Accountable Textual-Visual Chat Learns to Reject Human Instructions in Image Re-creation

Anonymous authors
Paper under double-blind review

Abstract

The recent success of ChatGPT and GPT-4 has drawn widespread attention to multimodal dialogue systems. However, the academia community lacks a dataset that can validate the multimodal generation capabilities of Visual Language Models (VLMs) in textual-visual chat tasks. In this paper, we construct two new multimodal datasets: the synthetic CLEVR-ATVC dataset (620K) and the manually pictured Fruit-ATVC dataset (50K), both featuring visual and text-based inputs and outputs. Additionally, to enable the multimodal system to reject human requests (i.e., demonstrate accountability), as in language-based ChatGPT conversations, we develop and incorporate specific rules into the datasets as supervisory signals. This allows the trained VLM to provide a yes or no answer after visual and textual reasoning, accompanied by a language explanation as to why the human instruction cannot be excuted. In our method, we propose a two-state training procedure to train the image auto-encoder and auto-regressive transformer from scratch. The first state involves a discrete variational autoencoder (dVAE) to compress each image into short tokens, which are then concatenated with text tokens as a single data stream to be fed into the decoder-based transformer for generating visual re-creation and textual feedback in the second state. We provide comprehensive analyses of experimental results in terms of re-created image quality, answer accuracy, and the model behavior when faced with uncertainty and imperfect user queries. We hope our explorations and findings contribute valuable insights regarding the accountability of textual-visual generative models.

1 Introduction

Recently, the most important breakthrough was made by ChatGPT OpenAI (2023a) and GPT-4 OpenAI (2023b), which unveiled the emerging potential of the conversation between human and artificial intelligence system. ChatGPT is served as a chatbot with language as both input and output, while GPT-4 is a multi-modal model that accepts both image and text inputs and produces text outputs. A successful multimodal generative model should be able to perform both textual and visual reasoning and generate high-quality text and image feedback. Visual ChatGPT Chenfei Wu & Duan (2023) is a pioneering work that combines ChatGPT and a series of pre-trained visual foundation models to enable text-iamge chat. FROMAGe Jing Yu Koh (2023) is also involves image-text inputs and outputs to achieve multimodal dialogue, in which a few linear layers are finetuned and the pre-trained LLM is frozen. However, previous datasets lack unique and definite ground truths to measure the quality of text and images generated in multimodal dialogue systems. Therefore, there is an pressing need for a dataset to evaluate the performance of multimodal generative models. In this paper, we would like to build new multimodal datasets that require models to output high-quality images and provide textual feedback while accepting a text-image pair. We construct one synthetic CLEVR-ATVC (620K) dataset and one real Fruit-ATVC (50K) dataset. The synthetic dataset is newly rendered using about 200 GPU days, whereas the real Fruit-ATVC dataset is manually pictured and annotated.

The other important contribution of this paper is that we consider the issue of responsibility of text-to-image generation models, specifically the need for models to learn to reject human instructions. Prior works Doshi-Velez et al. (2017); Loi & Spielkamp (2021) have discussed the importance of accountability in AI

