# 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 $\langle$object$\rangle$ present in the image?", where we sample $\langle$object$\rangle$ randomly (Random), among the most frequent objects in the dataset (Popular), or among the objects that frequently co-occur with $\langle$object$\rangle$ (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 $\alpha$ in the range $\{0, 0.01, 0.1, 0.3, 0.5, 0.8, 1.0\}$ and the optimal fading memory coefficient $\lambda$ in $\{0.001, 0.005, 0.01, 0.02, 0.03\}$. For PMI [25], we search $\tau$ in $\{0, 0.01, 0.1, 0.3, 0.5, 0.8, 1.0\}$ and $\mu$ in $\{0.1, 0.3, 0.5, 0.8, 1.5, 2.0\}$. For Contrastive Decoding [8], we search $\xi$ in $\{0.1, 0.3, 0.5, 0.8, 1.5, 2.0\}$ and $\psi$ 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 $2 \times 10^{-5}$, 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 $\mu$ 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 $p_{\text{exp}}$ to contain plausible and fluent text with the textual input prompt by amplifying its prediction difference with respect to a weaker *amateur* model $p_{\text{ama}}$. 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 $\gamma_t$. 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 $\lambda$ 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 $p_{\text{ref}}$, 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.

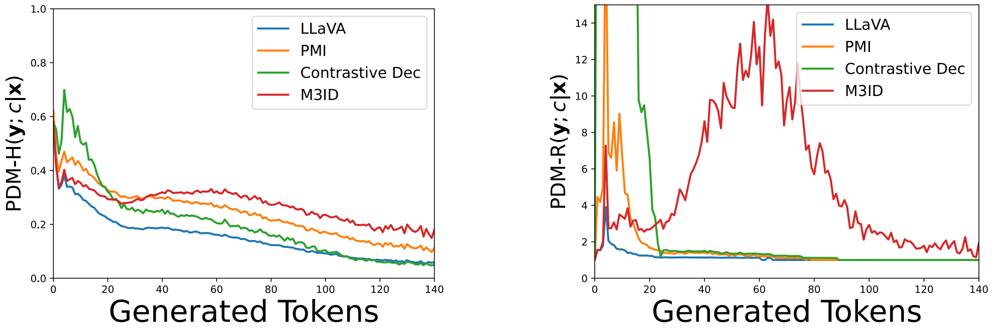


Figure 5. **PDMs with different distances.** We report how PDM varies with different choices of the distance function, i.e., PDM-H with Hellinger (left panel) and PDM-R with Rank (right panel). We compare the generation of a caption with M3ID, standard, and other context-aware decoding schemes. With 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 $\text{rank}_p$ 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

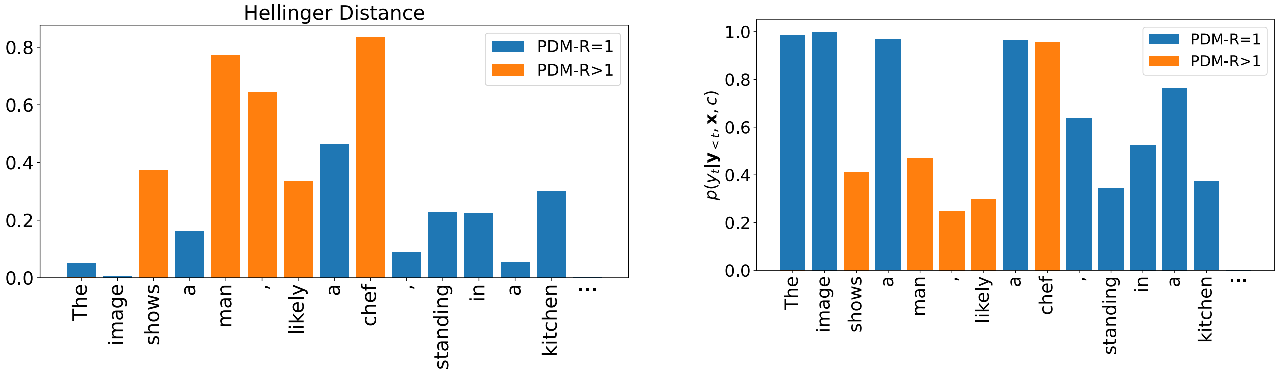


Figure 6. **How much does the conditioning prompt "surprise" the unconditioned predictor?** We report the PDM-H and the PDM-R for each token in the string "The image shows ...". Both distances increase on tokens for which visual information is required and decrease for punctuation, articles, and when 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 $\lambda$ 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

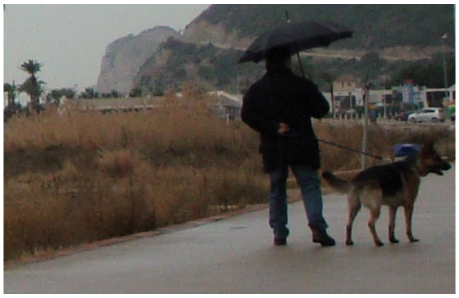
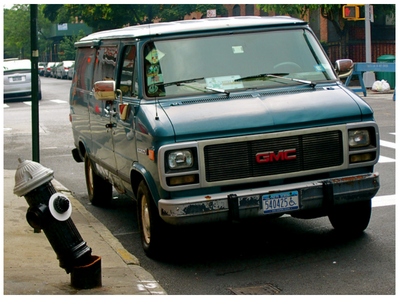
**Detailed captioning and stochasticity due to sampling.** Tab. 5 presents the CHAIR and cover metrics for detailed captioning, where LLaVA is prompted to generate longer captions resulting in an increase of the cover metric from approximately 55% to about 70%. In this scenario, although all methods 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<br>M3ID | correct<br>correct | correct<br>incorrect | incorrect<br>correct | incorrect<br>incorrect |
|---|---|---|---|---|
| LLaVA$_{7B}$ | 72.0 % | 3.0 % | 16.2 % | 9.8 % |
| LLaVA$_{13B}$ | 74.2 % | 2.4 % | 18 % | 5.4 % |

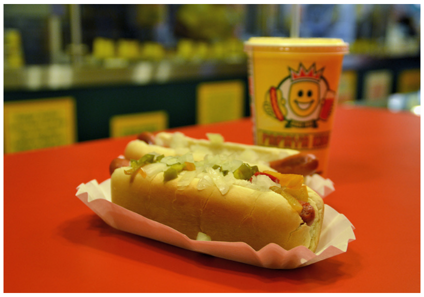
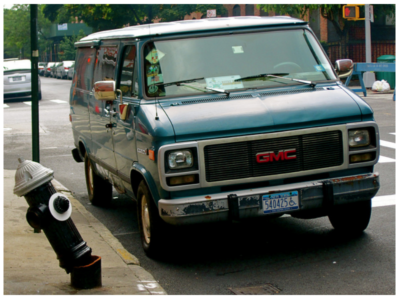
Table 7. **Examples of overcompensation.** Maximally surprising the VLM might lead to "overcompensation", a phenomenon affecting decoding interventions that amplify the logits difference between the conditioned and unconditioned models like M3ID. When M3ID's hyper-parameters force a strong correction 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 $\alpha$ ($\alpha = 1$ 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 $\alpha$. 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.