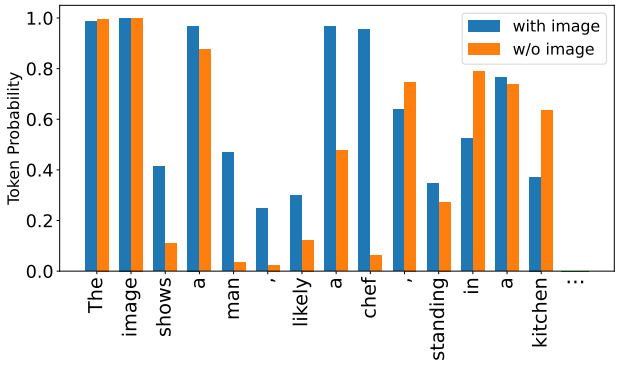

Figure 2. **How much does the conditioning prompt "surprise" the unconditioned predictor?** We report the likelihood that the conditioned and unconditioned models assign to each token in the string "The image ...". The probability gap increases on tokens for which visual information is required (e.g. objects and attributes), while it decreases for punctuation and articles. Note that both models mostly agree when sufficient visual information has already been extracted and is present in the caption (e.g. "kitchen" is very likely for the unconditioned model as well). See Sec. D.1 for a plot with our prompt dependency measure.

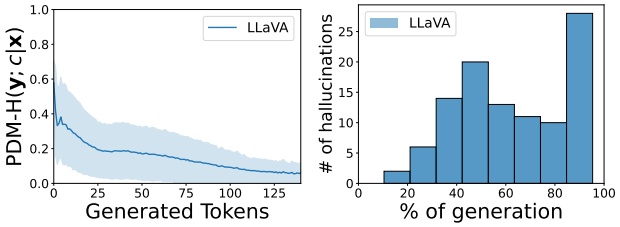

Figure 3. **The influence of the image conditioning decreases as we generate more tokens. (Left) Conditioning dilution.** We report the average of the prompt dependency measure (PDM-H) and its standard deviation for different synthetic captions generated by LLaVA on MS COCO's validation split. We see that the influence of the image over the next token prediction decreases as we generate more. This suggests that the information within VLM's visual prompt gets diluted and fades away as more tokens are generated. **(Right) Frequency of hallucinated objects as a function of the token position.** We report the number of non-existent objects present on the same synthetic captions as a function of the number of generated tokens. Note that very few objects are hallucinated for tokens near the visual prompt, while their number increases as more tokens are generated and with a smaller PDM.

lucinations in standard natural language processing tasks such as summarization. Mutual-information-based decoding techniques have proven to be helpful in steering the generation of LLMs to stay faithful to the given input text [25], or to promote diversity or relevance in neural dialogue models [7]. Our work is the first to use mutual information to improve multi-modal grounding and to vary the penalization of marginally likely outputs in a time-varying fashion to counteract the progressive forgetting of the visual prompt.

3. Analysis of hallucinations in VLMs

We start by investigating hallucinations in VLMs. We introduce a visual *prompt dependency measure* (PDM), which will later motivate our algorithms (Sec. 4). The main aim of this section is to assess whether a model output, such as a token, is hallucinated or not by comparing its likelihood with the likelihood of that same output being generated without the relevant conditioning factor, in this case, the image.

We consider decoding approaches for open-ended language generation of VLMs, where the model receives an input image and an input textual prompt and aims to generate a fluent and coherent continuation conditioned on both the image and the text. We denote probabilistic generative VLMs over a vocabulary of text tokens $\mathcal{V}$ with p and denote the probability of a token $y \in \mathcal{V}$ given a textual prompt $\mathbf{x}$ and an image context c as $p(y|\mathbf{x}, c)$. The goal of the VLM is to leverage the information contained in the input image c to provide a continuation $\mathbf{y} = [y_0, \dots, y_T]$ of the input prompt $\mathbf{x}$, e.g., "Describe this image in detail.". The probability of a valid sequence $\mathbf{y}$ can be computed as $p(\mathbf{y}|\mathbf{x}, c) = \prod_{t=1}^{T} p(y_t|\mathbf{y}_{<t}, \mathbf{x}, c)$, where $\mathbf{y}_{<t} \triangleq [y_0, \dots, y_{t-1}]$.

We propose to study hallucinations on VLMs using PDMs defined as follows

$$\text{PDM}(\mathbf{y}_{<t}; c|\mathbf{x}) \triangleq \text{dist}\Big(p(\cdot|\mathbf{y}_{<t}, \mathbf{x}, c),\ p(\cdot|\mathbf{y}_{<t}, \mathbf{x})\Big)$$

where dist is any distance measure between probability distributions, such as Hellinger, total variation, or KL. PDMs quantify how generic or context-specific a language model's output is. In particular, a high $\text{PDM}(\mathbf{y}_{<t}; c|\mathbf{x})$ indicates that the token y_t is strongly associated with a specific input prompt, while a low PDM suggests that the token is more prompt-neutral or prompt-agnostic. Depending on the choice of the distance function, PDMs highlight different aspects of the generative distribution. We will mainly use PDM-H based on the Hellinger distance defined as $H(p, q) = 2^{-1/2}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$, where $p = (p_i)_{i\in[k]}$ and $q = (q_i)_{i\in[k]}$ are discrete probability distributions. We refer the reader to the Supplementary Material for an analysis of the impact of other distances.

Contextual pressure. Fig. 2 shows the conditional and unconditional prediction likelihoods on a single caption. First, we note that tokens that are required for linguistic fluency, like prepositions and conjunctions, are highly predictable by both the conditioned and the unconditioned models. This pattern also emerges when the model generates tokens to compose "fine-grained" objects. For example, when the model generates "Peanut butter", both the VLM and the LLM can correctly predict the last token even without looking at the image and only having access to a sufficiently long truncation (e.g. "Peanut bu"). Therefore, in these circumstances, $p(y_t|\mathbf{y}_{<t}, \mathbf{x}, c)$ and $p(y_t|\mathbf{y}_{<t}, \mathbf{x})$ can be very close

irrespective of the number of generated tokens and without implying a higher hallucination risk. We will refer to this phenomenon as *contextual pressure*.

Conditioning dilution and hallucinations. In Fig. 3 we show that PDM-H decreases as more tokens are generated, indicating that the visual information gets diluted and neglected by the model throughout the generation process. This phenomenon suggests that the conditioned model distribution gets closer to the unconditioned one, the language prior, as we generate more tokens. In other words, the VLM places high probability mass on marginally likely tokens that are mostly explained by their language priors, $p(y_t|\mathbf{y}_{<t}, \mathbf{x}, c) \rightarrow p(y_t|\mathbf{y}_{<t}, \mathbf{x})$ as t increases. We refer to this phenomenon as *conditioning dilution* or *fading memory effect*. While conditioning dilution is not necessarily an issue (for instance, if the generated text has already extracted all of the relevant information from the context), one ideally would want to prevent the model from only relying on language priors based on the generated caption which may lack important details and facilitate hallucinations. Notice that whereas PDMs are useful for measuring the importance of the visual prompt, their values do not fully characterize the likelihood of generating ungrounded tokens. Therefore, in Fig. 3 we empirically test how much our prompt dependency measure correlates with the number of hallucinated objects generated by a SOTA VLM (LLaVA [13]) on the MS COCO dataset [10]. In the figure, we report the number of non-existent objects on captions generated by LLaVA as a function of the number of generated tokens. Note that very few objects are hallucinated near the input context, while their number increases as the PDM gets smaller. This motivates us to intervene in the generative distribution to maximize the PDM in order to reduce hallucinations.

4. Methods

Building on insights from Sec. 3, we formalize how to model the generation of VLMs as a fading memory process and how to intervene in the generation to prevent the VLM from "forgetting" the visual prompt c. In this section, we denote the log probabilities as $l(y_t|\mathbf{y}_{<t}, \mathbf{x}, c) \triangleq \log p(y_t|\mathbf{y}_{<t}, \mathbf{x}, c)$.

