If I understand the context, I will act accordingly: Combining Complementary Information with Generative Visual Language Models

Anonymous ACL submission

Abstract

The effectiveness of autoregressive LLMs has allowed many language and vision tasks to be reframed as generative problems. Generative visual language models (VLMs) have recently shown potential across various downstream tasks. However, it is still an open question whether, and to what extent, these models can properly understand a multimodal context where language and vision provide complementary information—a mechanism routinely in place in human language communication. In this work, we test various VLMs on the task of generating action descriptions consistent with both an image's visual content and an intention or attitude (not visually grounded) conveyed by a textual prompt. Our results show that BLIP-2 is not far from human performance when the task is framed as a generative multiple-choice problem, while other models struggle. Furthermore, the actions generated by BLIP-2 in an open-ended generative setting are better than those by the competitors; indeed, human annotators judge most of them as plausible continuations for the multimodal context. Our study reveals substantial variability among VLMs in integrating complementary multimodal information, yet BLIP-2 demonstrates promising trends across most evaluations, paving the way for seamless human-computer interaction.

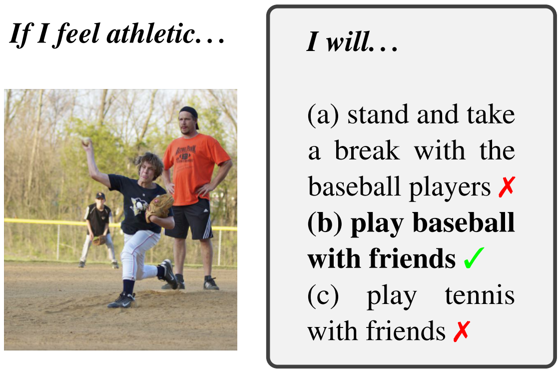

Figure 1: We test generative visual language models' (VLMs) abilities to combine *complementary* information brought into context by the two modalities. In this example from the BD2BB dataset (Pezzelle et al., 2020) (slightly edited for space reasons), only one of the actions on the right, (b), is consistent with both the textual prompt and the image on the left. As for (a) and (c), they are plausible based on the image or the textual prompt, respectively, but not on the combination of both.

1 Introduction

In recent years, transformer-based generative visual language models (VLMs) have shown outstanding results in many downstream tasks. Similar to what happened in NLP, where pre-trained generative models have supplanted previous architectures thanks to their flexibility and portability, VLMs have proven effective in solving language-and-vision tasks by turning them into generative problems. This is possible thanks to their massive multimodal pre-training, which typically builds on a pre-trained language model and image processing model. This has enabled systems that can, in zero-shot mode and without further fine-tuning, seamlessly describe the content of an image, answer questions about it, or engage in a dialogue (see Caffagni et al., 2024, for an overview). This might suggest that VLMs have skills similar to those needed for meaningful multimodal communication.

In real-life multimodal communication, human speakers continuously integrate complementary information from various modalities, including language and vision, to understand and convey messages and properly act in various situations (Partan and Marler, 1999; Benoît et al., 2000; Forceville, 2020). An example of such complementarity is shown in Figure 1: If someone observing the scene depicted in the image *feels athletic*, they would likely take an action that is consistent with both the visual content and their attitude or intention, i.e., *play baseball with friends*. In contrast, actions that are plausible given either the image or the textual intention, but not both, would not be considered.

Note that making this type of inference is also key for any multimodal model that aims to be communicatively plausible and useful. Consider the case of a virtual assistant that has access to the visual context and a spoken or written request from a user. If asked to recommend an appropriate activity to do in a specific context—e.g., *I feel adventurous. What do you recommend I do?*—the assistant should suggest something suitable for the user's location, and obviously in line with their attitude. Despite the relevance of the problem, only a few studies have investigated, to date, whether language-and-vision models master this ability. One notable exception is Pezzelle et al. (2020), which proposed the *Be Different to Be Better* (BD2BB) benchmark (see an example in Figure 1) to test the ability of multimodal encoders such as LXMERT (Tan and Bansal, 2019) to integrate complementary information. In that study, these models were shown to lag far behind human intuitions, leaving ample room for improvement in future systems. To the best of our knowledge, no subsequent work addressed whether generative VLMs have filled this gap.

In this research, we use the BD2BB benchmark and test how several generative VLMs deal with it. We do so employing two main experiments. First, we challenge the models to solve the task in its original multiple-choice format, i.e., by picking, for a given image, one among 5 candidate actions (*I will...*) that we give to the model via prompting together with the intention (*If I...*). We evaluate model performance in terms of accuracy, that we measure both *extrinsically* (considering the label, corresponding to a given action, that is output by the model) and *intrinsically* (looking at the probability assigned by a model to each action following the same intention). Second, we test VLMs in the setup that best suits them, that is, by letting them generate an action based on the image and the intention. In this case, we assess model performance using both a *reference-based*, automatic metric (we compute BERTScore similarity between the generated action and the target one from BD2BB) and a *reference-free*, human-based evaluation (we ask annotators to judge whether a certain action is good for a given <image, intention> pair).

The results of our first experiment show that, while most tested models hover around the chance level, BLIP-2 achieves fairly high accuracy, much closer to human performance than LXMERT (reported in Pezzelle et al., 2020). Similarly, in our second experiment, the actions generated by BLIP-2 are deemed plausible by human participants in most cases, which is not the case for other models. Taken together, these results highlight substantial variability across VLMs in their ability to combine complementary multimodal information. At the same time, the promising trends exhibited by BLIP-2 reveal that this model is capable of understanding—to some extent—the visual scene, the intention, and their complex interaction.