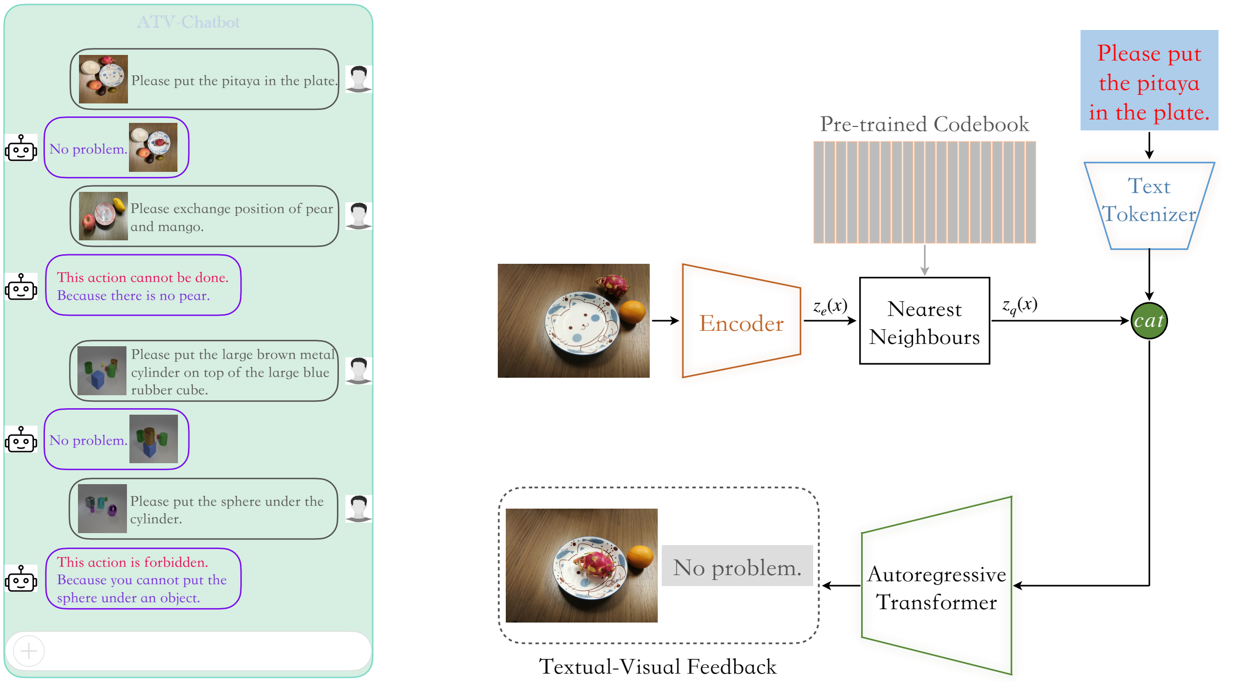


Figure 1: The left figure shows that a multimodal generative model can simultaneously recreate visual input and give textual feedback when faced with human instructions, especially it can reject some commands. The right figure illustrates the overall framework of our method. The model is required to generate a re-created image (**M**) and a textual feedback (**A**) conditioned on the visual input (**V**) and text-based user query (**T**), and the language-based explanation is also given for those instructions that cannot be executed and the prohibited instructions.

systems, including the ability to hold decision-makers responsible for adhering to procedural and substantive standards. The question of how to responsibly manage and deploy AI systems is an important and under-appreciated topic. In fact, ChatGPT OpenAI (2023a) implements a similar requirement at deployment time, retrieving rules set by OpenAI to limit certain requests. However, as far as we know, we are the first to consider the problem of responsibility in the textual-visual dialogue models. Previous text-to-image generation methods Ramesh et al. (2021a); Nichol et al. (2021); Wu et al. (2021); Ramesh et al. (2022); Saharia et al. (2022); Ding et al. (2021) have produced random and uncontrolled outcomes that lack human oversight. Although prior works have considered the controllability of image generation, they mainly focused on controlling the object categories or attributes (e.g., position, pose, color, etc.) Li et al. (2019); Niemeyer & Geiger (2021); Jiang et al. (2022); Patashnik et al. (2021). For example, StyleCLIP Patashnik et al. (2021) can manipulate the attributes of the visual images by using text. Therefore, in this paper, we hope the machine can reject human instructions and give explanations for its decisions to improve the accountability, especially those instructions that cannot be executed or prohibited instructions.

In order to serve the two research goals we discussed above, we propose a novel task, accountable text-based visual re-creation, to explore whether textual-visual chat can learn to reject human instructions, which requires the model to produce both visual re-creation and language-based feedback while accepting a text-image pair. Our datasets include rules as supervision signals to teach the model to reject certain instructions. As shown in Figure 1, when a chatbot receives instructions from a human, it will say "no" to those instructions that are forbidden or cannot be completed and explain why. The instruction that cannot be executed is because there are no corresponding objects mentioned in the text-based query in the visual input. We provide the details of prohibited instructions in Tabel 3. In our method, as shown in Figure 1, we use a multimodal generative model that consists of an image auto-encoder and an auto-regressive transformer, and propose a two-state training procedure to train them from scratch. The former one is a discrete variational autoencoder (dVAE) to compress each image into short tokens, then the text tokens are concatenated to be fed into the latter one. The decoder-based transformer is used to generate the visual re-creation and textual feedback.

In addition to the above features, we provide comprehensive analysis of our experimental results, including re-created image quality, answer accuracy, and the model's behavior when faced with uncertainty and imperfect user queries. We also compare two different image auto-encoders (VQVAE Oord et al. (2017) and VQGAN Esser et al. (2021)) and analyze the reasons for poor performance of VQGAN. All the datasets are publicly available for non-commercial use, and we will release the source code for our annotation tool, the evaluation tool, the implementation of baseline models, metric calculations, and thorough instructions.

In summary, the main contributions of this paper are summarized as follows:

- We construct two multimodal datsets, CLEVR-ATVC (620K) dataset and one Fruit-ATVC (50K), that include visual and textual inputs and outputs to validate the performance of multimodal generative models.

- We consider the accountability issue of multimodal generative models, and some pre-set rules are embedded in our datasets as supervised signals.