4.1. M3ID: Improving grounding at inference time

As we have shown in Sec. 3, within their prediction horizon, VLMs can be modeled as fading memory autoregressive systems [14, 30]: the more they generate the smaller the impact of the visual information on the generated tokens. As such, the longer the generation, the higher the reliance on the language priors which favor continuations not specifically grounded on the conditioning signal.

Preventing conditioning dilution. Under the fading memory assumption, we model the conditioned log probabilities $l(y_t|\mathbf{y}_{<t}, \mathbf{x}, c)$ (for example, LLaVA [13]) as an interpolation between the unconditioned model and a model l^* that does not "forget" old context as more tokens are generated.

$$l(y_t|\mathbf{y}_{<t}, \mathbf{x}, c) = \gamma_t \, l^*(y_t|\mathbf{y}_{<t}, \mathbf{x}, c) + (1 - \gamma_t) \, l(y_t|\mathbf{y}_{<t}, \mathbf{x}) \tag{1}$$

where $\gamma_t \in [0, 1]$ is a mixing coefficient which monotonically decreases over time. When γ_t is small, the conditional distribution is mostly explained without providing the input image, and the conditioned distribution "forgets" it, while for higher γ_t, the input image becomes more relevant. We take $\gamma_t \triangleq \exp(-\lambda t)$, λ models the rate of change of γ_t and defines the fading memory rate (*forgetting rate* [14, 30]).

Given observations from the conditioned distribution $l_c \triangleq l(y_t|\mathbf{y}_{<t}, \mathbf{x}, c)$ and the unconditioned one $l_u \triangleq l(y_t|\mathbf{y}_{<t}, \mathbf{x})$, our goal is to find an estimate $\hat{l}^*$ of the latent generative distribution l^* which does not forget the past.

To do so, we assume that l^* is a perturbation of the conditioned distribution, $l^* = l_c + \Delta$, where Δ can be assumed to be a random variable with zero mean and bounded variance. Therefore we rewrite our model in Eq. (1) as:

$$(1 - \gamma_t)(l_c - l_u) = \gamma_t \Delta. \tag{2}$$

Over the time index t, Eq. (2) is a stochastic process across tokens whose variance decreases over time. Hence, we can estimate the optimal intervention on l_c to counteract the fading memory effect and get closer to l^* as $\hat{l}^* = l_c + \hat{\Delta}$. We use measurements from Eq. (2) to estimate the correction term $\hat{\Delta}$. This brings us to the optimal intervention:

$$\hat{l}^* = l_c + \frac{1 - \gamma_t}{\gamma_t}(l_c - l_u). \tag{3}$$

Note that the optimal sampling distribution l^* is approximately proportional to l_c at the beginning of the generation (i.e., when γ_t is close to 1), which translates to sampling from the conditioned distribution alone. On the other hand, far from the input prompt (i.e., $\gamma_t \rightarrow 0$) the optimal sampling distribution becomes proportional to $l_c - l_u$. This amplifies tokens proposed by l_c that "surprise" the unconditional policy l_u and can be thought of as a way to ameliorate the conditioning dilution phenomenon described in Sec. 3. From an information theory viewpoint, the second scenario corresponds to maximizing the pairwise mutual information between the visual input and the text tokens instead of maximizing the log-likelihood of the text tokens alone. In fact, $\max_{\mathbf{y}} \log \frac{p(c,\mathbf{y}|\mathbf{x})}{p(c)p(\mathbf{y}|\mathbf{x})} = \max_{\mathbf{y}} \log \frac{p(\mathbf{y}|\mathbf{x},c)}{p(\mathbf{y}|\mathbf{x})} = \max_{\mathbf{y}} l_c - l_u$ [7, 25].

Accommodating for contextual pressure. Notice, however, that the contextual pressure forces l_c and l_u to be similar irrespective of the number of generated tokens. As such, penalizing the language prior l_u by amplifying $l_c - l_u$ indiscriminately would sometimes also penalize correct but

obvious tokens that do not require the input image to be inferred (like prepositions or conjunctions). To avoid such a scenario, we suppress our intervention in Eq. (3) when the contextual pressure is high. Specifically, if the conditioned model l_c is highly *confident* on the next token (i.e., it has the maximum probability above a threshold α [8]) we do not apply the correction term $l_c - l_u$.

Multi-Modal Mutual Information Decoding (M3ID). Putting everything together, our Multi-Modal Mutual Information Decoding Algorithm 1 for generation is given by:

$$y_t = \arg\max_{y \in \mathcal{V}} \hat{l}^*(y|\mathbf{y}_{<t}, \mathbf{x}, c) \tag{4}$$

where $\hat{l}^* = l_c + \mathbb{1}\Big[\max_k (l_c)_k < \log \alpha\Big] \frac{1-\gamma_t}{\gamma_t}(l_c - l_u)$. M3ID can be applied to different search algorithms like greedy search (as in Eq. (4)) or beam search.

4.2. M3ID+DPO to learn more grounded policies

With access to compute and model weights, we can optimize the model to output continuations that are more grounded to the image content. In this section, we reframe this goal into a *preference optimization problem*, where the objective is to prefer grounded continuations over ungrounded ones and fine-tune the VLM using this objective.

Multi-modal preference optimization. Consider two different continuations $\mathbf{y}_w$ and $\mathbf{y}_l$ of the same prompt $\mathbf{x}$ for the image c, with $\mathbf{y}_w$ being preferred over $\mathbf{y}_l$. Direct Preference Optimization (DPO) [19] is an alignment technique that aims at learning a generative policy that is more likely to generate continuations *similar* to $\mathbf{y}_w$ rather than $\mathbf{y}_l$. In particular, given a dataset $\mathcal{D}$ containing preference data pairs $(\mathbf{y}_w, \mathbf{y}_l)$, DPO minimizes the following loss function,

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(c,\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{p_\theta(\mathbf{y}_w|c,x)}{p_{\text{ref}}(\mathbf{y}_w|c,x)} - \beta \log \frac{p_\theta(\mathbf{y}_l|c,x)}{p_{\text{ref}}(\mathbf{y}_l|c,x)}\right)\right],$$

where p_{ref} denotes the VLM before DPO training and p_θ the model being trained. In short, the intuition behind this objective is to increase the likelihood of the completion $\mathbf{y}_w$ with respect to the base model p_{ref} while decreasing the likelihood of generating $\mathbf{y}_l$ relative to the base model.

Generating multi-modal preference data The success of DPO hinges on the quality of the pairs in the dataset $\mathcal{D}$. Therefore, we propose to fine-tune a pre-trained VLM using the DPO objective while ensuring that the preferred continuations are more grounded to the visual information. To do so, we generate preferred continuations by sampling from the pre-trained conditioned distribution improved with M3ID $\mathbf{y}_w \sim p^*(\mathbf{y}|\mathbf{x}, c)$, while we generate negative continuations by sampling from the unconditioned distribution $\mathbf{y}_l \sim p(\mathbf{y}|\mathbf{x})$. However, notice that generating text continuations from $\mathbf{x}$ (e.g., "Describe the image.") with the unconditioned model leads to negative continuations $\mathbf{y}_l$ that are often very different from the image caption and thus provide a little signal in the DPO objective. To overcome this limitation we append to $\mathbf{x}$ the first sentence generated by the conditioned VLM to restrict the set of possible continuations.

Algorithm 1: M3ID

Input: VLM p, textual prompt $\mathbf{x}$, image c, threshold on confidence α, forgetting rate λ
Output: Generated string $\mathbf{y}$ conditioned on $\mathbf{x}$ and c
$y_0 = \text{BOS}$, $t = 1$
while $y_t \neq \text{EOS}$ **do**
  $\gamma_t \leftarrow \exp(-\lambda t)$
  $l_c \leftarrow \log p(y|\mathbf{y}_{<t}, \mathbf{x}, c)$
  $l_u \leftarrow \log p(y|\mathbf{y}_{<t}, \mathbf{x})$
  $\hat{l}^* = l_c + \mathbb{1}\Big[\max_k (l_c)_k < \log \alpha\Big] \frac{1-\gamma_t}{\gamma_t}(l_c - l_u)$
  $y_t = \arg\max_{y \in \mathcal{V}} \hat{l}^*(y|\mathbf{y}_{<t}, \mathbf{x}, c)$
end

5. Experiments

We evaluate our methods on standard captioning and VQA benchmarks on MS COCO [10], an object detection and captioning dataset with annotations for 80 object categories. We use ground-truth annotations to measure the number of ungrounded objects that are predicted by our models.