2 Related Work

2.1 Generative Language-and-Vision Models

With the introduction of Transformers (Vaswani et al., 2017), NLP research has experienced unprecedented development. This, in turn, influenced the work on language and vision processing, which followed the same 'evolutionary' steps. First, based on Masked Language Models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), the community proposed many multimodal encoders, either single-stream (i.e., jointly processing language and vision from the beginning), such as UNITER (Chen et al., 2020), VL-BERT (Su et al., 2019), and VisualBERT (Li et al., 2019), or dual-stream (i.e., processing language and vision separately, and later combining them through a series of multimodal layers), such as LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019).

Later, in the wake of the success of autoregressive Large Language Models (LLMs) such as GPT (Radford et al., 2019), OPT (Zhang et al., 2022) or LLaMA (Touvron et al., 2023), the language-and-vision community has taken a generative direction. With such an approach, answering questions about an image (VQA) or describing its content (IC) can be done by simply feeding the model with the image and the appropriate prompt. Various generative language-and-vision models have been proposed in recent years, such as BLIP-2 (Li et al., 2023), Flamingo (Alayrac et al., 2022), FROMAGe (Koh et al., 2023), MAPL (Mañas et al., 2022), and IDEFICS (Laurençon et al., 2023). In general, a common feature of all these models is that they leverage a pre-trained text-only LLM and a visual encoder, on top of which a relatively lightweight trainable network is learned. Such a network—which can consist of a bunch of Transformer (BLIP-2, Flamingo, IDEFICS), fully connected (MAPL), or linear layers (FROMAGe)—is responsible for connecting the two modalities and making the model capable of solving multimodal tasks. Using this strategy, generative language and vision models have achieved results never approached before (e.g., when introduced, Flamingo was the best-performing model on 16 multimodal tasks). Furthermore, their architecture makes these models much more flexible and portable than their predecessors, as they can be applied, without any fine-tuning, to virtually any unseen task.

2.2 Complementary Language and Vision

The models described above have been quite extensively tested in various downstream tasks, such as Visual Question Answering (Antol et al., 2015) and Image Captioning (Bernardi et al., 2016), which typically require dealing with *aligned* information from language and vision. To illustrate, these tasks challenge the models to locate a phrase or sentence in the image, retrieve information from it, or verify that what is depicted complies with a description. Comparably less attention has been paid to assessing whether, and to what extent, they can genuinely combine *complementary* information from the two modalities—something necessary, e.g., to generate a plausible action for the example in Figure 1.

This ability is certainly necessary for tasks such as Visual Dialog (Das et al., 2017; Mostafazadeh et al., 2017) or Visual Storytelling (Huang et al., 2016; Hong et al., 2023). In the former, multimodal models are asked to maintain a meaningful conversation starting from the contents of an image, which requires more than simply describing visible aspects. As for the latter, the goal is to produce a story based on a sequence of images. Again, this task requires not only understanding the visual content (which is, however, crucial; see Surikuchi et al., 2023), but also making inferences over people's emotions and feelings, understanding social dynamics, and so on. These are challenging tasks for large multimodal models, which were recently shown to have little social awareness and struggle with recognizing subtle and culturally diverse emotions (Deng et al., 2023). Similarly, these models face difficulties in handling semantically underspecified language (where the language signal needs to be complemented by extra information, e.g., visual info; see Pezzelle, 2023); moreover, they have trouble understanding humor (Hessel et al., 2023), an aspect of multimodal language use that can only be mastered by going beyond the literal (i.e., image-aligned) meaning of a sentence.

To explore more complementary scenarios, various directions have been taken. These include approaches to Image Captioning that are sensitive to the context and communicative purpose of the captions (Kreiss et al., 2021, 2022); tasks that challenge the models to predict something *external* to the multimodal sample, such as the motivation or intent of a social media post (Kruk et al., 2019), or the cause or consequence of an event (Hessel et al., 2022); datasets to test complex inference abilities in multimodal setups, such as predicting the next utterance or frame in a comic strip (Iyyer et al., 2017). BD2BB (Pezzelle et al., 2020) also belongs to this latter category, as it challenges models to predict *what comes next* based on both grounded (the image contents) and non-grounded information (the textual intention). In this work, for the first time, we study how generative visual language models deal with complementary multimodal information.

3 Methods

3.1 Data

We use the BD2BB dataset and corresponding multiple-choice task (Pezzelle et al., 2020). The task is exemplified in Figure 1: given an image and a textual intention (*If I...*), a model must select the correct action (*I will...*), i.e., the one that complies with both the visual and textual information. Note that, in BD2BB (and differently from what is shown in the figure), each sample comes with 5 candidate options—two that are valid given the image only (so-called *visual decoys*), two that are valid given the intention only (*language decoys*), and the correct one. The images in BD2BB come from a subset of COCO images (Lin et al., 2014) depicting at least one person.[1] The dataset, collected via crowdsourcing and further post-processed, includes around 10K <image, intention, candidate actions> samples. In this work, we test models in a zero-shot setup (without training or fine-tuning them) on the test set, which includes 4081 samples.

3.2 Models

We experiment with four state-of-the-art, open-source generative VLMs, i.e., MAPL, FROMAGe, BLIP2, and IDEFICS. As mentioned in Section 2.1, these models are all based on a similar architecture that leverages two frozen pre-trained unimodal models (a language and a vision one) and learns

[1]This choice is meant to increase the likelihood of interacting with these images by performing some action.