- We propose a two-state training procedure to train the image auto-encoder and auto-regressive transformer. The models are expected to learn to reject human instructions after trained on our datasets.

- We provide extensive qualitative and quantitative results for the generated image quality and answer accuracy. We also evaluate the ability of the model to handle re-creation uncertainty and incomplete queries.

The remainder of this paper proceeds as follows. We first present our dataset information in Section 2. Section 3 then introduces our models and training process. To validate our method, we construct extensive experiments and provide detailed analysis in Section 4. We introduce the related work in Section 5. Section 6 concludes the paper. In the appendix, we provide additional details of our datasets.

2 Datasets

In this section, we first introduce the difference between our datasets and other existing data. Then, the collecting process of datasets is presented. Next, we show the information contained in the annotation files for our datasets, and more details can be found in the appendix. Finally, the statistics of the datasets are summarized in the tables.

2.1 Definition of ATVC

Our proposed dataset and task differ from previous methods in two main ways. Firstly, our multimodal dataset consists of image-text inputs and outputs, which can help us design better multimodal generative models. Secondly, our datasets include preset rules that enable the model to learn to reject human instructions, i.e., the model needs to know which kinds of actions are forbidden even if they can technically be done. The resulting accountable text-based visual re-creation (ATVC) task can be defined as follows.

Given a Visual input (**V**) and the Text-based query (**T**), the multimodal generative model is required to output a re-created image (**M**) and provide language-based explanation (**A**) for its decisions. A successful ATVC model is expected to learn entailment of the action given by the user-query and able to reason language cues to the visual input, as well as the appropriate feedbacks of the visual results it has achieved.

Table 1: The summary information of CLEVR-ATVC dataset.

CLEVR-ATVC				
Size	Train Pairs	Test size	Textual-Visual Feedback	Data type
68 GB	619,741	5,000	Yes	Synthetic

Please <action1> {size color material} [object] <action1> {size color material} [object].
Please <action2> of {size color material} [object] and {size color material} [object].

Action1	Action2	Size	Color	Material	Object
put on top	exchange color	big	red, green, blue, cyan	metal	cylinder
put under	exchange position	small	purple, brown, yellow	rubber	cube, sphere

Table 2: The summary information of Fruit-ATVC dataset.

Fruit-ATVC				
Size	Train Pairs	Test size	Textual-Visual Feedback	Data type
168 GB	27,503	1,872	Yes	Real

Please <action1> [object].
Please <action2> of [object] and [object].
Please <action3> [object] <action3> [container].

Action1	Action2	Action3	Object	Container
remove	exchange position	put in	apple, coconut, lemon etc.	plate, bottle, etc.

2.2 Data Collection

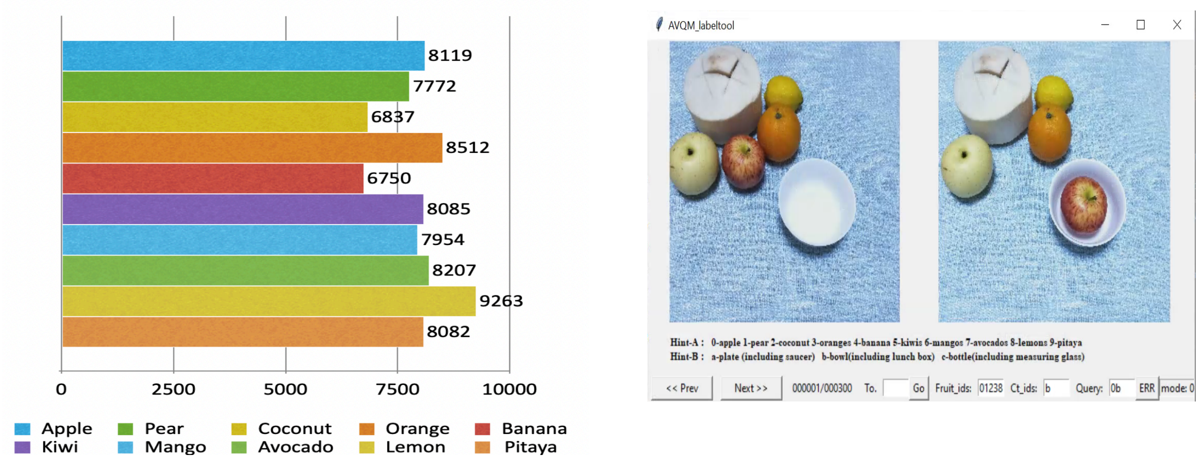
We validate the proposal of this paper on the following two datasets. The first CLEVR-ATVC is a large-scale synthetic dataset for comprehensively evaluating the proposed task, and the Fruit-ATVC is for evaluating our method on real scenarios.