Architecture. Our method applies to any VLM, as long as it is possible to drop the visual conditioning to compute the language-only prior. In the following, we shall mainly use the LLaVA architecture [13]. LLaVA is an open-source VLM connecting a frozen pre-trained open-set visual encoder with a pre-trained large language model via a trainable linear mapping layer. As a vision encoder, we use CLIP ViT L-14 [18]. As a large language model, we use the 7B and 13B versions of Vicuna v1.3 [3], which are instruction-tuned from LLaMA [24]. All our base models have been trained on LCS-558K [12] and LLaVA-Instruct-80K [13]. We follow the default query format used in LLaVA-v1.3 for the input data [13]. Furthermore, in Sec. D.3 we extend our evaluations to InstructBLIP [4], which introduces a Q-Former to bridge the vision and language modalities.

Baselines. For our baseline methods, we consider generation using standard multinomial sampling [13, 24] where the next token is randomly sampled from the prediction distribution with a given temperature parameter (0.2 for LLaVA [13]). Furthermore, we compare M3ID with other training-free methods for autoregressive models like PMI [25] and Contrastive Decoding [8]. PMI is developed for text summarization and improves the grounding of text gen-

Table 1. Evaluation of vision-language grounding on the validation set of MS COCO [10]. Captioning results are obtained by prompting the model with the task "Describe the image". *CHAIRi* and *CHAIRs* [20] denote the percentage of hallucinated objects and captions respectively, with lower values corresponding to fewer hallucinations. *Cover* indicates the percentage of annotated objects that are mentioned in the captions. We use LLaVA-v1.3.

	Captioning		
	CHAIRi ↓	CHAIRs ↓	Cover ↑
LLaVA_{7B}	8.1	17.5	53.3
LLaVA_{7B} PMI [25]	6.7	16.2	51.5
LLaVA_{7B} Contrastive [8]	6.3	14.8	**55.2**
LLaVA_{7B} M3ID	**5.9**	**13.8**	55.1
LLaVA_{13B}	7.4	18.5	**55.2**
LLaVA_{13B} PMI [25]	6.5	16.6	54.7
LLaVA_{13B} Contrastive [8]	6.5	14.5	54.7
LLaVA_{13B} M3ID	**5.5**	**13.2**	54.0
LLaVA_{13B} + LURE [31]	6.4	27.1	–
mPLUG-Owl + LURE [31]	5.4	18.8	–
LLaVA_{7B} M3ID + DPO	5.7	13.5	**55.8**
LLaVA_{13B} M3ID + DPO	**5.3**	**12.6**	54.2

eration by increasing the likelihood of generating tokens that are related to the text to be summarized. Differently from our method, the weight assigned to the realization of marginally likely continuations is time-independent. Similarly, Contrastive Decoding [7, 8] has been developed to foster generations of an *expert* LLM to contain plausible and fluent text with the textual input prompt by amplifying its prediction difference with respect to a weaker model. To ensure a fair comparison, we replace the text to be summarized with the input image for PMI and set the weaker model in Contrastive Decoding as the unconditioned model (see Sec. B for more details). We also compare with training-based methods that have been proposed to improve the grounding of VLMs on the visual prompt like LLaVA-RLHF [17], Robust mPLUG-Owl [11] and LURE [31]. However, all these methods require some form of annotations, like preference data [17] or positive/negative examples [11, 31], which M3ID+DPO does not require.

Evaluation. We evaluate our algorithm both on captioning and VQA benchmarks. On captioning, we measure object hallucinations by using the CHAIR (Captioning Hallucination Assessment with Image Relevance) metrics [20]. These metrics assess the quality of captions by comparing the mentioned objects to the annotated objects present in an image. In particular, CHAIR comprises two variants: one computes the fraction of hallucinated objects in the whole caption (CHAIRi) and the other evaluates the fraction of captions with at least one object hallucination (CHAIRs),

$$\text{CHAIRi} = \frac{\text{\# hallucinated objects}}{\text{\# generated objects}},$$

$$\text{CHAIRs} = \frac{\text{\# hallucinated captions}}{\text{\# generated captions}}.$$

Furthermore, we introduce the *Cover* metric to quantify the comprehensiveness of the captions by measuring the fraction of ground-truth annotated objects mentioned by the model,

$$\text{Cover} = \frac{|\{\text{correct mentioned objects}\}|}{|\{\text{annotated objects}\}|}.$$

To evaluate object perception, we use the POPE (Polling-based Object Probing Evaluation) VQA benchmark [9]. In particular, POPE recasts the evaluation of object hallucinations into a binary classification problem with yes/no questions of the type "Is a ⟨object⟩ present in the image?".

DPO training details. We use Hugging Face's DPO implementation in the TRL library [26] and train LLaVA for 5 epochs on 10,000 self-generated preference pairs with LoRA and cosine decaying learning rate with 2×10^{-5} peak. Following [19], we set $\beta = 0.1$ in the DPO loss.

5.1. VLM grounding on captioning

In Tab. 1 we test how much our method can prevent the generation of hallucinated objects in image captions. In particular, we compare against training-free and training-based baselines. We prompt all baseline methods with the instruction "Describe the image." and let the models generate until the EOS token is obtained.

First, we compare M3ID with other decoding strategies, PMI [25] and Contrastive Decoding [8]. The main difference between M3ID and these baselines is that M3ID increasingly counteracts the language prior as more tokens are generated. As such, M3ID reduces ungrounded generations compared to all other training-free baselines both on the large LLaVA13B and on the smaller LLaVA7B. Furthermore, it has uniformly good improvements for different model sizes over standard multinomial decoding, in particular, M3ID achieves 27%/21% relative improvement over LLaVA7B and 26%/29% over LLaVA13B on the CHAIRi and CHAIRs metrics. Importantly, this improvement does not come at the price of high reductions of the Cover metric, which actually improves on the 7B model and decreases by less than 2.2% for the larger 13B model[1]. Lastly, we note that M3ID shows an improvement in absolute performance that correlates with model size, suggesting that further gains could be obtained as larger models are used. In Sec. D.2, we also show that M3ID still improves grounding when longer captions (twice as long on average) are generated and where other decodings do not perform as well.

In Tab. 1 we also report training-based results and show that pairing DPO with M3ID leads to a smaller number

[1]Note that CHAIR metrics can be hacked by simply returning shorter captions that do not attempt to predict any object in the image.

Table 2. Evaluation on the POPE VQA hallucination benchmark [9]. $*$; † indicate results taken from [11] and [23] respectively. *SUP.* Indicates that the method involves supervised training on annotated data, in contrast to our approaches. POPE is divided into three subsets, *Random*, *Popular*, and *Adversarial*. We report the VQA accuracy of the model using the following template: "Is a ⟨object⟩ present in the image?", where we sample ⟨object⟩ randomly (Random), among the most frequent objects in the dataset (Popular), or among the objects that frequently co-occur with ⟨object⟩ (Adversarial). *Acc.* is the binary classification accuracy and *Yes* is the percentage of "Yes" answers.

