Controlling Multimodal LLMs via Reward-guided Decoding

Supplementary Material

6. Experiments

6.1. Details on evaluation metrics

We evaluate object precision and recall with standard metrics from the corresponding benchmarks, defined as follows.

CHAIR$_i$ (C$_i$) [38], CHAIR [46]. Measure the fraction of hallucinated objects in the generated captions.

$$\mathrm{C}_i/\mathrm{CHAIR} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}$$

CHAIR$_s$ (C$_s$) [38], Hal. [46]. Measure what fraction of generated captions include a hallucinated object.

$$\mathrm{C}_s/\mathrm{Hal.} = \frac{|\{\text{captions with a hallucinated object}\}|}{|\{\text{all captions}\}|}$$

Recall (Rec.), Coverage (Cov.) [46]. Measure the fraction of ground-truth objects covered in the generated captions.

$$\mathrm{Rec./Cov.} = \frac{|\{\text{correct objects}\}|}{|\{\text{all ground-truth objects}\}|}$$

6.2. Details on reporting results of existing methods

In Table 1, we report results for existing hallucination mitigation methods from the best source available. Unless otherwise specified, values are directly copied from the corresponding papers. For HA-DPO [55] and EOS [51], values are copied from Sarkar et al. [39] since their evaluation setup matches ours. For LLaVA-RLHF [42] and VCD [21], we compute results by generating captions with the original code and evaluating them on CHAIR [38] and AMBER [46], since the original papers do not report hallucination results on these benchmarks. For CGD [12], we also run the original code to generate captions for both AMBER and the full standard set of 5000 examples in the CHAIR benchmark (instead of the 500-example subset used by Deng et al. [12]).

6.3. Prompting baseline

We propose multimodal reward-guided decoding (MRGD) as a method to control the behavior of MLLMs at inference time. A common approach to steer the behavior of LLMs at inference time is prompting [8]. Here, we apply the same idea to MLLMs as an alternative approach to control their behavior. To mitigate visual hallucinations in image captioning, we use the instruction "`{captioning instruction}. Provide an accurate and objective description, focusing on verifiable visual elements such as colors, textures, shapes, and compositions. Avoid making assumptions, inferences, or introducing information not present in the image`", where the captioning instruction is the one described in Section 4.1: "`Describe this image in detail`" for LLaVA-1.5 and "`Describe this image in a few sentences`" for Llama-3.2-Vision. We maintain greedy decoding for the prompting baselines. In Tables 1 and 2, we observe that prompting slightly reduces object hallucinations compared to greedy decoding for LLaVA-1.5, while for Llama-3.2-Vision, surprisingly, it does not help much and, in fact, it increases the sentence-level hallucination rate (CHAIR$_s$ and Hal.). Instead, with LLaVA-1.5 on COCO, for a similar level of object recall ($\sim$81%), MRGD with w=0.25 achieves better object precision by $\sim$5.8% CHAIR$_i$ and $\sim$11.4% CHAIR$_s$ compared to prompting. This suggests that prompting is not a very effective strategy to steer MLLMs towards complex behaviors such as reducing visual hallucinations.

6.4. Ablation studies

Using SigLIP for r_{hal}. CGD [12] can be viewed as a particular instance of MRGD when using off-the-shelf SigLIP as the reward model for object hallucinations and removing the combination of multiple reward models (i.e., setting w=1.0). Therefore, we also conduct an ablation of MRGD replacing PaliGemma fine-tuned on preference data (Section 3.1.1) with off-the-shelf SigLIP-SoViT-400m [8]. Due to SigLIP's limited context length of 64 tokens, we only evaluate the last generated sentence, unlike PaliGemma which receives the full prefix response (which may contain several sentences). To ensure that the scores from multiple reward models are comparable and can be combined effectively, we normalize their ranges. In particular, since the effective range of SigLIP scores is much narrower than that of the reward model for object recall ($r_{\text{rec}} \in [0,1]$), we linearly rescale SigLIP scores $r \in \mathbb{R}^k$ to cover the range $[0,1]$: $r = (r - \min(r))/(\max(r) - \min(r) + \epsilon)$, where min and max are computed across the set of candidate samples Y, and ϵ is a small value to avoid division by zero (in case all candidates obtained the same score). In Table 4, we observe

[8] `google/siglip-so400m-patch14-384`

Table 4. Additional results for LLaVA-1.5$_{7B}$. MRGD with k=30 and T=1. MRGD$_{\text{PaliGemma}}$ indicates MRGD using PaliGemma fine-tuned on preference data for r_{hal}, MRGD$_{\text{SigLIP}}$ indicates MRGD using off-the-shelf SigLIP for r_{hal}.