CLEVR-ATVC. We follow the default environmental setting (object settings, rendering options, etc) of CLEVR Liu et al. (2019) to render the synthetic images. For example, the size of rendered images is 256, one image contains 3 to 6 objects. The only difference is that we algorithmically filter out scenes with objects on the border of the image before rendering, which is used to ensure that all the objects in the original images and the re-created images are located within the boundary to facilitate the evaluation of experimental results. The synthetic CLEVR-ATVC dataset is newly rendered by about 200 GPU days.

Fruit-ATVC is a dataset for real scenario, in which all the images are manually pictured by the authors using mobile phones and tripods. The images are taken with different categories of mobile phones, including iphone, huawei, xiaomi, etc. In this dataset, a total of five researchers helped picture the visual inputs and re-created images. In addition, we adopt 12 different scenes to enrich the diversity of this dataset. The figures provided in the appendix show the diversity of collection scenes.

2.3 Generation of Textual Query and Feedback

On CLEVR-ATVC dataset, all visual inputs, text-based queries, re-created images, and textual feedbacks are automatically generated simultaneously by the program we designed. After this program, the visual inputs and re-created images are saved to the corresponding folders, while text-based queries, textual feedbacks and other data information are also automatically saved to a annotation file. For one visual input, we generate ten different text-based queries and corresponding textual feedbacks. For example, six human instructions can be executed by the AI system and re-created the images accordingly, two instructions cannot be executed and two instructions are forbidden. The instructions that cannot be executed are because we insert objects to be re-created in the text-based queries that are not in the visual input, while the prohibited instructions are manually specified.



Mention frequency

Annotation interface

Figure 2: The mention frequency of different fruits and the annotation interface for Fruit-ATVC dataset.

Table 3: The detailed information of textual feedback.

Textual Feedback		
Answer Type	Reason for prohibition	Explanation
1) No problem.	/	/
2) Cannot be done.	/	no mentioned objects
3) Forbidden.	cannot put certain object under an object	no mentioned objects
4) Forbidden.	cannot put an object on top of certain object	no mentioned objects

The generation of text-based query and textual feedback on Fruit-ATVC dataset is different with the above process, because the visual inputs and re-created images need to be manually pictured, and the text-based queries and textual feedbacks are also generated semi-automatically through our designed labeling tool. Therefore, we first collect and organize the visual inputs and recreated images according to the rules we have established, such as a visual input has a query that can be executed and a query that cannot be executed. Next, the annotation interface illustrated in Figure 2 is used to help us generate text-based queries and textual feedbacks and automatically save them to an annotation file.

2.4 Data Annotation

In this section, we introduce the main format of the data annotation file. We follow the COCO dataset Lin et al. (2014) to construct our annotation file. The main format shown in Figure 2.4 includes four parts: text-based query (T), visual input (V), recreated image (M) and textual feedback (A). Since each visual input corresponds to multiple text-based queries, we add "idx" to T, M and A to map them one by one. The subkey of "objects" lists which objects are contained in this visual input. The "explanation" here includes the reasons for the prohibition and the explanation of which objects do not exist. We show more details of textual feedback in Table 3. There are three types of answers: a) no problem; b) action cannot be done; and c) action is forbidden. There are two prohibited instructions: One is put certain object under an object, the other is put an object on top of certain object.

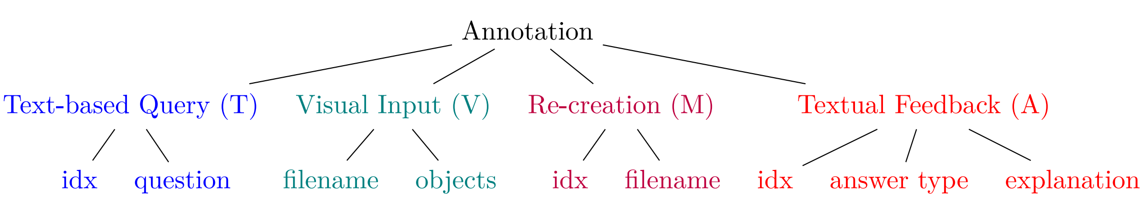


Figure 3: The main format visualization of data annotation file. The details can be found in the appendix.

2.5 Data Statistics

The summarized information of our constructed datasets is illustrated in Table 1 and Table 2.