	POPE							
	Random		Popular		Adversarial		All	
	Acc. ↑	Yes (%)	Acc. ↑	Yes (%)	Acc. ↑	Yes (%)	Acc. ↑	Yes (%)
Robust mPLUG-Owl-7B* [11] SUP.	**86.0**	–	73.0	–	65.0	–	74.7	–
LLaVA-RLHF$^\dagger_{7B}$ [23] SUP.	84.8	39.6	**83.3**	41.8	**80.7**	44.0	**82.9**	41.8
mPLUG-Owl-7B* [29]	52.0	–	57.0	–	60.0	–	67.3	–
LLaVA$_{7B}$	74.8	75.1	61.8	86.7	58.1	90.1	64.9	84.0
LLaVA$_{7B}$ M3ID	76.0	67.7	69.3	73.3	65.8	77.6	70.3	72.9
LLaVA$_{7B}$ M3ID + DPO	**81.2**	65.6	**73.9**	67.3	**68.2**	75.4	**74.4**	69.4
LLaVA-RLHF$^\dagger_{13B}$ [23] SUP.	**85.2**	38.4	**83.9**	38.0	**82.3**	40.5	**83.8**	39.0
MiniGPT4-13B* [32]	73.0	–	67.0	–	62.0	–	74.7	–
LLaVA$_{13B}$	67.9	80.6	63.8	83.2	59.8	87.3	63.8	83.7
LLaVA$_{13B}$ M3ID	84.3	55.6	77.0	61.6	71.3	68.2	77.5	61.8
LLaVA$_{13B}$ M3ID + DPO	**85.2**	53.4	**79.1**	57.5	**73.2**	67.5	**79.2**	51.1

of hallucinated objects and improved Cover numbers compared both to our training-free approach and LLaVA+LURE [31] a concurrent training based method to improve the grounding of VLMs that relies on GPT-3.5 annotations.

5.2. VLM grounding on VQA

Differently from the MS COCO captioning task, the POPE VQA benchmark only requires the generation of "Yes/No" tokens. As such, one does not expect the dilution effect to play a key role. However, we highlight that the image tokens and the output of the VLM are separated by the input template used to prime the model for the VQA task, which, as we show in Sec. D.4, introduces a non-negligible dilution effect.[2] To address this, we take an offset into account when using M3ID on POPE, and select $t = t_0$ where t_0 is the number of tokens in between the output and the image.

In Tab. 2, we report the results on the POPE VQA hallucination benchmark [9]. Hallucinations (wrong answers) tend to correlate with the tendency of the VLM to reply using the "Yes" token (see the disproportionately high percentage rate, 84%/83.7%, of "Yes" answers for the LLaVAv1.3 base model). This tendency can be counteracted at inference time and without any training by simply using M3ID. In our experiments, M3ID reduces the Yes ratio to 72.9%/61.8% for 7B and 13B models, respectively, which leads to relative accuracy improvements over standard LLaVA decoding by 8% and 21%, respectively.

To test whether our training-based approach increases the VLM's grounding on the visual prompt we trained a model with our DPO objective on the MS COCO captioning dataset and then tested it on the POPE benchmark. Note that improving caption generation does not directly imply improvements on the POPE VQA benchmark, in fact, even if the images are both from MS COCO, the output format is quite different: open-ended generation on the former and binary classification on the latter. However, intuitively, one expects that by improving the reliance on the visual prompt, the model could make better use of the visual information regardless of the output format required to solve the task at hand. Tab. 2 shows that M3ID+DPO further improves performance over M3ID's inference time intervention. Specifically, M3ID+DPO achieves 15% and 24% accuracy improvements over the LLaVA 7B and 13B models respectively. For completeness, we also compare with other training-based baselines that are fine-tuned on labeled [11] and preference data [17]. While M3ID+DPO does not have access to any labeled information we show that it is close to these baselines without requiring any annotations.

5.3. Ablations

Conditioning dilution and "overcompensation". In Fig. 4 and Tab. 3 we show that when the forgetting factor λ is high, corresponding to the assumption that the input image gets forgotten quickly, M3ID effectively maintains a large PDM-H throughout the generation. However, while deviating from the unconditioned probability is often a desired behavior, an excessive deviation from it across the whole generation could result in "overcompensation", which, as we show in Tab. 3 can be counterproductive and can lead to

[2] We use `[Img][Model Prompt][Question]`, see Sec. D.4.

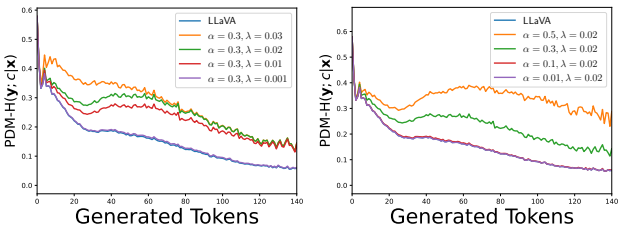

Figure 4. **Sensitivity analysis on hyper-parameters.** We vary both the forgetting factor λ and the thresholding parameter α. Small λ and α result in minor effects. Instead, either a high forgetting factor λ or a high α results in a stronger intervention which leads to higher PDM-H.

higher hallucination rates. In particular, observe that higher corrections detrimentally impact the cover metric. This suggests that while M3ID tries to find tokens that diverge from the language prior the most, it is more likely to fail in captioning elements that are inherently predictable by the language prior alone. For instance, in describing *an image of a dog on a leash accompanied by a man,* M3ID might overlook mentioning the presence of the man, a token that the language prior could anticipate without necessitating any visual information only from context clues. We report examples of this behavior in the Sec. D.2.

Importance of the confidence threshold. When we set a high threshold α, the indicator function remains active more frequently along the generation, leading M3ID to consistently prioritize maximizing the distance of the generation from the language prior. Hence, similarly to the scenario with a high λ, M3ID is likely to overcompensate and increase the rate of multi-modal hallucinations. However, too high α disrupts linguistic fluency. We report qualitative examples of this in Sec. D.2. In Tab. 4 we also show that adding the term to compensate for contextual pressure does not significantly change the results over LLaVA decoding in the POPE benchmark since it is essentially framed as a binary classification over the "Yes"/ "No" tokens and not open-ended generation like captioning.

6. Conclusions

We introduced M3ID, a new approach designed to combat multi-modal hallucinations by maximizing the mutual information between the text generated by VLMs and the corresponding visual context. M3ID operates at inference time and can be seamlessly integrated with any pre-trained autoregressive VLM. This makes M3ID a cost-effective and flexible solution to enhance vision-language grounding. Furthermore, for settings where model training is feasible and higher visual grounding is expected, we also paired M3ID with Direct Preference Optimization (DPO), showcasing reduced hallucinatory behaviors. Interestingly, our findings suggest that object hallucinations in VLMs result from excessive reliance on the language prior rather than a poor understanding of the visual modality. In fact, M3ID at inference time is sufficient to significantly reduce the amount of generated hallucinations without any training.

A limitation of M3ID is that it requires two forward passes at inference time, one for the conditioned and one for the unconditioned prediction. A possible solution to not increase inference time, but at the expense of higher memory consumption, is to use two batched queries with one having masked visual tokens. Furthermore, as we observed in Tab. 1, sometimes M3ID may prevent the generation of objects that are highly likely under the unprompted language prior (due to context clues). Interestingly, a similar observation has been reported in [22], where, people asked to provide a 10-word list of objects contained in a given image often failed to report the most obvious objects while mainly focusing on secondary ones. While we already showed that this can be mitigated with proper hyper-parameter selection Tab. 3 we believe that an interesting avenue for future research is to directly encode this preference within the model's weights by favoring the generation of structured captions that get progressively more detailed while still being grounded to the image.

In general, integrating human- or AI-annotated preference pairs, assessed based on their level of grounding, constitutes a promising avenue for future investigation.

Table 3. **Sensitivity to hyper-parameters.** λ is the forgetting factor used to counteract conditioning dilution while α is the threshold used to mitigate overcompensation when the contextual pressure is high. We report results on LLaVA-7B.