	MAPL	FROMAGe	BLIP-2	IDEFICS
Publication year	2022	2023	2023	2023
Underlying language model	GPT-J	OPT	OPT / FlanT5	LLaMA
Underlying vision model	Vit-L14	Vit-L14	Vit-L14 / Vit-G14	OpenClip[5]
Mapping network's architecture	Fully connected layers	Linear layers	Transformer	Transformer
# trainable parameters	3.4M	5.5M	188M	1.4B
Generated output	Text	Text / Image	Text	Text
Trained with COCO?	no	no	yes	no
Visual model trained with COCO?	no	no	no	no

Table 1: A comparison of the four VLMs used in this work concerning some of their main features.

a relatively lightweight mapping network on top of them. Below, we briefly describe these models from smallest to largest in terms of learnable parameters. For convenience, we provide an overview of their most important features in Table 1. We refer the reader to the original papers for further details on each model's architecture, training data, and optimization strategies.

MAPL (Mañas et al., 2022) builds on CLIP (Radford et al., 2021) and GPT-J (Wang and Komatsuzaki, 2021) as a visual and language frozen model, respectively. The trainable network to map visual features into token embeddings consists of a few fully connected layers with ReLU activations (Nair and Hinton, 2010) and dropout regularization (Srivastava et al., 2014). With only trainable 3.4M parameters, this network is the lightest of the four we use in this work.

FROMAGe (Koh et al., 2023) leverages CLIP Vit-L14 (Radford et al., 2021) and OPT (Zhang et al., 2022) as its frozen visual and language model, respectively. The projection of the image and text representations into a common latent space is done through several trainable linear layers. This makes this model lightweight, with only 5.5M trainable parameters. Among the four models we use, FROMAGe is the only one capable of producing outputs including both text and images.

BLIP2 (Li et al., 2023) bootstraps language-and-vision representations from the underlying frozen pre-trained unimodal models via a Transformer-based network. It allows using various underlying frozen models: CLIP Vit-L14 (Radford et al., 2021) or Vit-G14 from EVA-CLIP (Fang et al., 2023) on the vision side; OPT (Zhang et al., 2022) or FlanT5 (Chung et al., 2022) on the language side (here, we use the version with FlanT5 and Vit-G).

The multimodal mapping is carried out by a trainable Querying Transformer (Q-Former) network. The Q-Former includes two transformer submodules sharing self-attention layers: an image transformer interacting with the frozen image encoder for visual feature extraction, and a language transformer serving as both a text encoder and decoder. It is worth noting that, among the four models here considered, BLIP-2 is the only one also trained with images from COCO (Lin et al., 2014), i.e., the images used to build the BD2BB dataset. Though the model has not seen the BD2BB data, it could still have an advantage over other architectures.

IDEFICS (Laurençon et al., 2023) is the most recent model among the four we tested in this work. It is an open-access re-implementation of the Flamingo model (Alayrac et al., 2022) which leverages LLaMA as the language model (Touvron et al., 2023) and OpenClip[5] (a model pre-trained with a contrastive text-image approach, similar to CLIP Radford et al., 2021) as the vision model. Similar to BLIP-2, IDEFICS uses a Transformer-based architecture to connect language and vision. In particular, it employs a Perceiver Resampler module to map varied-size vision features to a few tokens, which are then used to condition the frozen LM through cross-attention layers. We employ the 9B parameter instructed version with 1.4B trainable parameters, nearly 10 times more than BLIP-2. This makes IDEFICS the largest model we consider.

3.3 Experimental Settings

We test the four models in two experiments: a multiple-choice experiment (Section 4) and an open-ended generative experiment (Section 5). In both experiments, we test the pre-trained models in a zero-shot manner.[2] That is, we do not further

[2]The pre-trained models can be downloaded from: https://github.com/octarinesec/MAPL (MAPL) https://github.com/kohjingyu/fromage (FROMAGe) https://huggingface.co/docs/transformers/en/model_doc/blip-2 (BLIP-2) https://huggingface.co/docs/transformers/en/

train or fine-tune them.[3] We ran the models on an A1000 GPU using their default hyperparameters to ensure deterministic results. We also conducted the multiple-choice experiment with other hyperparameter settings (see Appendix A).

4 Multiple-Choice Experiment

We test the four generative models in the original BD2BB multiple-choice classification task. Here, together with the intention and the image, we provide the model with the five candidate actions and task the model to select the correct one. We evaluate model performance in terms of accuracy, which we measure both *intrinsically* and *intrinsically*. Below, we describe the two evaluations in more detail.

Extrinsic evaluation Given an <image, intention, actions> sample, we ask the models to provide the correct action via prompting. Since we present the candidate actions as options preceded by an alphabet letter *(A-E)*, models are expected to output the letter corresponding to the action they consider correct. To elicit model responses, we used the following template, filled with the intention, the five actions, and a prompt describing the task: "[intention], [prompt]: A. [action$_1$] B. [action$_2$] C. [action$_3$] D. [action$_4$] E. [action$_5$]". Given this template, we experiment with 30 prompts (provided in Appendix B) and compute average accuracy and standard deviation over them. An example of a template filled with all information for one dataset's sample is the following (we give the prompt in italic): "If I feel adventurous, *what should I do? Choose the best option from the following:* A. I will ride an elephant. B. I will merely watch my friend fly an animal kite. C. I will go bird watching on an outdoor public patio. D. I will ride a horse like the man. E. I will stand and observe the zebras."