CLEVR-ATVC includes a total of 620K pairs for training, and the testing set has 500 visual inputs and 10 queries per image, resulting in total 5000 pairs. In this dataset, we incorporate four action types, including "exchange color", "exchange position", "put under", and "put on top". The rendered images include 3 object shapes, 7 object colors and 2 object materials, etc. Considering the different attributes of the objects mentioned above, there will be hundreds of millions of combinations. The pattern of the text-based query includes two different types: The first template is for "put on top" and "put under" actions, and the second template is for "exchange color" and "exchange position" actions.

Fruit-ATVC includes 27,503 training pairs for training, and the testing set has 1872 pairs. There are ten kinds of fruits involved in this dataset, including apple, pear, coconut, orange, banana, kiwi, mango, avocado, lemon, and pitaya. The mention frequency of different fruits is shown in Figure 2. This dataset includes 14 types of containers. As for queries, we have three types of actions, including "put a fruit in a container", "exchange positions of the fruits", and "remove the container".

3 Method

In our method, the generative model includes an image auto-encoder and an auto-regressive transformer Vaswani et al. (2017) to predict tokens of re-created image and textual feedback. The framework of our method is shown in Figure 1. Our method is a two-stage training procedure, similar to Ramesh et al. (2021a).

Image Auto-encoder This stage forces the model to learn the sequential discretized latent space of high dimensional images to utilize the expressive transformer architecture, *i.e.*, learning a codebook to encode the visual input. Given an image of resolution 256×256, a quantization model encodes it into $32 \times 32 = 1024$ discretized latent codes where the codebook size is 8192. We have tried two alternatives, including Vector Quantized Variational Autoencoder (VQVAE) Oord et al. (2017) and VQGAN Esser et al. (2021). As illustrated in Figure 1, the VQVAE can learn a latent space matrix with discrete learnable variables, which is trained end-to-end via straight-through estimation. The codebook and the VQVAE model can be trained end-to-end using the following objective:

$$\begin{aligned}\mathcal{L}_{VQ}(G,C,E) = &\underbrace{\|sg[E(v)]-e\|_2^2}_{\mathcal{L}_{codebook}} + \underbrace{\beta\,\|sg[e]-E(v)\|_2^2}_{\mathcal{L}_{commit}} \\ &+ \underbrace{\|v-G(e)\|_2^2}_{\mathcal{L}_{rec}},\end{aligned} \tag{1}$$

where v is the visual input, e is the embedding for learning the codebook-indices, C is the learned codebook, E is the image encoder, $G(e)$ is the decoder for reconstruction. sg is the stop-gradient operation, and β is a hyper-parameter to balance the commitment loss Oord et al. (2017). The quantized codebook index is determined by looking up the nearest codebook vector of input features $E(v)$ in terms of the Euclidean distance. The VQGAN Esser et al. (2021) is a variant of VQVAE, which adds a perceptual loss and an adversarial loss produced by patch-based discriminator. It can achieve high resolution while significantly reducing the sequence length. The complete objective for the compression model Q^* is:

$$\begin{aligned}Q^* = &\arg\min_{G,C,E}\max_{D}\mathbb{E}_{x\sim p(v)}[\mathcal{L}_{VQ}(G,C,E) \\ &+\lambda\mathcal{L}_{GAN}(\{G,C,E\},D)],\end{aligned} \tag{2}$$

where $\mathcal{L}_{GAN}$ is to learn the differences between real and reconstructed images:

$$\mathcal{L}_{GAN}(\{G,C,E\},D) = [\log D(v) + \log(1-D(\hat{v}))]. \tag{3}$$

Auto-regressive Transformer Based on the first stage, the images can be highly compressed by the codebook-indices of their embeddings. In this state, the image encoder E is fixed. Therefore, the $\mathbf{V}$, $\mathbf{T}$, $\mathbf{M}$, and $\mathbf{A}$ can be represented by the embedding $\mathbb{S}$ of the sequence of tokens. We adopt the axial positional embedding $\mathbb{P}$ to process the codebook-indices generating by the image reconstruction module $\mathbb{R}$, which is practically effective for multi-dimensional data. In our implementation, the sequence T_{seq} of the transformer is sequentially formed by $\mathbf{T}$, $\mathbf{V}$, $\mathbf{M}$, and $\mathbf{A}$, which is defined as follows:

$$T_{seq} = \mathrm{Concat}(\mathbb{S}(\mathbb{O}(T)), \mathbb{P}(\mathbb{R}(V)), \mathbb{P}(\mathbb{R}(M)), \mathbb{S}(\mathbb{O}(A))), \tag{4}$$

which focuses to generate the re-created result and textual feedback. $\mathbb{O}$ represents the tokenize operation. The transformer is a pure decoder, where text tokens can be attended by each image token in any one of the self-attention layers. Causal mask and full attention mask are used for the text-to-text and image-to-image, respectively. During training, each token is asked to predict the distribution of the next token, which formulates as an autoregressive next-index prediction. Therefore, it is equivalent to maximize the log-likelihood of the data representations.