	Captioning		
	CHAIRi ↓	CHAIRs ↓	Cover ↑
LLaVA_{7B}	8.0	17.8	53.5
M3ID $\alpha = 0.3, \lambda = 0.02$	**5.8**	**13.6**	**55.0**
M3ID $\alpha = 0.3, \lambda = 0.001$	7.3	17.4	53.7
M3ID $\alpha = 0.3, \lambda = 0.03$	6.2	14.7	52.0
M3ID $\alpha = 0.3, \lambda = 0.1$	7.2	16.3	45.5
M3ID $\alpha = 0.5, \lambda = 0.02$	6.1	14.9	54.6
M3ID $\alpha = 0.1, \lambda = 0.02$	6.9	15.4	53.1
M3ID $\alpha = 0.01, \lambda = 0.02$	6.9	16.6	52.5

Table 4. **Ablation studies.** We ablate the components of M3ID on both model sizes. Removing the amplification term to counteract conditioning dilution leads to more hallucinations.

	POPE	Captioning		
	Accuracy ↑	CHAIRi ↓	CHAIRs ↓	Cover ↑
LLaVA_{7B}	64.9	8.0	17.8	53.5
w/ context pressure	65.5	6.9	16.7	54.5
w/ conditioning dilution	**77.5**	6.4	14.1	53.9
LLaVA_{7B} M3ID	**77.5**	**5.8**	**13.6**	**55.0**
LLaVA_{13B}	63.8	15.3	18.2	**55.3**
w/ context pressure	64.9	6.5	14.4	55.0
w/ conditioning dilution	**70.3**	5.9	14.8	53.7
LLaVA_{13B} M3ID	**70.3**	**5.4**	**13.0**	54.0

References

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 2

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 12

[3] 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. 5

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 5, 16

[5] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 2

[6] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021. 2

[7] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, 2016. Association for Computational Linguistics. 1, 3, 4, 6, 11

[8] 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, 2023. 2, 5, 6, 11, 14

[9] 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. 2, 6, 7, 11, 17

[10] 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. 4, 5, 6, 11, 14, 15

[11] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2, 6, 7

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 5

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 4, 5, 11

[14] M.B. Matthews and G.S. Moschytz. The identification of nonlinear discrete-time fading-memory systems using neural network models. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 41(11):740–751, 1994. 4

[15] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023. 2

[16] Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi, and Luis A Lastras. Pointwise mutual information based metric and decoding strategy for faithful generation in document grounded dialogs. *arXiv preprint arXiv:2305.12191*, 2023. 2

[17] 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. 1, 6, 7

[18] 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. 5

[19] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023. 2, 5, 6, 11, 12

[20] 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, Brussels, Belgium, 2018. Association for Computational Linguistics. 2, 6, 11, 14, 17

[21] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*, 2023. 2

[22] Merrielle Spain and Pietro Perona. Measuring and predicting object importance. *Int. J. Comput. Vis.*, 91(1):59–76, 2011. 8

[23] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2, 7

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

[25] Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive sum-

marization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 2, 3, 4, 5, 6, 11, 14

[26] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020. 6

[27] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023. 2

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 1

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

[30] Luca Zancato and Alessandro Chiuso. A novel deep neural network architecture for non-linear system identification. *IFAC-PapersOnLine*, 54(7):186–191, 2021. 19th IFAC Symposium on System Identification SYSID 2021. 4

[31] 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. 2, 6, 7

[32] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 7

Multi-Modal Hallucination Control by Visual Information Grounding

Supplementary Material

A. Experimental details

In this section, we provide details on our experimental setup.

Computing CHAIR metrics. We select 5,000 images from the 2014 validation split of MS COCO [10] and, for each image, prompt the models with one of the following questions: "Describe the image.", "Give an explanation of the image.", "Provide a description of the given image.". To measure hallucinated objects we follow [20] and complement the set of MS COCO annotated objects with their synonyms and automatically detect object hallucinations by comparing the objects mentioned in the captions against the annotated ones.

Computing POPE metrics. We use the official sets of questions introduced in [9], consisting of 3,000 *random*, 3,000 *popular*, and 3,000 *adversarial* questions on images taken from the 2014 validation split of MS COCO [10]. We report the VQA accuracy using the following template: "Is a ⟨object⟩ present in the image?", where we sample ⟨object⟩ randomly (Random), among the most frequent objects in the dataset (Popular), or among the objects that frequently co-occur with ⟨object⟩ (Adversarial).

Decoding hyper-parameters. To find the optimal hyper-parameters for M3ID, PMI, and Contrastive Decoding we compute the CHAIR metrics on 500 images sampled from the MS COCO validation set that do not overlap with the 5,000 images used to measure the CHAIR metrics. For M3ID, we search α in the range $\{0, 0.01, 0.1, 0.3, 0.5, 0.8, 1.0\}$ and the optimal fading memory coefficient λ in $\{0.001, 0.005, 0.01, 0.02, 0.03\}$. For PMI [25], we search τ in $\{0, 0.01, 0.1, 0.3, 0.5, 0.8, 1.0\}$ and μ in $\{0.1, 0.3, 0.5, 0.8, 1.5, 2.0\}$. For Contrastive Decoding [8], we search ξ in $\{0.1, 0.3, 0.5, 0.8, 1.5, 2.0\}$ and ψ in $\{0, 0.01, 0.1, 0.3, 0.5, 0.8, 1.0\}$.

DPO hyper-parameters. We use the same set of hyper-parameters as the ones used to fine-tune LLaVA with SFT [13]. In particular, we train using the AdamW optimizer with batch size 128, learning rate 2×10^{-5}, cosine annealing scheduler, warm-up ratio 0.03, and DeepSpeed ZeRO-2. We use LoRA with rank 64, scaling factor 16, and dropout 0.05. Following [19], we set $\beta = 0.1$ in the DPO loss.

Hardware. Experiments are run on 8 NVIDIA Tesla A100 GPUs.

B. Context-dependent decodings

In this section, we review precedent context-dependent decoding strategies developed for large language models that we use as baselines in our experiments and highlight the differences with M3ID.

PMI Decoding. Pointwise Mutual Information (PMI) Decoding [25] is a decoding strategy developed for improving the grounding of summary generation by increasing the likelihood of generating tokens that are related to the text to be summarized. Specifically, PMI optimizes for the mutual information of the source text and target token when the model exhibits uncertainty – quantified with the Shannon entropy of the output distribution. Let p denote an LLM, $\mathbf{x}$ a prompt and c the source text to be summarized, PMI selects tokens as follows,

$$y_t = \arg\max_{y \in \mathcal{V}} \log p(y|\mathbf{y}_{<t}, \mathbf{x}, c) - \mu \mathbb{1}\left[H(p(\cdot|\mathbf{y}_{<t}, \mathbf{x}, c)) \geq \tau\right] \log p(y|\mathbf{y}_{<t}, \mathbf{x})$$

where $H(p(\cdot|\mathbf{y}_{<t}, \mathbf{x}, c)) = -\sum_{y \in \mathcal{V}} p(y|\mathbf{y}_{<t}, \mathbf{x}, c) \log p(y|\mathbf{y}_{<t}, \mathbf{x}, c)$ denotes the Shannon entropy. In the multi-modal case, we replace the text to be summarized with the input image. Notice that, differently from our method, in the PMI intervention the weight μ assigned to the realization of marginally likely continuations is constant across time.

Contrastive Decoding. Contrastive Decoding [7, 8] has been developed to foster generations of an *expert* LLM p_{exp} to contain plausible and fluent text with the textual input prompt by amplifying its prediction difference with respect to a weaker *amateur* model p_{ama}. Firstly, Contrastive Decoding applies a *plausibility constraint* in order to mask tokens to which the expert model assigns low probability, i.e.,

$$\mathcal{V}_{\text{plausible}} = \left\{ v \in \mathcal{V},\ (\log p_{\text{exp}})_v \geq \log \psi + \max_{k \in \mathcal{V}} (\log p_{\text{exp}})_k \right\}.$$

Secondly, Contrastive Decoding applies a penalty to the amateur logits,

$$y_t = \arg \max_{y \in \mathcal{V}_{\text{plausible}}} (1+\xi) \log p_{\text{exp}}(y|\mathbf{y}_{<t}) - \xi \log p_{\text{ama}}(y|\mathbf{y}_{<t}).$$

In the multi-modal case, we set the expert model as the VLM, i.e., $p_{\text{exp}}(\cdot) = p(\cdot|\mathbf{x}, c)$ and the amateur model as the unconditioned model, i.e., $p_{\text{ama}}(\cdot) = p(\cdot|\mathbf{x})$. Notice that, also in this case, the penalization is time-independent.

