Behind the Magic, MERLIM: Multi-modal Evaluation Benchmark for Large Image-Language Models

Supplementary Material

To complement the experimental assessments in Section 4, we provide a summary of comparisons across a subset of tasks and evaluation metrics in Figure 9. This is followed by an in-depth analysis of the effects of our inpainting rationale and a direct comparison between IT-LVLMs and their LLM-only counterparts in MERLIM. Additionally, we offer insights into language biases by examining the average number of nouns generated for each of the five proposed prompts in Section 3.1. Lastly, we delve into the object hallucination issues by presenting the full Precision and Recall metrics for the models across the original image set, using the same five prompts from Section 3.1. Additionally, we analyze the gradient percentage corresponding to each input (Image, Question, and Answer) to understand their contributions to the generated answers.

6. Analysis of Inpainted Images

To analyze the visual changes in the images after the inpainting process, we first assess that the selected inpainting strategy induces a minimal change in the global image features and that the inpainted masks effectively hide the object instance from object detectors. We use YOLOv7 [50] to verify the absence of the object in the edited image and ResNet50 [19] to evaluate the global feature similarity with the original image. For each image in MERLIM, we enforce two essential criteria: (1) YOLOv7 must not generate a detection box with an Intersection over Union (IoU) greater than 0.7 relative to the ground-truth bounding box of the removed instance, and (2) the cosine similarity between the ResNet50 features of the edited image and the original image must exceed 0.7.

For a more in-depth assessment of the visual changes in the inpainted images, we use another box detector, Mask R-CNN [20], to verify if each predicted box in the inpainted image has a corresponding box with the same class and a similar spatial location (i.e., high Intersection over Union) in the original image.

We perform this analysis for bounding boxes with confidence scores above 0.7 and 0.5, summarizing the results in Table 3.

Out of a total of 762782 predicted boxes in the inpainted image set, we found that only 4455 boxes (about 0.59%) could not be mapped to the boxes detected in the original image with the same class and an overlap of at least 0.5 IoU. In comparison, 726969 boxes (96.8%) in the inpainted images have a matching box with the same class and a nearly identical spatial location (IoU $>$ 0.7) in the original image.

Table 3. **Analysis of Inpainted Images.** We analyze the Intersection over Union (IoU) of each box predicted by Mask R-CNN in the edited image set of MERLIM. Table a) presents the analysis for bounding box predictions with a confidence score greater than 0.7, while Table b) shows the analysis for box predictions with a confidence score greater than 0.5. We find that up to 3290 boxes, representing 0.7% of the total, do not exhibit a high overlap with the original boxes.

IoU	Percentage	# of boxes
above 0.9	85.37%	259454
0.8-0.9	9.3%	28250
0.7-0.8	3.1%	9428
0.6-0.7	1.4%	4260
0.5-0.6	0.44%	1324
bellow 0.5	0.38%	1165

(a)

IoU	Percentage	# of boxes
above 0.9	80.68%	360595
0.8-0.9	10.97%	49054
0.7-0.8	4.49%	20097
0.6-0.7	2.23%	10003
0.5-0.6	0.84%	3768
bellow 0.5	0.7%	3290

(b)

Moreover, in Figure 5, we also provide visual examples of the resulting images after the inpainting process. As can be seen, the inpainting images are highly similar to the original ones. The editions are visually indistinguishable even when large objects are removed. Despite the inpainting strategy being relatively under-explored, we conclude that the observed performance gaps in MERLIM should not be attributed to the minimal discrepancies identified by representative image networks (ResNet50, YOLOv7, Mask R-CNN) but rather to hallucination events in the IT-LVLM.