$$\mathcal{L}_{Transformer} = \mathbb{E}_{x\sim p(x)}[-\log p(T_{seq})], \tag{5}$$

where p is the full representation of the likelihood of the possible next indices. Note that for the "cannot" and "forbidden" pairs, as there is no re-created ground truth, the loss of the its prediction will not be back-propagated. We have tried to predict a black image or the original visual input for these two cases, but both of them perform much worse than simply ignoring their loss. In the testing procedure, we only need to input the $\mathbf{T}$ and $\mathbf{V}$ for pixel-wise iterative generation until the required length is achieved.

4 Experiment

In this section, we introduce the experimental details, evaluation metrics, and analyse on the quantitative and qualitative results.

4.1 Experimental Details

All experiments are all conducted with Pytorch 1.8.1. The model parameters are updated through Adam Diederik P. Kingma (2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We first trains the image reconstruction module for 200 epochs with an initial learning rate of 0.001 and a decay rate of 0.99 per epoch. This stage takes about half a day to train for 200 epochs. The number of attention heads, the attention key and value dimensions, the number of layers, and the dimensions of the models are set to 8, 64, 4, and 512, respectively. The text tokenizer operates on a lower-cased byte pair encoding (BPE) representation of the text with a 49,408 vocab size Sennrich R & A (2015). The text sequence is bracketed with [SOS] and [EOS] tokens. The maximum length of the text sequence is set to 64, and the output length of the image codebook is 1024, which results in a total 2112 sequence lengths for the transformer model. The second stage is distributively trained over 200 epochs with a fixed learning rate of 0.0003. Unless specified, all the parameters of the architecture for the following experiments will remain the same. The model is trained for approximate 900 GPU days on CLEVR-ATVC dataset and 350 GPU days on Fruit-ATVC dataset.

4.2 Evaluation Metrics

We evaluate the proposed method using the following two types of metrics.

Image Quality Metric. The first two are the Peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM), which are commonly used to measure the similarity and quantify reconstruction quality of images. We also adopt the FSIM metric Zhang et al. (2011), which could be more suitable for the human visual system (HVS) using phase congruency (PC) and gradient magnitude (GM).

Human Evaluation. Following the previous methods Lee et al. (2020); Ding et al. (2021); Dong et al. (2017); Zhang et al. (2017; 2018; 2020); Kayser et al. (2021), we also adopt Human Rank (HR) to precisely reveal the performance. HR can be used to evaluate whether the synthesized image conforms to subjective

Table 4: Qualitative results of CLEVR-ATVC dataset.

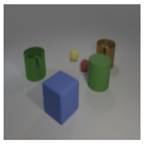
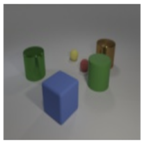
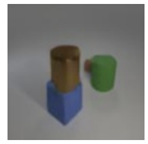
Visual (V)	Text (T)	Re-created (M)	Answer (A)	**FM-Score**
	Please exchange the positions of the large brown metal cylinder and the large blue rubber cube.		No problem.	**1.0**
	Please put the large brown metal cylinder on top of the large blue rubber cube.		No problem.	**0.75**
	Please put the small gray rubber cylinder on top of the large yellow metal cube.	/	This action cannot be done. Because there is no large yellow metal cube.	**1.0**
	Please put the small gray metal sphere under the small purple rubber cylinder.	/	This action is forbidden. Because you cannot put the sphere under an object, and there is no small gray metal sphere and no small purple rubber cylinder.	**1.0**

Table 5: Image re-creation evaluation on the CLEVR-ATVC dataset.

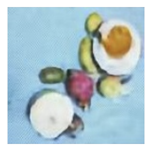
PSNR	SSIM	FSIM	Human Rank (%)				
			A	B	C	Score	FM-Score
55.0	0.9944	0.669	25.9	16.2	58.0	39.0	52.6

effects (authenticity, matching degree with text, etc.). We set three different classes for the HR: a) if the action is perfectly done, the result would be ranked "A", representing score 1; b) if the action is correct, but other parts are affected, *e.g.*, mistakenly change the color or erasing other irrelevant objects, the result would be ranked "B", representing score 0.5; c) if the action is incorrect, the result is "C", representing score 0. Therefore, the final HR score is between 0 and 1, and the Full-Match score (FM-Score) represents that both the re-creation and the answer are correct. The higher scores of all metrics, the better performance is represented. The "accountable" part is totally automatic, for which we use the full string matching to measure the accuracy between ground truth answers and the final predictions. We also take into account that the textual feedback in the explanation part will change, such as that both explanations "There is no object1 and object2" and "There is no object2 and object1" are correct. We also designed tools to help with human evaluation, and we provide its interface and evaluation criteria in the appendix. In addition, we assign 100 sets of results to 5 works for scoring, we empirically find that our task usually has an explicit answer which is not easily affected by human subjective opinions.