Multi-Modal Mutual-Information Decoding (M3ID). Our decoding algorithm M3ID optimizes the vision-language grounding of VLM generations by maximizing the mutual information between the generated textual tokens $\mathbf{y}$ and the provided visual context c. Specifically, when the contextual pressure is low (see Sec. 4), M3ID applies a penalty to the log probabilities unconditioned on the image, i.e., it penalizes the language-only prior of the VLM. In contrast with the previous strategies and motivated by our fading-memory modeling assumption, in M3ID the strength of this penalization increases as more tokens are generated and it is controlled by the time-dependent parameter γ_t. Let p denote the VLM and $\mathbf{x}$ a textual prompt, M3ID selects new tokens as follows,

$$\begin{cases} y_t = \arg\max_{y \in \mathcal{V}} \log p(y|\mathbf{y}_{<t}, \mathbf{x}, c) - \mathbb{1}\left[\max_k (p(\cdot|\mathbf{y}_{<t}, \mathbf{x}, c))_k < \alpha\right] \dfrac{1-\gamma_t}{\gamma_t} \left(\log p(y|\mathbf{y}_{<t}, \mathbf{x}, c) - \log p(y|\mathbf{y}_{<t}, \mathbf{x})\right) \\ \gamma_t = e^{-\lambda(t+t_0)}, \end{cases}$$

where the parameter λ controls the decay rate of the fading memory and t_0 controls the strength of the language-prior penalization at the beginning of the generation. In all our captioning experiments, we set $t_0 = 0$, whereas when evaluating on POPE we find that using $t_0 = 10$, approximately equal to the length of the question $\mathbf{x}$ which separates the image c and the beginning of the answer $\mathbf{y}$, improves the results (see also Sec. D.4).

C. Multi-modal Direct Preference Optimization

In this section, we present the theoretical formulation of multi-modal Direct Preference Optimization (DPO) and our algorithm for aligning VLMs on self-generated data.

Theoretical formulation. Given an image context c, let $\mathbf{y}_w$ and $\mathbf{y}_l$ be two different continuations to the same prompt $\mathbf{x}$, with $\mathbf{y}_w$ preferred over $\mathbf{y}_l$ ($\mathbf{y}_w \succcurlyeq \mathbf{y}_l$) judging based on grounding with respect to c. Consider the common assumption that the preference is governed by a latent Bradley-Terry preference model [2] with the reward given by $r^*(c, \mathbf{x}, \mathbf{y})$ and higher reward corresponding to better vision-language grounding. Thus, the preference distribution can be written as a sigmoid of the rewards' difference,

$$p^*(\mathbf{y}_w \succcurlyeq \mathbf{y}_l|c, \mathbf{x}) = \sigma\left(r^*(c, \mathbf{x}, \mathbf{y}_w) - r^*(c, \mathbf{x}, \mathbf{y}_l)\right).$$

The preference optimization objective is to find the optimal policy $\hat{p}^*$ that maximizes the expected reward of the generated continuations and minimizes the KL divergence with the reference policy p_{ref}, which is the policy at initialization[3],

$$\hat{p}^* = \arg\max_p \mathbb{E}_p\, r^*(c, \mathbf{x}, \mathbf{y}) - \beta D_{\text{KL}}(p, p_{\text{ref}}),$$

where $\mathbf{y} \sim p(\,\cdot\,|c, \mathbf{x})$. Leveraging an analytical mapping between r^* and p^*, Direct Preference Optimization (DPO) [19] allows to directly optimize the policy on a preference dataset $\mathcal{D} = \{(c^i, \mathbf{x}^i, \mathbf{y}_w^i, \mathbf{y}_l^i)\}_i$ using the loss

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(c,\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta \log \frac{p(\mathbf{y}_w|c,\mathbf{x})}{p_{\text{ref}}(\mathbf{y}_w|c,\mathbf{x})} - \beta \log \frac{p(\mathbf{y}_l|c,\mathbf{x})}{p_{\text{ref}}(\mathbf{y}_l|c,\mathbf{x})}\right)\right].$$

Optimizing for this loss increases the likelihood of the preferred and better-grounded completions $\mathbf{y}_w$ and decreases the likelihood of the poorly-grounded completions $\mathbf{y}_l$.

Preference data generation and alignment. Our alignment procedure is as follows.

1. We start from n images $\{c^i\}_{i\in[n]}$ and obtain one positive caption $\mathbf{y}_w^i$ per image prompting a VLM with the task $\mathbf{x}^i =$ "Describe this image." In order to maximize vision-language grounding, we use M3ID to generate the answers.
2. To generate the negative captions $\{\mathbf{y}_l^i\}_{i\in[n]}$, we use the same set of images and prompt an unconditioned VLM (i.e., the VLM with masked visual tokens) to complete the first sentences generated for the positive captions. In such a way, the first sentence is well-grounded in the image content, while the continuations are dictated by the language prior of the VLM alone and serve as informative negative examples.

[3] This regularization term is often added to avoid reward hacking.

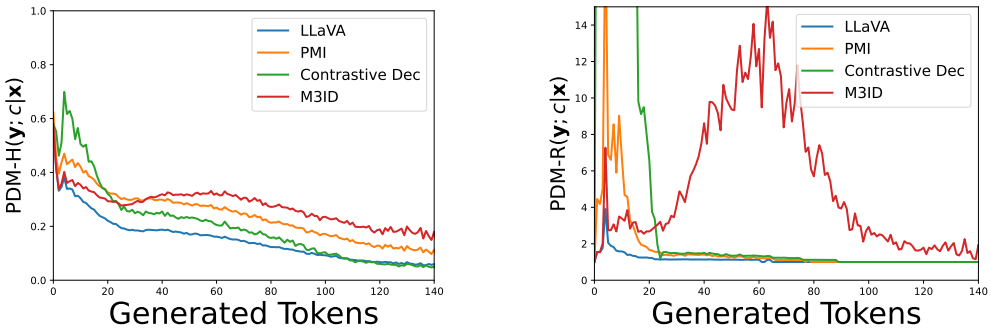

Figure 5. **PDMs with different distances.** We report how PDM varies with different choices of the distance function, i.e., PDM-H with Hellinger (left panel) and PDM-R with Rank (right panel). We compare the generation of a caption with M3ID, standard, and other context-aware decoding schemes. With both distance functions, M3ID maximizes the separation of the conditioned and unconditioned distribution, effectively counteracting the language prior even throughout long caption generations.

3. We train the VLM on the self-generated preference dataset $\mathcal{D} = \{(c^i, \mathbf{x}^i, \mathbf{y}_w^i, \mathbf{y}_l^i)\}_i$ with the DPO loss and the hyper-parameters described in Sec. A.

D. Further results

In this section, we present further results complementing the experiments presented in the main text.