7. Object Hallucination

In Figure 6, we provide further evidence of instruction bias in IT-LVLMs. Despite the semantic equivalence of the proposed prompts, all methods, except for BLIP-2 [30] and Kosmos-2 [43], generate varying numbers of nouns for each prompt. Figure 7 illustrates that while BLIP-2 produces the shortest answers (in terms of noun count), it consistently achieves higher precision compared to other methods, albeit with lower recall. This indicates that BLIP-2 is less prone to hallucination, likely because it was trained to generate concise captions rather than detailed descriptions. On the other

Figure 5. **Examples of Inpainting images.** We present eight examples demonstrating the results of removing an object from the original image using the inpainting model proposed by [31]. As illustrated, the edits are visually seamless, even when large objects are removed, making them nearly indistinguishable from the original.

hand, larger models trained with instructional data, such as InstructBLIP and MiniGPT-4 with Vicuna-13B [9, 58], LLaVa [36], and xGen-MM [52], tend to generate longer answers on average. This increases the likelihood of hallucinations, particularly when visual grounding is insufficient, resulting in higher recall but lower precision compared to BLIP-2. As shown in Figure 8 shows that even a proprietary model such as GPT-4o-mini [41] presents a high hallucination rate.

8. LLMs vs IT-LVLMs

In Table 4, we compare the performance of IT-LVLMs to their corresponding LLMs without visual input. Since the LLMs lack the capability to process visual information, we allow them to answer "do not know." Using the original questions from the visual relationship task (denoted as **Question**), we prompt an LLM as follows: *"*p_{llm} = **Question**. *Please answer yes, no, or do not know"*. While

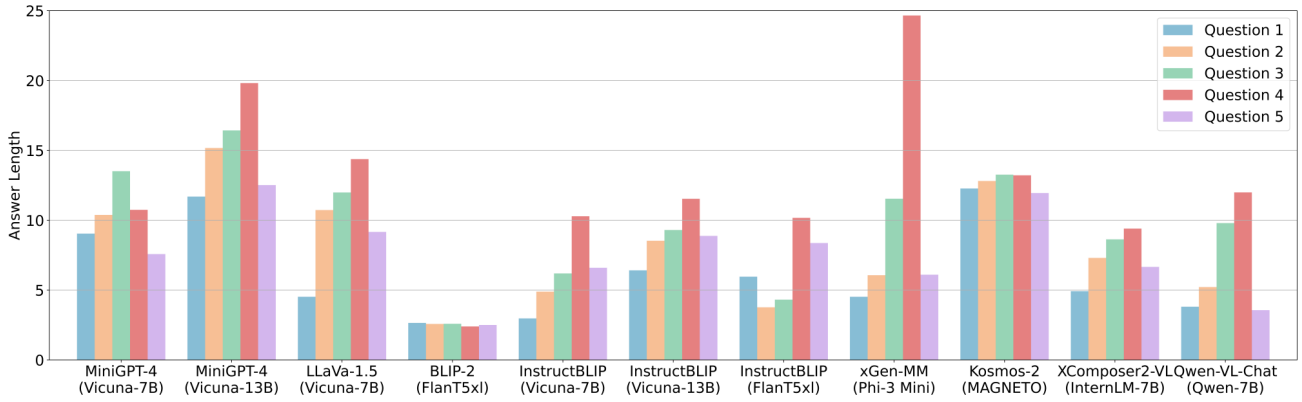
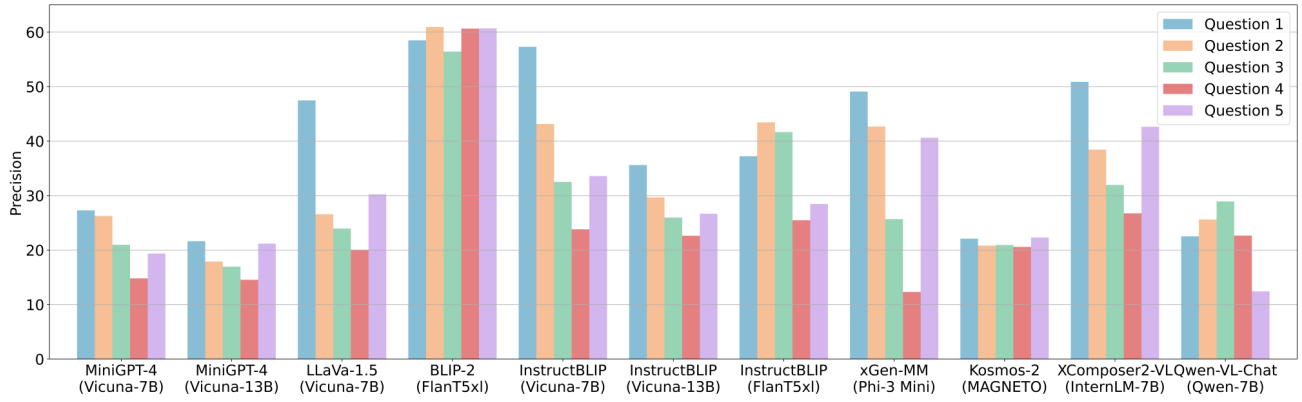
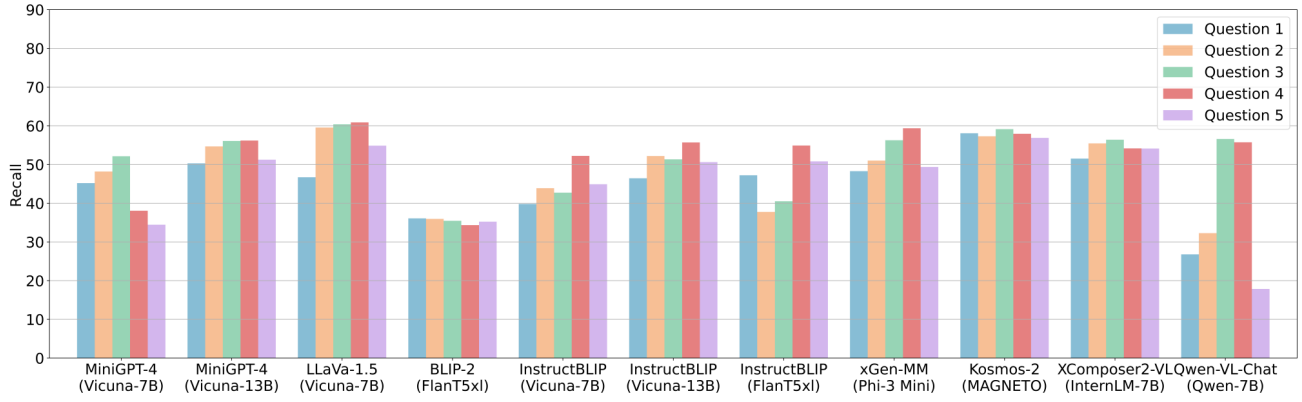