Intrinsic evaluation Given an <image, intention, actions> sample, we consider its 5 <intention, action> pairs and compute the cross-entropy loss between each of these sequences (we concatenate the intention and the action) and the image. To do so, we first obtain the logits from the model's final hidden layer for the current input sequence. Then, we calculate the cross-entropy loss between these logits and the target tokens. The total cross-entropy loss for a sequence is the sum of the losses at each word position. The sequence with the lowest cross-entropy loss is selected as the model answer. These predictions are used to compute model accuracy.

model	accuracy	
	intrinsic	*extrinsic*
LXMERT*	62.2±2.2	
CLIP	53.2	
MAPL	63.1	22.0±0.8
FROMAGe	47.9	20.0±0.5
BLIP-2	42.0	**75.7±0.8**
IDEFICS	**63.7**	35.5±7.2
Humans*	79.0	

Table 2: Multiple-choice experiment. Intrinsic and extrinsic model accuracy. Numbers in bold are the highest in the column. * Results from Pezzelle et al. (2020).

4.1 Results

In Table 2, we report the extrinsic and intrinsic accuracy of each tested model. We compare our results with those by humans and the pre-trained LXMERT (Tan and Bansal, 2019) (best-performing in Pezzelle et al., 2020), as they are given in the BD2BB paper. As an additional baseline, we report the results by CLIP (Radford et al., 2021), which we obtain by computing the CLIPScore (Hessel et al., 2021) (quantifying the plain degree of alignment between the visual and textual inputs) between the image and each of the <intention, action> pairs, fed to the model as a sequence. By looking at the numbers in the table, we identify a few key findings, that we summarize below.

BLIP-2 approaches human performance in the extrinsic evaluation The first key finding of our experiment concerns the performance of BLIP-2 in the extrinsic evaluation: the model achieves an average accuracy of 75.7%, i.e., only 3-accuracy points far from human performance. This means that, for more than 3 samples out of 4, the model identifies the correct action for a given <image, intention> pair. This result is even more remarkable considering that the other three models do not fare much better than chance in this evaluation setting. As mentioned in Section 3.2, BLIP-2 is the only model of the leaderboard trained with COCO images. Moreover, it is the only one leveraging a language model, FlanT5, which was instruction-finetuned on a mixture of tasks. Therefore, it is reasonable to hypothesize both these aspects could give an advantage to BLIP-2 over the other models.

`model_doc/idefics` (IDEFICS)

[3]Data and code to reproduce our results will be made available at: `https://anonymized/repo/`

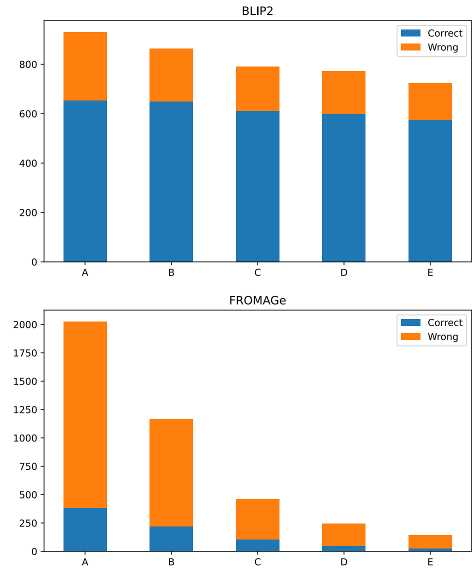

Figure 2: Multiple-choice experiment. Distribution of correct and wrong answers by BLIP-2 (top) and FROMAGe (bottom) against their position (A-E) in the template. While BLIP-2 has only a minor bias toward first-position answers, FROMAGe is heavily biased.

	BLIP-2	Humans*
multimodal	75.7±0.8	79.0
language-only	59.1±0.4	50.0
vision-only	57.0±2.5	72.3

Table 3: BLIP-2 and human accuracy in three settings: multimodal, language-only, and vision-only, evaluated extrinsically. *From Pezzelle et al. (2020).

Some VLMs are biased towards early-presented options Upon manual inspection of the model-generated outputs in the extrinsic evaluation, we noticed a bias of MAPL, FROMAGe, and IDEFICS toward predicting the actions presented earlier in the template; that is, these models appeared to prefer A over E. To quantify this effect, we calculated, for each model, the percentage of predicted responses based on their position. In Figure 2, we visualize the results for FROMAGe (MAPL and IDEFICS exhibit a very similar pattern), which we plot against the behavior of BLIP-2. As can be seen, FROMAGe is heavily biased toward the first positions/letters in the template, while BLIP-2 is not, or to a much lesser extent. This striking difference highlights that, while BLIP-2 can treat each action in the template (almost) equally, this is not the case for the other models. This is likely one of the reasons for the success of this model.

VLMs do not overtly outperform LXMERT in the intrinsic evaluation When evaluated intrinsically on the task, generative VLMs do not exhibit a generalized advantage over the previous-generation models. While MAPL and IDEFICS do perform slightly better than LXMERT (see Table 2), this is not the case for FROMAGe and BLIP-2 (note, though, that in an additional experiment, we found that BLIP-2 with underlying OPT achieves better accuracy: 62.4%). This suggests that generative VLMs may not, by default, be necessarily better encoders than previous models, in line with what was discussed by BehnamGhader et al. (2024) for text-only LMs. At the same time, all VLMs except FROMAGe outperform CLIP, which reveals that the cross-modal scores we obtain from them encode more than simple image-text alignment, which is all that CLIP captures. This provides indirect proof that VLMs can, to some extent, combine complementary information from the two modalities.