D.1. Prompt dependency measures

Impact of different distance functions. Depending on the choice of the distance function, PDMs highlight different aspects of the generative distribution. For example, by choosing the Hellinger (PDM-H) distance, we consider the whole generative distribution over the vocabulary tokens. However, notice that the generative process is mainly determined by its high probability modes, especially when greedy decoding or low-temperature sampling is used. Hence PDM-H relies on a high number of irrelevant tokens. To account for this, we complement PDM-H with the following PDM which only depends on the model's preference in generating the most likely token:

$$\text{PDM-R}(\mathbf{y}, c|\mathbf{x}) \triangleq \text{rank}_{p(\mathbf{y}|\mathbf{x})}(\arg\max \text{rank}_{p(\mathbf{y}|\mathbf{x},c)}) \tag{5}$$

where rank_p is the ranking of the tokens in the vocabulary according to the distribution p. Note that when PDM-R $= 1$, the highest-ranking token with the conditioning image is also the highest-ranking token without the image. So a greedy

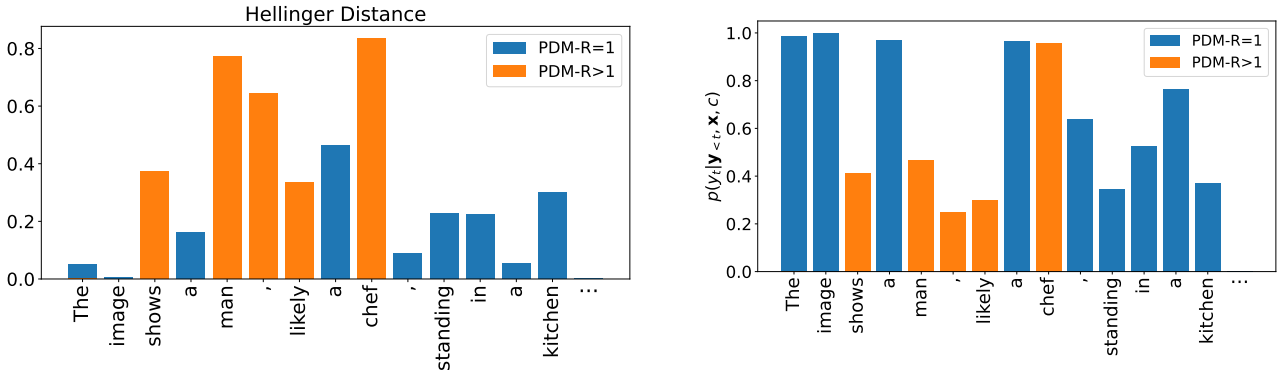

Figure 6. **How much does the conditioning prompt "surprise" the unconditioned predictor?** We report the PDM-H and the PDM-R for each token in the string "The image shows ...". Both distances increase on tokens for which visual information is required and decrease for punctuation, articles, and when sufficient visual information has already been extracted and is present in the caption (e.g., "kitchen" is already very likely given the text alone).

Table 5. Evaluation of vision-language grounding on the validation set of MS COCO [10]. Captioning results have been obtained by prompting the model with the task "Describe the image", and detailed captioning results with the task "Describe the image in detail." *CHAIRi* and *CHAIRs* [20] denote the percentage of hallucinated objects and captions respectively, with lower values corresponding to fewer hallucinations. *Cover* indicates the percentage of annotated objects that are mentioned in the captions.

	Captioning			Detailed Captioning		
	CHAIRi ↓	CHAIRs ↓	Cover ↑	CHAIRi ↓	CHAIRs ↓	Cover ↑
LLaVA_{7B}	8.1 ± 0.5	17.5 ± 0.8	53.3 ± 0.7	10.1 ± 0.3	35.6 ± 0.7	**71.1** ± 0.5
LLaVA_{7B} PMI [25]	6.7 ± 0.7	16.2 ± 1.2	51.5 ± 0.6	9.8 ± 0.4	34.0 ± 0.9	70.1 ± 1.1
LLaVA_{7B} Contrastive [8]	6.3 ± 0.1	14.8 ± 0.1	55.2 ± 0.1	10.1 ± 0.1	33.7 ± 0.3	69.9 ± 0.1
LLaVA_{7B} M3ID	5.9 ± 0.4	13.8 ± 0.8	55.1 ± 0.4	8.8 ± 0.4	28.8 ± 0.8	67.4 ± 0.7
LLaVA_{7B} M3ID + DPO	**5.7** ± 0.4	**13.5** ± 0.7	**55.8** ± 0.2	**8.5** ± 0.6	**28.1** ± 0.6	68.9 ± 0.9
LLaVA_{13B}	7.4 ± 0.4	18.5 ± 1.0	**55.2** ± 0.4	8.5 ± 0.5	31.8 ± 1.7	67.0 ± 0.7
LLaVA_{13B} PMI [25]	6.5 ± 0.5	16.6 ± 1.3	54.7 ± 0.6	8.6 ± 0.3	28.5 ± 1.5	67.1 ± 0.7
LLaVA_{13B} Contrastive [8]	6.5 ± 0.1	14.5 ± 0.2	54.7 ± 0.1	8.1 ± 0.2	28.7 ± 0.4	**68.2** ± 0.2
LLaVA_{13B} M3ID	5.5 ± 0.3	13.2 ± 0.8	54.0 ± 0.4	7.8 ± 0.6	27.5 ± 0.8	**68.2** ± 0.6
LLaVA_{13B} M3ID + DPO	**5.3** ± 0.2	**12.6** ± 0.7	54.2 ± 0.3	**7.5** ± 0.4	**26.9** ± 0.7	67.7 ± 0.5

generation without the specific conditioning signal would result in the same continuation.

In Fig. 5, we show how the average PDM-H and PDM-R vary during the generation of 5000 captions using LLaVA's standard decoding, M3ID, and the context-dependent decoding baselines. PDMs with standard decoding decrease as more tokens are generated in both cases. In particular, we observe that after approximately 30 tokens have been generated, PDM-R approaches 1, signaling that, on average, the VLM effectively behaves as an LLM which is unconditioned on the visual prompt. PMI and Contrastive Decoding are effective in increasing the PDM distance at the beginning of the captions but their effect becomes negligible near the middle and end of the generation, when more multi-modal hallucinations are observed. In contrast, M3ID maximizes both PDMs and successfully counteracts the language prior even when generating longer captions. Notice that PDM-R has a spike near the center of the generated captions signaling that M3ID selects, on average, tokens that are ranked among the top 15 by the model without conditioning. We attribute this phenomenon to the fact that there exist several ways to continue the captions after the beginning and before the end of the captions, which are more "constrained" (e.g., the majority of captions start as "In the image there are ...").

Decay rate of conditioning dilution. In Fig. 3 and Fig. 5 we report PDMs for the LLaVA model. Note that to reduce the computational complexity of the hyper-parameter search of our method it is possible to estimate the decay rate λ directly from these plots. In fact, estimating the decay rate from Fig. 3 with linear regression in logarithmic coordinates, leads to $\lambda = 0.016$, which is close to the optimal value 0.02 found with cross-validation.

Surprising the unconditioned model. Fig. 6 shows the PDM-H and the PDM-R for each token in the string "The image shows a man, likely a chef, standing in a kitchen ...". We color code the bars according to PDM-R, in blue when the conditioned and unconditioned models share the same prediction and in orange when they differ. Note that both distances increase on tokens for which visual information is required (e.g. objects and attributes), while they decrease for articles. Also, note that both models mostly agree even when sufficient visual information has already been extracted and is present in the caption (e.g., "kitchen" is already very likely given the text alone). On the right of Fig. 6, we show that tokens like "the" and "a" have high probability according to the VLM and have PDM-R $= 1$.

D.2. Captioning

Detailed captioning and stochasticity due to sampling. Tab. 5 presents the CHAIR and cover metrics for detailed captioning, where LLaVA is prompted to generate longer captions resulting in an increase of the cover metric from approximately 55% to about 70%. In this scenario, although all methods exhibit larger CHAIR values compared to standard captioning tasks, M3ID and M3ID+DPO achieve the best results in minimizing object hallucinations, substantiating the effectiveness of our approaches. Additionally, this table includes the standard deviations resulting from stochastic sampling, which we omitted in the main text for the sake of clarity.

Introduced errors. To further illustrate the effectiveness of M3ID in reducing hallucinations, in Tab. 6 we reconsider the

Table 6. **M3ID introduced error metrics.** Frequency at which M3ID either modifies or maintains the hallucinations that LLaVA produces in the task of captioning MS COCO images [10] (same setting as in Tab. 1).

LLaVA M3ID	correct correct	correct incorrect	incorrect correct	incorrect incorrect
LLaVA_{7B}	72.0 %	3.0 %	16.2 %	9.8 %
LLaVA_{13B}	74.2 %	2.4 %	18 %	5.4 %

Table 7. **Examples of overcompensation.** Maximally surprising the VLM might lead to "overcompensation", a phenomenon affecting decoding interventions that amplify the logits difference between the conditioned and unconditioned models like M3ID. When M3ID's hyper-parameters force a strong correction over the language prior, the resulting captions tend to overlook elements that are inherently predictable by the language prior based on the text tokens alone, e.g., the man in the left picture and the van in the right picture.