Figure 6. **Number of nouns predicted.** We extract the nouns using the spaCy library from the answers of the models from all the formulated prompts on the original image set. It is worth noting that BLIP2 predicts a consistent number of nouns across all the prompts, unlike the other methods.



(a)

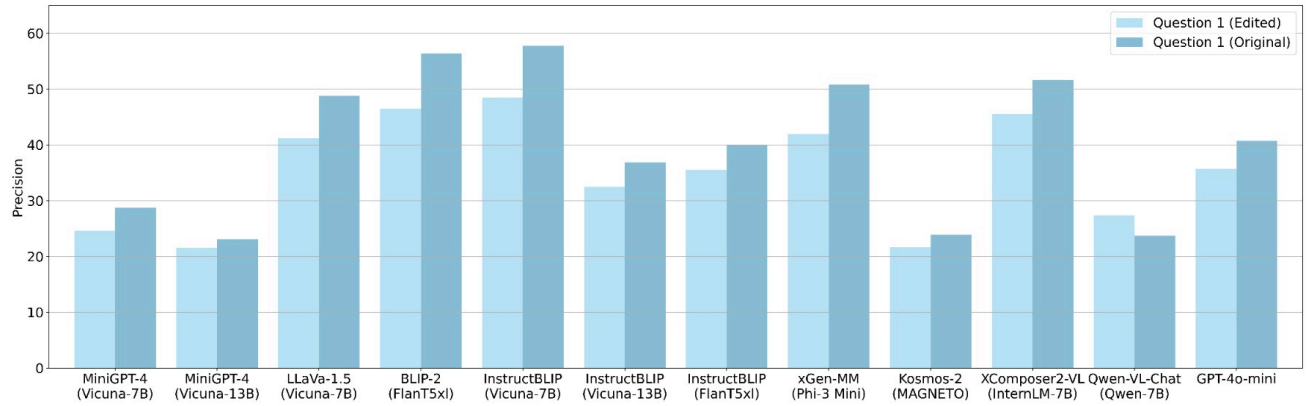


(b)