4.2 Is BLIP-2 Using the Multimodal Context?

As discussed above, BLIP-2 achieves near-human accuracy in the multiple-choice experiment when evaluated extrinsically. In this analysis, we explore whether this performance is due to genuine integration of language and vision or biases and shortcuts exploited in one of the two modalities. To do so, we run the same experiment in two additional settings: (1) a language-only one, where we provide the model with the intention and the actions, but not the image; (2) a vision-only one, where we provide the model with the image and the actions, but not the intention(See the prompts in Appendix(**?**)). If the model genuinely leverages the two modalities, it should perform worse in both these settings than the multimodal one, where both the image and the intention are given as input. The results of this analysis are presented in Table 3.

As can be seen, the model fed with the multimodal input neatly outperforms both unimodal settings. This reveals that jointly leveraging information conveyed by the image and the intention is beneficial to solving the task, a pattern that is also observed in human behavior. Compared to humans, however, BLIP-2 exhibits a slight advantage in the language-only setting and a large disadvantage in the vision-only setting. This pattern suggests, on the one hand, that the underlying FlanT5 language model might be driven by some biases and default choices when performing the inference task; on the other hand, its image processor is less capable than humans to understand the subtleties of a scene and which actions it pragmatically licenses.

In Appendix D, we present the results of an additional analysis that further investigates whether, and when, the model leverages complementary information or simply counts on a single modality.

5 Open-Ended Generative Experiment

In the multiple-choice experiment, only BLIP-2, but none of the other models, is extrinsically good. At the same time, most VLMs can assign a higher probability to the correct action in many cases. This discrepancy is likely due more to how the different models have been trained and designed than to what the models do or do not know. Moreover, we acknowledge that a multiple-choice scenario is not the most naturalistic way to interrogate these models. To overcome these issues, in the second experiment, we feed the VLMs with the image and the intention and let them generate an open-ended continuation. This is a more straightforward way to assess the models, but it poses challenges on the evaluation side. Below, we describe the two methods we use to evaluate model performance.

Reference-based evaluation In this evaluation, we take the continuation generated by a model and compare it to each of the five candidate actions in the sample. We make the simplistic assumption that, if the generated action is good, it should be more similar to the correct action than the decoy actions. This assumption allows us to compute model accuracy: we consider the model correct every time the similarity between the generated and correct actions is the highest in the batch.

Intuitively, the choice of the prompt to use to elicit a continuation from a model plays a big role. Indeed, we noticed that some prompts may be effective for some models, but not for others. After a careful, manual exploration of prompts, we focused on four that appeared to be good-performing across models. We provide further details about this exploration and the actual prompts in Appendix **??**.

To compute similarities, we used various common NLG metrics, including BLEU4 (Papineni et al., 2002), ROUGE (Lin, 2004), CIDER (Vedantam et al., 2015), Meteor (Banerjee and Lavie, 2005), and the more recent BERTScore (Zhang et al., 2019). While the scores by various metrics can be different, we observed that various metrics led to similar patterns. Therefore, from now on, we only focus on BERTScore and refer the reader to Appendix E for further details on other metrics.

model	accuracy
MAPL	32.9±8.7
FROMAGe	32.7±4.8
BLIP-2	**49.5±2.6**
IDEFICS	31.5±10.9

Table 4: Open-ended generative experiment. Reference-based accuracy is computed using BERTScore similarity. Average and std. over results for 4 different prompts.

Reference-free evaluation Evaluating model outputs using automatic, reference-based metrics is simplistic as it assumes that only an action that is similar to the target one is a good one. To evaluate the plausibility of the actions in a reference-free manner, we therefore carried out a human evaluation. We sampled 50 <image, intention, generated action> datapoints per model and presented them, one at a time, to six participants.[4] We asked them to judge whether the second part of the sentence (displayed in bold), i.e., the generated action, was a plausible continuation of the first part, i.e., the ground-truth intention, based on the contents of the image. As the question was binary, they could choose between the options *Yes* or *No*. To ensure the quality of human annotations, we added 20 clear-cut cases to the data (10 correct, 10 wrong), that we used as a control group. All participants achieved high accuracy ($\geq 75\%$) on these control samples. In total, each participant assessed 220 samples (200 model-generated + 20 control ones).

5.1 Results

Table 4 and Figure 3 report, respectively, the results of the reference-based and reference-free evaluation. Below, we summarize the main findings.

BLIP2 is the best-performing model according to both evaluations Based on the results of both evaluations, BLIP-2 appears to be the best-performing model in this experiment. Indeed, this model achieves the highest average reference-based accuracy (49.5%) across the board, outperforming the other models by nearly 20 accuracy points. As for the reference-free evaluation, human participants judge BLIP-2's generated actions as plau-

[4]Participants were recruited among colleagues at our institution and carried out the annotation voluntarily. They were informed about the use of the annotations they provided and agreed to their use through informed consent.

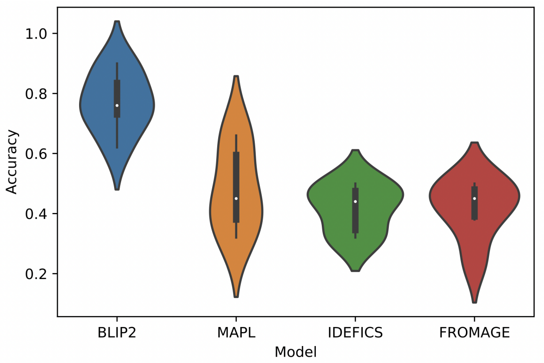

Figure 3: Open-ended generative experiment. Reference-free accuracy is based on human judgments, 300 per model (i.e., one per assessed sample).

sible in 77% cases. This is a remarkably higher accuracy than the one obtained by the other models, whose accuracy ranges between 40 and 45%. These results confirm the superiority of BLIP-2 in generating actions consistent with both a visual context and a non-grounded textual intention.