Table 6: Texual feedback evaluation on the CLEVR-ATVC dataset. Type Acc. represents that the answer can correctly recognize the query that cannot be done or forbidden. Exp Acc. represents the explanation is correct in both answer type and the reasons.

Can		Cannot			Forbidden			Score (%)
Num	Acc.	Num	Type Acc.	Exp Acc.	Num	Type Acc.	Exp Acc.	
1662	71.6%	1997	49.4%	46.5%	1341	99.7%	58.2%	66.2

Table 7: Qualitative results of the Fruit-ATVC dataset.

Visual (V)	Text (T)	Re-created (M)	Answer (A)	**FM-Score**
	Please remove the plate.		No problem.	**1.0**
	Please put the lemon in the bowl.		No problem.	**0.75**
	Please exchange the position of the banana and the coconut.	/	This action cannot be done. Because there is no coconut.	**1.0**
	Please remove the bottle.	/	This action cannot be done. Because there is no kiwi and no bottle.	**0.75**

4.3 Results and Analyses

In this section, we evaluate the proposed method on the CLEVR-ATVC and Fruit-ATVC datasets, whose results include the image re-creation, textual feedback, qualitative visualizations and the model behavior when facing uncertainty.

4.3.1 Results on CLEVR-ATVC

For the CLEVR-ATVC dataset, the re-created results are shown in Table 5. We can see that the machine achieves FM-Score 52.6% (re-creation and textual feedback are completely correct), which also shows the difficulty of the task. The high SSIM 0.9944 represents the model can obtain high image reconstruction quality. Based on our observation, we find that the "exchange color" queries are mostly correct, while "exchange positions" queries are somewhat limited by the image reconstruction.

The quantitative results of textual feedback are shown in Table 6. We can see that there are 1662 pairs for the results on "can" queries; among them, 71.6% pairs are answered correctly. For the queries about "cannot", 49.4% pairs can be accurately recognized that the query is not doable, and 46.5% pairs can not only answer the type but correctly answer the reasons. For the queries about "forbidden", 99.7% can be correctly recognized, but only 58.2% can correctly explain the reasons. The score is the weighted average of

Table 8: Image re-creation evaluation on the Fruit-ATVC dataset.

PSNR	SSIM	FSIM	Human Rank (%)				
			A	B	C	Score	FM-Score
44.1	0.9272	0.420	12.8	29.4	57.9	27.5	46.1

Table 9: Textual feedback evaluation on the Fruit-ATVC dataset.

Can		Cannot			Score
Num	Acc.	Num	Type Acc.	Exp Acc.	
950	78.3%	922	77.2%	25.0%	64.7%

the above results. As there are numerous possible answers, a random guessing result is less than 1%, and thus it should be safe to conclude that the proposed method has shown impressive results.

Example qualitative results are shown in Table 4. We can conclude that the model can cay "No" to instructions and give accurate language-based explanations. In addition, it is worth mentioning that the model also implicitly learns the depth of the image and size of the objects, and thus it knows whether the object is occluded and how much it is occluded when you manipulate some objects.

4.3.2 Results on Fruit-ATVC

For the Fruit-ATVC dataset, the re-created results are shown in Table 8. The final score for re-creation is 46.1%, which results can produce both correct re-creation as well as the textual feedback. These results also demonstrate the the difficulty of the task, especially in real scenarios.

The quantitative results of answers are shown in Table 9. We can see that 78.3% of "can" queries can be answered correctly, and 25.0% queries about "cannot" can be correctly explained. In the "exchange position" queries, we find that although it can perform correct re-creation, the reconstruction for other objects is challenging. For the "put in" queries, the errors usually occur by creating a new target fruit without erasing it in the image. The re-creation results of the "remove" are the best among three actions. All the results suggest the challenges of this dataset of real scenes. Example qualitative results are shown in Table 7.

4.3.3 VQVAE vs. VQGAN

In this section, we evaluate the image auto-encoder methods, VQVAE Oord et al. (2017) and VQGAN Esser et al. (2021), on the CLEVR-ATVC sub-dataset. We only select one "put on top" action without textual feedback for evaluation. As shown in Table 11, although VQGAN-based image auto-encoder can output a large resolution and at 4 times faster of training than the VQVAE counterpart, the latter is 7.4% better. Based on our observations shown in the appendix, we find conclude VQGAN-based image encoder always changes the color of objects, erases objects and add irrelevant objects on the re-created images.