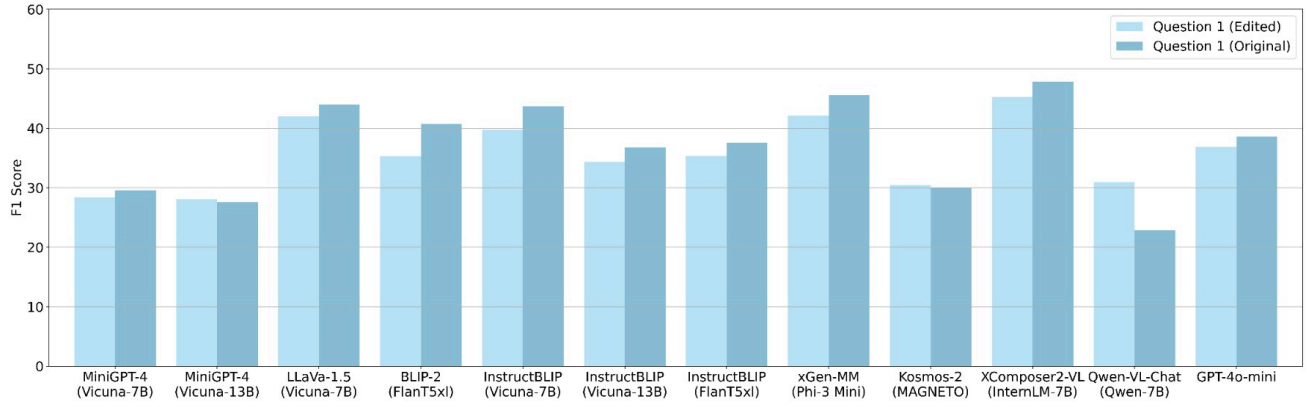
Figure 7. **Precision and Recall on the original image set.**. We compare the Precision and the Recall of IT-LVLMs on the original image set across the five proposed prompts P in the sub-figure **a)** and **b)**, respectively. It is worth noting that the biggest models, such as InstructBLIP and MiniGPT4 with Vicuna13B and LlaVa, get the highest recall and lowest precision.

LLMs sometimes respond with "I do not know", they typically base their answers on the knowledge acquired during language pre-training. With three response options, the random chance of selecting the correct answer is 33%.

For additional evaluation (shown on the right side of the table), we consider only the answers where the LLM



(a)



(b)

Figure 8. **Comparing GPT-4o-mini with the open-source models on the original and edited sets.** (a) We compare the precision of IT-LVLMs on the edited and original image sets using one prompt. To analyze the hidden hallucination problem, we focus on the subset where the image inpainting removes an entire category. Notably, all methods lose performance on the edited image set. (b) We also report the F1 Score on the original and edited images using the same question.

Table 4. **Textual Biases.** We compare the accuracy of IT-LVLMs in the relationship understanding with the performance of LLMs. Since the LLMs lack visual context we allow the model to reply 'yes', 'no' or 'I Don't know'. We report two accuracy values, one where we penalized every 'I don't know' as a false prediction (left side of /) and the other considering only the questions answered with yes or no (right side of /). FlanT5xl outperforms the other opensource LLMs.

Model	LLM	Random Set		Curated Set	
		$\mathbf{Acc_{org}} \uparrow$	$\mathbf{Acc^{neg}_{org}} \uparrow$	$\mathbf{Acc_{org}} \uparrow$	$\mathbf{Acc^{neg}_{org}} \uparrow$
Random Baseline	N/A	33% / 50%	33% / 50%	33% / 50%	33% / 50%
LLM Only	ChatGPT 3.5	21.27% / 80.01%	8.05% / 28.83%	13.06% / 55.80%	11.57% / 47.78%
LLM Only	Vicuna-7B v1.1	6.59% / 44.72%	27.53% / 66.86%	11.11% / 43.05%	26.42% / 55.54%
LLM Only	Vicuna-13B v1.1	7.92% / 75.22%	2.07% / 25.23%	7.63% / 55.60%	4.92% / 47.76%
LLM Only	FlanT5xl	**41.82%** / 69.47%	**40.06%** / 55.84%	**36.10%** / 54.65%	**50.74%** / 65.80%
BLIP-2	FlanT5xl	83.62%	48.80%	67.84%	49.77%
InstructBLIP	Vicuna-7B v1.1	**91.17%**	19.60%	**76.75%**	45.31%
InstructBLIP	Vicuna-13B v1.1	90.32%	16.24%	75.99%	43.52%
InstructBLIP	FlanT5xl	80.69%	**77.16%**	70.15%	**63.90%**