Decoding strategy	COCO				AMBER		
	C$_i$ (↓)	C$_s$ (↓)	Rec. (↑)	Len.	CHAIR (↓)	Hal. (↓)	Cov. (↑)
Greedy	15.05	48.94	81.30	90.12	7.6	31.8	49.3
MRGD$_{\text{PaliGemma},w=1.0}$	4.53	18.19	76.04	95.90	3.4	15.9	52.4
MRGD$_{\text{PaliGemma},w=0.75}$	4.76	19.28	76.84	96.17	3.2	17.3	56.7
MRGD$_{\text{PaliGemma},w=0.5}$	5.34	22.54	78.63	97.96	4.4	25.4	60.8
MRGD$_{\text{PaliGemma},w=0.25}$	7.67	32.63	81.56	105.34	6.5	37.7	63.8
MRGD$_{w=0.0}$	24.20	73.42	85.23	108.92	14.8	65.0	64.3
MRGD$_{\text{SigLIP},w=1.0}$	7.19	28.00	73.71	92.73	6.0	30.1	48.5
MRGD$_{\text{SigLIP},w=0.75}$	7.57	29.58	74.30	93.17	6.1	30.3	50.0
MRGD$_{\text{SigLIP},w=0.5}$	8.17	32.88	75.96	94.93	6.3	33.3	53.4
MRGD$_{\text{SigLIP},w=0.25}$	10.84	43.58	79.50	99.57	8.5	46.2	57.8

Table 5. Ablation results for LLaVA-1.5-7B. MRGD with k=30, T=1, w=0.5. MRGD$_{\text{PG2}}$ means using PaliGemma-2 instead of PaliGemma for r_{hal}, MRGD$_{\text{+RLAIF-V}}$ indicates removing RLAIF-V from the original data mix for r_{hal}, MRGD$_{\text{+RLAIF-V}-\text{POVID}}$ means adding RLAIF-V and removing POVID from the original data mix, MRGD$_{\text{DETR}}$ denotes using DETR instead of OWLv2 as object detector for r_{rec}, and MRGD$_{\tau=x}$ denotes using x instead of 0.5 as semantic similarity threshold for r_{rec}.

Decoding strategy	COCO				AMBER		
	C$_i$ (↓)	C$_s$ (↓)	Rec. (↑)	Len.	CHAIR (↓)	Hal. (↓)	Cov. (↑)
Greedy	15.05	48.94	81.30	90.12	7.6	31.8	49.3
MRGD	5.34	22.54	78.63	97.96	4.4	25.4	60.8
r_{hal} variants							
MRGD$_{\text{PG2}}$	5.88	27.07	78.76	105.25	4.1	25.0	59.6
MRGD$_{\text{+RLAIF-V}}$	7.83	29.68	77.54	94.26	6.3	33.2	57.1
MRGD$_{\text{+RLAIF-V}-\text{POVID}}$	8.17	34.08	79.03	104.04	5.1	29.3	59.9
r_{rec} variants							
MRGD$_{\text{DETR}}$	5.37	23.76	82.04	99.24	4.0	19.8	53.5
MRGD$_{\tau=0.2}$	5.89	24.46	78.09	106.86	4.3	22.8	54.5
MRGD$_{\tau=0.9}$	5.00	20.96	78.36	98.09	4.0	22.3	61.2

that when using a SigLIP-based r_{hal}, our MRGD strategy is still effective in reducing object hallucinations and enabling the user to trade off object precision and recall on-the-fly at inference time. However, SigLIP does not allow to reach the same level of object precision, and the trade-off with object recall is also worse. For instance, when w=1.0, MRGD$_{\text{PaliGemma}}$ achieves better object precision by ~2.7% CHAIR$_i$ and ~9.8% CHAIR$_s$, and better Recall by ~2.3% compared to MRGD$_{\text{SigLIP}}$.

Preference data mix for r_{hal}. To understand the impact of different preference data compositions on the quality of r_{hal}, we conduct an ablation over the datasets used for its training. Our base reward model is trained on a mixture of LLaVA-RLHF [42] (9.4k), RLHF-V [49] (5.7k), and POVID [56] (17k). We consider a new preference dataset, RLAIF-V [50] (83k), which contains 2.6× more examples than all previous datasets combined. We train two additional variants: (1) adding RLAIF-V and (2) adding RLAIF-V while removing POVID. As shown in Table 5, both adding RLAIF-V and removing POVID lead to notable performance degradation, highlighting the importance of carefully choosing the preference data mix to train r_{hal}.

6.5. MRGD's robustness to reward models' quality

To further assess the robustness of MRGD to variations in reward model quality, we evaluate the performance of our approach for a variant of r_{hal} with a different model backbone (PaliGemma-2$_{3B}$[9] [41] instead of PaliGemma), and several variants of r_{rec}: different object detector (DETR [10] instead of OWLv2) and different semantic similarity thresholds (τ). In Table 5, we observe that (1) upgrading r_{hal}'s backbone yields similar performance, (2) using DETR for r_{rec} performs similarly for COCO and decreases both hallucinations and coverage for AMBER, and (3) MRGD remains effective when varying τ. Overall, all ablated variants significantly outperform the greedy search baseline on most metrics, demonstrating the effectiveness and robustness of our approach across reward design choices.

[9] `google/paligemma2-3b-pt-224`