BLIP-2's abilities can also be appreciated by looking at cases where it generates actions that are judged implausible by human annotators, as the one in Figure 4. Here, given the intention *If I want to socialize*, the model generates a good action, which is also consistent with the scene content—a pool in the foreground and several people standing around it. However, in this case, this action is *pragmatically* implausible, as the people in the image are busy playing video games. From this single example, it appears that the strengths of BLIP-2 lie in its ability to understand the scene, the intention, and their complex interaction. On the other hand, there is room for improvement in understanding the dynamics of events and relationships between people conveyed by an image. Improving this aspect can be a good direction to develop semantically valid and pragmatically plausible models.

Other models perform similarly (poorly) As for MAPL, IDEFICS, and FROMAGe, it can be noted that their performance is similar according to both evaluations. This is interesting as the models build on different language and vision models, have varying sizes, and are trained with different data. Once again, this observation seems to reiterate the peculiarity of BLIP-2 compared to other architectures, from which it differs by the instruction-tuned LM and the presence of COCO in the training data.

If I want to socialize…

Ground-truth
I will play the Wii with my friends

BLIP-2
I will play pool with the guys ✗

Figure 4: An example of an action generated by BLIP-2. In this case, the human annotators considered this action implausible given the intention and the image.

6 Conclusion

In this work, we focused on the problem of combining complementary information brought to a context by language and vision. We used a benchmark proposed for previous-generation multimodal models, i.e., language-and-vision encoders based on the Masked Language Modeling objective, and tested, for the first time, how state-of-the-art generative visual language models deal with it. Through two experiments, we found that the BLIP-2 performs consistently and significantly better than competing models. While most generative VLMs struggle, this model achieves both near-human accuracy in the multiple-choice experiment and high human judgments in the open-ended generative experiment. This reveals the superiority of this model on the task, likely due to instruction finetuning and having seen COCO images in training. These two ingredients appear to be key for the model, which exhibits a deep understanding of the image, the textual intention, and the complex interaction between them. Based on these findings, we conjecture that this recipe—and, particularly, instruction finetuning—may help models develop better generalized semantic and pragmatic abilities. These skills are crucial to language-mediated communication; future work might extend our investigation to other scenarios, including more naturalistic ones.

We argue that future work should focus on building more datasets and resources that encompass complex interactions between image content and its accompanying text. This implies taking a more communicative perspective on the study of language in multimodal contexts, which is what is needed to develop linguistic technologies ready to communicate seamlessly with human users.

Limitations

Our investigation is limited to one (English) dataset and a handful of models. This narrows the scope of the findings we presented. While our approach can be easily applied to other resources, languages, and models, we acknowledge that the claims made in this paper may not necessarily generalize. Another limitation is the choice of prompts used to elicit the responses from the models. There is growing evidence of the significant role of prompt wording on model generation, that we fully recognize. Although we believe we conducted a fairly comprehensive prompt search, our results can only speak for the prompts we used. Furthermore, the human evaluation we conducted is arguably small-scale as it involves few participants and a relatively small number of samples. We cannot fully exclude that the reported patterns may not replicate when increasing the number of participants and stimuli.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

C. Benoît, J. C. Martin, C. Pelachaud, L. Schomaker, and B. Suhm. 2000. Audio-visual and multimodal speech-based systems. pages 102–203.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The (r)evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Katherine Deng, Arijit Ray, Reuben Tan, Saadia Gabriel, Bryan A. Plummer, and Kate Saenko. 2023. Socratis: Are large multimodal models emotionally aware?

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

C. Forceville. 2020. Introduction. *Visual and Multimodal Communication*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*.

Elisa Kreiss, Fei Fang, Noah D Goodman, and Christopher Potts. 2021. Concadia: Towards image-based text generation with a purpose. *arXiv preprint arXiv:2104.08376*.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv e-prints*, pages arXiv–1908.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2022. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*.

Nasrin Mostafazadeh, Chris Brockett, William B Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sarah Partan and Peter Marler. 1999. Communication goes multimodal. *Science*, 283(5406):1272–1273.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12098–12112, Toronto, Canada. Association for Computational Linguistics.

Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. Be different to be better! a benchmark to leverage the complementarity of language and vision. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2751–2767.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Aditya Surikuchi, Sandro Pezzelle, and Raquel Fernández. 2023. GROOViST: A metric for grounding objects in visual storytelling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3331–3339, Singapore. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv e-prints*, pages arXiv–2205.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Different Parameter Combinations for Multiple-Choice Experiment

We tested each model's most promising prompt with 4 different parameter combinations. The result shows in below table 5.