responded with either "yes" or "no." Although Vicuna performs below the random baseline, its performance improves significantly when focusing solely on yes/no responses, especially for Vicuna-13B. FlanT5xl exceeds random chance and further narrows the performance gap between LLMs and IT-LVLMs.

We conclude that the language model serves as a strong prior for resolving visual relations, even without image information. These textual priors can act as "shortcuts" in MERLIM, contributing to the performance difference between the Random and Curated sets, where these biases are less effective. While these textual "shortcuts" can occasionally align with visual grounding (Hidden Hallucinations), they are often revealed when visual grounding contradicts language-based intuitions.

9. Gradient Analysis

To further investigate the visual grounding of IT-LVLMs, we take inspiration from [47] and compute the proportion of the total gradient attributable to each specific type of input (Image, Question, Answer) for each predicted token. We analyze the model's output logits before token selection and propagate the gradient to the input by identifying the maximum activation value per output token. Specifically, for each output token, we determine the gradient contribution from each input type and then average these contributions across all answers. To simplify our analysis, we focus on architectures with autoregressive LLMs, such as Vicuna, which explicitly use previous output tokens as inputs for predicting subsequent tokens.

As shown in Figure 5, the gradient of the visual input is significantly lower than that of the language inputs (Question and Answer). This indicates that LLMs prioritize language tokens over visual ones when predicting answers, leading to hallucinations based on language biases. Additionally, we observe that only a few tokens account for most of the visual gradients. For instance, in InstructBLIP with Vicuna-7B, just 10 out of 30 tokens represent nearly 73% of the visual gradients, making it challenging for specific and strong visual information to be adequately considered. These findings support the results of our previous experiments.

10. Number of unsuitable outputs

On the Object Recognition task models will occasionally provide responses lacking any valid nouns, such as "There are many objects" or "It is a beautiful scene". In such cases, we set both recall and precision metrics to 0. The results in Table 6 outline that MiniGPT-4 with Vicuna-7B v0 and BLIP-2 with FlanT5xl produce the largest number of unsuitable answers. Conversely, the InstructBLIP methods consistently generate valid responses with object lists.

Table 5. **Study of the relevance of the inputs.** We analyze the relevance of tokens from each kind of input (Image, Question, and Answer tokens) by computing the portion of the total gradient produced in the input that belongs to the image, question, and answer tokens relative to the outputs. Specifically, for each output, we calculate the gradient contribution from each input type and then average these contributions across all answers.

Model	Vis. Tokens	Que. Tokens	$\nabla_{\text{vis}} \uparrow$	$\nabla_{\text{que}} \uparrow$	$\nabla_{\text{ans}} \uparrow$
LLaVA-1.5	575 (All)	52 (All)	27.68%	44.75%	28.67%
LLaVA-1.5	top 10	top 10	11.47%	53.28%	36.68%
InstructBLIP	32 (All)	10 (All)	24.89%	47.45%	29.04%
InstructBLIP	top 10	top 10	19.98%	52.01%	29.41%