Input Image		
Input Instruction	*Describe the image.*	
LLaVA_{7B}	The image shows a man walking a dog on a leash, with the dog wearing a raincoat. They are walking down a sidewalk, and the man is holding an umbrella to protect himself from the rain.	The image shows a blue van parked on the side of a street, with a fire hydrant nearby. The van is parked on the side of the road, and there is a fire hydrant located close to it.
LLaVA_{7B} overcompensation	The image features a dog on a leash, walking down a paved road or pathway. [*No man is mentioned*]	The image features a city street with a sidewalk, a fire hydrant, and a pole. There is also a street sign visible in the scene. [*No van is mentioned*]
LLaVA_{7B} M3ID	The image shows a man walking a dog on a leash in a city street. The man is holding an umbrella over the dog to protect it from rain. They are walking along a dirt road near mountains.	The image features a blue van parked on the side of a city street next to a fire hydrant and on the curb.

metrics presented in Tab. 1 before aggregration. Specifically, we detail the proportion of hallucination-free captions generated by both the base model and M3ID, alongside the proportion of captions that contain hallucinations exclusively in one of the two cases, and the proportion of captions where hallucinations occur under both models. For the 13B model (respectively 7B model), M3ID corrects hallucinations in 18% (16.2%) of instances, while introducing new hallucinations in only 2.4% (3.0%) of cases.

Overcompensation. Tab. 7 illustrates instances of "overcompensation", a phenomenon affecting decoding interventions that amplify the logits difference between the conditioned and unconditioned models. For example, M3ID, while trying to select tokens that deviate from the language prior, tends to overlook elements that are inherently predictable by the language prior alone (see Sec. 5.3). In describing the image of a dog on a leash with a man (left panel in Tab. 7), when M3ID is applied with an excessively large penalty applied to the language prior, it fails to mention the presence of the man. This oversight occurs because the "man" token could be inferred from context clues (first tokens in the caption) by the language prior alone without necessitating any visual information. A similar behavior is observed in the second example (right panel), where M3ID fails

Table 8. **Examples of poor language fluency.** Using a high threshold value α ($\alpha = 1$ in the examples) disrupts language fluency. Intensifying the strength of the correction term results in the model initially losing syntactical accuracy and subsequently generating apparently random tokens, including non-English ones.

Input Image		
Input Instruction	*Describe the image.*	
LLaVA_{7B}	The image shows a hot dog on a bun, accompanied by a cup of soda, sitting on a red table.	The image shows a blue van parked on the side of a street, with a fire hydrant nearby. The van is parked on the side of the road, and there is a fire hydrant located close to it.
LLaVA_{7B} mild overcompensation	The image features two footlong foot hotdogs accompanied by napkins on red surface such as disposable sheets laid over bright stried yards Kontrola Conference Red pl attached signs. [...]	The image shows vehicles parked on opposite sides of the street street. Specifically, attention is focused onungsseiteways van stopped at an empty firearchivi coming halfViewByIdium curbALSEwer. [...]
LLaVA_{7B} strong overcompensation	The image features a footlong hot dog sitting wrapped in paper on red surface [cyrillic tokens] table or bench[ˆ], next to [cyrillic tokens] genomsnitt queen empty paper.](phia cup company genomsnitt.) [...]	The image shows a van parked on the side of a city street next to fire meters and standing curbs. Street signs Kontrola Bez Praza (”stop required prohibitively during Kontrola Bez Bez Praza Activated Here Only Stopton Blue Vanlet Transport Motor [mixed cyrillic tokens until the EOS]
LLaVA_{7B} M3ID	The image shows a hot dog and a drink sitting on a red table or counter.	The image features a blue van parked on the side of a city street next to a fire hydrant and on the curb.

to mention the presence of the van, despite it being the dominant element in the image.

Preserving language structure. Tab. 8 demonstrates the detrimental effects of using a high threshold value α. In such cases, the indicator function activates more frequently throughout the generation, leading M3ID to consistently prioritize maximizing the distance from the language prior and thereby disrupting language fluency and structure. Specifically, setting $\alpha = 1$ and incrementally intensifying the strength of the correction term leads to “overcompensation”. This results in the model initially losing syntactical accuracy and subsequently generating apparently random tokens, including tokens in non-English languages, such as Cyrillic.

D.3. InstructBLIP

Our method applies to any VLM, regardless of whether the base LLM has been fine-tuned or not, so long as it is possible to drop the visual conditioning to compute the language prior. In this section, we apply M3ID to the InstructBLIP model [4] and evaluate it on the CHAIR benchmark. Compared to the LLaVA architecture, InstructBLIP connects the vision encoder and the LLM using a Q-Former. As such, differently from LLaVA, to obtain the unconditioned model predictions we mask the image tokens in the Q-Former. As reported in Tab. 9, also for this architecture, M3ID significantly reduces hallucinations on CHAIR with respect to standard generation, showcasing its broad applicability.

Table 9. **InstructBLIP + M3ID.** Evaluation of vision-language grounding on MS COCO (as in Tab. 1) using InstructBLIP and masking the image tokens in the Q-Former for the unconditioned log probabilities. Captioning results are obtained by prompting the model with the task "Describe the image.". *CHAIRi* and *CHAIRs* [20] denote the percentage of hallucinated objects and captions respectively, with lower values corresponding to fewer hallucinations. *Cover* indicates the percentage of annotated objects that are mentioned in the captions.

	CHAIRi ↓	CHAIRs ↓	Cover ↑
Instruct BLIP $_{7B}$	7.9	21.2	**59.2**
InstructBLIP$_{7B}$ M3ID	**6.6**	**17.6**	56.3
InstructBLIP$_{13B}$	7.3	18.6	**54.1**
InstructBLIP$_{13B}$ M3ID	**6.4**	**14.0**	53.6

Table 10. **Conditioning dilution in VQA.** Evaluation on the POPE VQA hallucination benchmark [9] (as in Tab. 2) using templates of different lengths: *short* prompts the models with "Is a ⟨object⟩ present in the image?", *long* in addition specifies the format of the answer and details on the evaluation. *Acc.* denotes the binary classification accuracy. Longer prompts result in higher errors due to the conditioning dilution phenomenon, which M3ID effectively reduces by introducing an offset t_0 that takes into account the length of the template.

	POPE All	
	Short prompt template ($t_0 = 10$) ↑	Long prompt template ($t_0 = 50$) ↑
LLaVA $_{7B}$	64.9	55.6
LLaVA$_{7B}$ M3ID	**70.3**	**65.7**
LLaVA$_{13B}$	63.8	57.9
LLaVA$_{13B}$ M3ID	**77.5**	**69.8**

D.4. Conditioning dilution in VQA

As outlined in the main body of the text, VQA binary classification tasks, such as POPE, do not require the generation of multiple tokens. Nonetheless, the distance between the "yes/no" response and the image tokens is influenced by the specific input template chosen to prime the model for VQA. In our experiments, we employ the format `[Img][Model Prompt][Question]`, where the prompt can be an empty string or any specific instruction such as: "Answer the following question using the image content. Respond with yes or no.". Consequently, the length of both the prompt and the question adversely affects the visual prompt dependency measure. Indeed, as demonstrated in Tab. 10, the dilution effect can manifest even in binary VQA tasks using prompts of different lengths. In particular, the longer system prompt is obtained by adding "neutral" sentences like: "You must answer with either Yes or No. You will be evaluated with Accuracy, Precision, and Recall as the evaluation metrics.".

To mitigate this issue, when using M3ID on POPE, we introduce an offset t_0, i.e., $\gamma_t = e^{-\lambda(t-t_0)}$, corresponding to the number of tokens in-between the output token and the image tokens. Our findings in Tab. 10 indicate that incorporating this offset allows M3ID to significantly improve performance as the output token gets pushed far from the image content.