-	FROMAGe	BLIP2	IDEFICS
Default parameter	20.5	77.1	42.1
Temp = 0.8,topp = 0.8	19.7	75.9	33.1
Temp = 0.8,topp = 0.9	17.8	74.3	30.0
Temp = 0.8,topp = 0.8	17.7	75.2	30.5
Temp = 0.8,topp = 0.8	17.7	73.5	28.3

Table 5: The result for 3 generative models in discriminative setting across different parameter combinations

B 30 Prompts for Multipul-Choice Experiments

1. What am I supposed to do <intention>? <candidate actions>
2. What am I supposed to do <intention>? Pick the best option from the following ones: <candidate actions>
3. Question: What am I supposed to do <intention>? Pick the best option from the following ones: <candidate actions> Answer:
4. Given this visual context, what am I supposed to do <intention>? Pick the best option from the following ones: <candidate actions>
5. What am I supposed to do given this visual context, <intention>? Pick the best option from the following ones: <candidate actions>
6. Suppose I am faced with this situation. What am I supposed to do <intention>? Pick the best option from the following ones: <candidate actions>
7. What action will I perform <intention>? Pick the best option from the following ones: <candidate actions>
8. Suppose I am faced with this situation. What actions should I perform <intention>? Pick the best option from the following ones: <candidate actions>
9. Imagine I am in this visual scenario. What action should I perform <intention>? Pick the best option from the following ones: <candidate actions>
10. What actions should I perform <intention>? Considering both my feeling and the image, pick the best option from the following ones: <candidate actions>
11. Considering both my feeling and the image, what will I do <intention>? Pick the best option from the following ones: <candidate actions>
12. What will I do <intention>? Pick the best option from the following ones: <candidate actions>
13. <intention>, what's the best step to take? Select the optimal choice from the options provided.<candidate actions>
14. <intention>, what's the most suitable action to choose? Select the best option from the following list.<candidate actions>
15. <intention>, what's the recommended action to perform? Choose the best option from the following ones.<candidate actions>
16. Which action is most advisable <intention>? Choose from the following options. <candidate actions>
17. What's the best step to follow <intention>? Pick from the following options. <candidate actions>
18. <intention>, what's the most appropriate decision to make? Select the right choice from the options provided. <candidate actions>
19. What's the recommended response <intention>? Choose from the following options. <candidate actions>
20. What should I do <intention>? Choose from the following options considering both my feeling and the image.<candidate actions>
21. Taking into account both my feeling and the image, what's the recommended step to take <intention>? Pick from the following options. <candidate actions>
22. Considering equally the image and my feeling, what should be my action <intention>? Choose the right action from the following list.<candidate actions>

23. With both the image and my feeling in mind, how should I react <intention>? Select the best choice from the following options. <candidate actions>

24. With both the image and my feeling in mind, what action is most advisable <intention>? Choose from the following options. <candidate actions>

25. When taking both the image and my feeling into consideration, what's the best step to follow <intention>? Pick from the following options.

26. <intention>. <candidate actions>

27. <intention>, what happens next? <candidate actions>

28. <intention>, what will I do? <candidate actions>

29. What will I do <intention>? <candidate actions>

30. How should I behave <intention>? <candidate actions>

C 4 Prompts For Open-Ended Generative Experiment

1. What am I supposed to do <intention>? I will...
2. <intention>, what will I do? I will...
3. What should I do <intention>? I will...
4. What action will I perform <intention>? I will...

For this open-ended generative experiment, it is important to note that the optimal prompt may vary for each model. For instance, the prompt "What am I supposed to do + [intention]+? " can yield results for BLIP2 but did not work well for the MAPL and FROMAGe models. For the MAPL model, "Question:... Answer:", and for the FROMAGe model, "Q:... \nA: " are the template prompts provided by the model developer. Additionally, adding "I will" at the end of the prompt is proved to be effective for both models. After a careful manual inspection of several prompts and their outputs, we focused on the 4 most promising ones as in this appendix.

Actions generated using these prompts also need to be further processed to ensure they conform to the same format as the target action and other optional actions. For example, IDEFICS consistently generates sentences prefixed with "Assistant:". To calculate the similarity score of these answers with other actions, it is necessary to remove the "Assistant:" prefix and retain only the main action, which typically begins with a verb.

D Error Analysis

We performed an error analysis aiming to compare the outputs of the three versions of BLIP2: multimodal, language-only, and vision-only. By doing so, we aimed to gain insights into how, and when, BLIP2 effectively leveraged information from language and vision to achieve better performance in the task. We observed that, in 1,350 cases (33%), all three model versions provided a true prediction. In such cases, the model could make a correct assessment by relying only on one single modality, which suggests that, in these cases, the information conveyed by the multimodal input may be redundant.

In 221 cases (around 5%), only the multimodal BLIP2 could correctly predict the right answer, while no unimodal model versions could. In these cases, BLIP2 genuinely leveraged complementary information from the two modalities, which was necessary but not sufficient on their own to perform the task.

The entire test dataset, comprising 4,081 samples, was categorized into eight different groups based on the consensus of model predictions under three conditions. The categories are as follows:

- TTT: The model correctly produces the answer in LV, L, and V.
- TTF: The model correctly produces the answer in LV, L, but not in V.
- ...and so on for the remaining categories.

For each category, a manual inspection of 100 cases was conducted to identify the sources of errors in the models. The results of this analysis are summarized in Table 6.

This error analysis table reveals a wealth of information. The second and third rows of the table indicate that when there is correct information in one modality, the multimodal model knows how to utilize it effectively. Furthermore, the examples in the fourth row demonstrate that these cases can only be predicted correctly using complementary information.