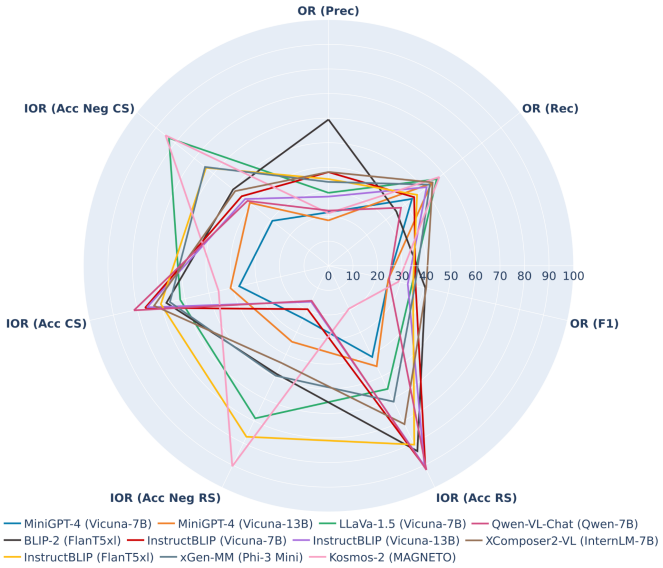


Figure 9. **Results on MERLIM.** We showcase a subset of the performance evaluations on MERLIM. Our evaluation metrics include Precision (Prec), Recall (Rec), and F1 Score (F1) for the Object Recognition Task (OR). We also measure the Accuracy at identifying inter-object relationships (IOR) in two sets: one set generates negative examples through random sampling (Random - Sample RS), and the second set has curated relations (Curated Set - CS) where a commercial LLM discards impossible associations, thus forcing the IT-LVLM to use the visual data. We further verify the IT-LVLMs consistency, as we calculate the Accuracy for both affirmative (Acc) and negative (Acc Neg) versions of the instructions describing the object relations.

11. Visual Examples of MERLIM's tasks.

Figure 10 presents additional visualizations of the tasks evaluated by MERLIM. It shows the InstructBLIP outputs for both the original and edited images across the three tasks. Notably, the model consistently produces visually ungrounded predictions (hallucinations) in all scenarios, highlighting the complexity of the hallucination problem in these instructional models.

12. Impact Statements

MERLIM is a benchmark designed to assess the performance of Instruction Tuning Large Vision and Language Models (IT-LVLMs). Besides the empirical evaluation, MERLIM's primary aim is to identify and quantify instances

Table 6. **Number of unsuitable outputs.** The IT-LVLMs occasionally provide unsuitable answers to the five proposed questions for the Object Recognition Task. That is answers without valid nouns, for instance, *"There are many objects"* or *"It is a beautiful scene"*. For such instances, we establish both recall and precision metrics as 0.

Model	LLM	Num. unsuitable outputs									
		Question 1		Question 2		Question 3		Question 4		Question 5	
		Original	Edited	Original	Edited	Original	Edited	Original	Edited	Original	Edited
MiniGPT-4	Vicuna-7B v0	17	110	6	28	2	26	10	61	81	414
MiniGPT-4	Vicuna-13B v0	0	2	0	0	0	0	1	10	0	1
LLaVA-1.5	Vicuna-7B v1.5	0	1	0	0	0	0	0	0	0	0
BLIP-2	FlanT5xl	11	73	6	58	5	65	21	209	15	96
InstructBLIP	Vicuna-7B v1.1	0	1	0	2	0	0	0	0	0	0
InstructBLIP	Vicuna-13B v1.1	0	2	0	0	0	0	0	0	0	0
InstructBLIP	FlanT5xl	0	0	0	1	0	1	0	0	0	0
xGen-MM	Phi-3 Mini 3.8B	0	0	0	0	0	0	0	0	0	0
Kosmos-2	MAGNETO	0	0	0	0	1	12	0	0	0	0
XComposer2-VL	InternLM-7B	4	36	12	180	0	1	0	1	4	2
Qwen-VL-Chat	Qwen-7B	6	38	2	4	0	1	0	4	3	22

of "Hallucination" events in the textual responses generated by IT-LVLMS. Consequently, MERLIM represents a tool with a potentially positive societal impact by encouraging advanced IT-LVLM models to be more robust to hallucination events and, therefore, bring more factual and informative responses.

Figure 10. **Visual Examples.** We present additional visual examples illustrating the tasks evaluated by MERLIM. Sub-figures **(a)**, **(b)**, and **(c)** showcase comparisons of InstructBLIP outputs for original and edited images, focusing on the tasks of Object Recognition, Inter-Object Relationship Understanding, and Object Counting, respectively.