Is the prediction correct?	Number of Cases	Percentage	Comments
BLIP_LV: T BLIP_V: T BLIP_L: T	1350	0.3308	No errors were found in these cases, indicating that they may be too easy for the multimodality model to handle.
BLIP_LV: T BLIP_V: T BLIP_L: F	581	0.1424	The model in the L setting gave incorrect predictions due to the absence of image information.
BLIP_LV: T BLIP_V: F BLIP_L: T	808	0.1980	The model in the V setting gave incorrect predictions due to the absence of intention information.
BLIP_LV: T BLIP_V: F BLIP_L: F	222	0.0544	Only multimodality setting can give true predictions.
BLIP_LV: F BLIP_V: T BLIP_L: T	11	0.0027	The model's incorrect predictions can be attributed to the following reasons: 1. Problematic/borderline cases; 2. Wrong object detection; 3. Failure to understand the intention; 4.Only considering one modality;
BLIP_LV: F BLIP_V: T BLIP_L: F	221	0.0542	
BLIP_LV: F BLIP_V: F BLIP_L: T	117	0.0287	
BLIP_LV: F BLIP_V: F BLIP_L: F	771	0.1889	

Table 6: Error Analysis Table: Each row provides information on some specific cases, indicating whether the BLIP2 model can produce a correct prediction under three different conditions and the potential reasons for such results.

E Different Metrics to Calculate Similarity

We tested different metrics to conduct the Reference-based evaluation for the open-ended generative experiment. We tested in three settings: multimodal, language-only, and vision-only.

F Degree of Visual Grounding

In our previous analysis, we evaluated the BLIP2 model's performance in the BD2BB task by examining the accuracy of the generated actions. However, accuracy alone does not fully capture the model's ability to utilize the information from two modalities. Therefore, we can also evaluate the model from a different perspective by considering its ability to incorporate information only from the image. We assumed that if the model successfully utilizes the image information, it will explicitly mention objects from the image in the generated actions. This indicates that the action is grounded in the visual content.

Thanks to the labeling of golden nouns in the image data, we can easily determine whether the generated action mentions any objects from the image. Based on how many actions are grounded in the visual content, we can calculate the grounding rate by following the formula:

$$\text{grounding\_rate} = \frac{N_{\text{grounded}}}{N_{\text{total}}} \tag{1}$$

We calculated the grounding rate for generated actions using 15 manually selected prompts. These prompts were carefully crafted to vary in their focus: some directed the model's attention toward language aspects, others toward visual elements, and some involved variations in linguistic forms. The prompts we use are shown in Table 8. The grounding rate varied across different prompts. Interestingly, we found that by changing the prompt, we could easily influence the grounding rate while accuracy remain stable. Although we cannot suppress a modality by altering the prompt (prompt 6), we can effectively focus selectively on one modality by being explicit (prompts 7, 8, 9, and 14). Figure 5 is the bar plot about both accuracy and grounding rate. The pink bar represents the accuracy of different prompts in an ascending order. It

Setting	BERTScore	BLEU-4	CIDER	METEOR	ROUGE
LV	0.53	0.54	0.52	0.48	0.51
L	0.39	0.49	0.38	0.29	0.36
V	0.42	0.37	0.41	0.38	0.4

Table 7: Results for BLIP2 in three different conditions(row) using different similarity scores(column).

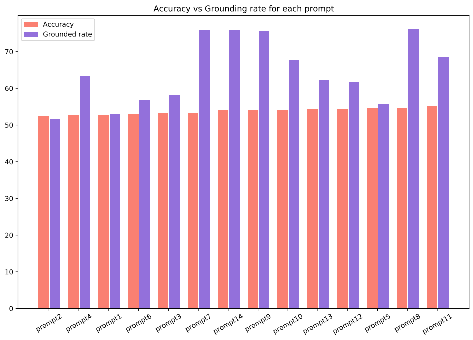

Figure 5: A bar plot illustrates the accuracy and grounding rate across different prompts. Accuracy increases from left to right. The bar with the highest grounding rate corresponds to Prompt 8: "Suppose I am in this specific visual context. What actions should I take [intention]? Please mention at least one object from the image." The bar with the lowest grounding rate corresponds to Prompt 2: "Q: What am I supposed to do [intention]? A:" This variation highlights the impact of prompt design on model performance.

is shown that the grounding rate does not always align with the accuracy.

By introducing the grounding rate as an evaluation metric, we provide an additional perspective to assess the performance of multimodal models. It highlights the model's ability to ground its generated actions in visual information, offering insights beyond accuracy alone (Surikuchi et al., 2023).

	Prompt	Accuracy	Grounding rate
prompt1	Question: What am I supposed to do + [intention]? Answer:	52.66	53.10
prompt2	Q: What am I supposed to do + [intention]? \nA:	52.34	51.58
prompt3	What am I supposed to do + [intention]?	53.20	58.25
prompt4	What am I supposed to do + [intention]+? I will	52.63	63.44
prompt5	What am I supposed to do + [intention]+? Answer in the format "I will"	54.57	55.67
prompt6	What am I supposed to do + [intention]+? Please provide an answer based solely on the intention, without considering the image.	53.08	56.82
prompt7	What action should I take +[intention]+? Please base your response solely on the image. Additionally, kindly mention at least one object visible in the image.	53.32	75.97
prompt8	Suppose I am in this specific visual context. What actions should I take+ [intention]+? Please mention at least one object from the image.	54.76	76.06
prompt9	Imagine I am in the given visual scenario. What actions should I take regarding +[intention]+'? Please mention at least one object from the image.	54.06	75.74
prompt10	Imagine yourself in this specific visual context. Considering both the intention and the image, what actions should be taken +[intention]+?	54.06	67.78
prompt11	Considering both the intention and the image, what will you do +[intention]+?	55.16	68.41
prompt12	What will I do +[intention]+?	54.47	61.67
prompt13	What will you do +[intention]+? I will	54.37	62.23
prompt14	What will you do +[intention]+? Please give a plausible reason by mentioning at least one object from the image.	53.96	75.89

Table 8: The accuracy and grounding rate across different variations of the prompt.