4.3.4 Uncertainty of Image Re-creation

In this section, we would like to explore whether the model can handle the uncertainty appeared in the query. The results show that our method deals with emerging uncertainties well. As shown in Table 12, all the image re-creation results get ranking A, except for the result of the last row. The score B is due to an object being erased during image reconstruction, not due to uncertainty. For example, as shown in the fourth row of Table 12, our model only exchanges the position of one "small yellow rubber cube", which is consistent with the settings in our constructed dataset. The "exchange position" only needs to process one of the objects, but "exchange color" needs to exchange the colors of all satisfied objects. Existing models are able to make different decisions for the above different situations, so our method shows a good behavior.

Table 10: Evaluation results on the CLEVR-ATVC sub-dataset.

Methods	Image	PSNR	SSIM	FSIM	Human Rank (%)			
					A	B	C	Score
Ours w/ VQVAE	128×128	**56.6**	**0.9963**	**0.693**	**66.4**	14.8	18.8	**73.8**
Ours w/ VQGAN	256×256	53.9	0.9947	0.604	52.7	**27.4**	**19.9**	66.4

Table 11: Evaluation results on the CLEVR-ATVC sub-dataset.

Methods	Text	Image	PSNR	SSIM	FSIM	Human Rank (%)			
						A	B	C	Score
Ours w/ VQVAE	Short	128×128	56.3	0.9957	0.689	65.9	14.4	**19.7**	71.9
Ours w/ VQVAE	Long	128×128	**56.6**	**0.9963**	**0.693**	**66.4**	**14.8**	18.8	**73.8**

4.4 Single Action without Answer

In this section, we firstly evaluate the image auto-encoder method VQVAE Oord et al. (2017) with short query on the CLEVR-ATVC sub-dataset. We only select one "put on top" action without textual feedback for verification and evaluation.

4.4.1 Short Query for Re-creation

We would like to verify the results on the shorter query and the imperfect query on the CLEVR-ATVC sub-dataset. We only select one "put on top" action without textual feedback for verification and evaluation. The former removes redundant text in the query, such as "please", "put", and "the" are removed, while the latter randomly removes keywords (such as "color", "size" or "shape"). The quantitative results shown in Table 11 show that the intact queries (long text) perform slightly better than the removed version (short text).

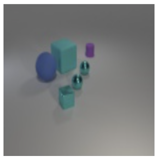
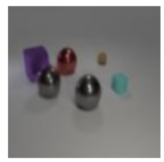
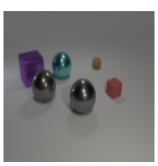
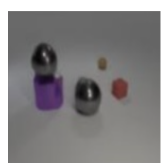
4.4.2 Imperfect Query for Re-creation

The visualization results are shown in Figure 4, with the observations on the imperfect queries: a) If the "blue" is removed from query, the generated cylinder seems to be shorter; b) If the query contains a nonexistent "blue" adjective for the cube, the model creates one; c) If both the nonexistent "blue" cube and "cyan" cylinder coexist in the same query, the model seems to mistakenly recognize the cyan cube, and we can find the small gray cylinder is also erased, which is incorrect; d) If we simply remove the "cube", the model just dyes a litter red on top of the cylinder; e) If we remove the "please put", the result just remains the original figure; f) If we add the unseen word, *e.g.*, "amazing", in the query, the manipulation result can still be achieved. We further test a lot of images, which show that the unlearned "create" and "dye" abilities commonly appear but not always. The above results also suggest that if we don't add restriction rules to the dataset, then the model will satisfy human instructions by creating new objects, and this uncontrollability is not the desired outcome. Therefore, it is necessary to regulate the behavior of the model by adding pre-set rules as supervision signals.

4.4.3 Analysis on VQGAN Results

The results shown in Figure 5 and B4 show that VQGAN-based image encoder always changes the color of objects, erases objects and add irrelevant objects on the re-created images. These results also demonstrate that VQGAN can produce high-quality images using fewer computational resources, but it appears to have lost its ability to discriminate on object properties in latent space.

Table 12: Uncertainty evaluation on CLEVR-ATVC dataset.

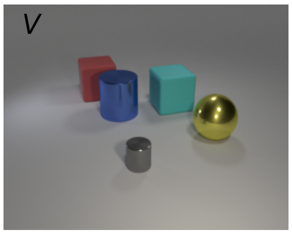
Visual (V)	Text (T)	Re-created (M)	Answer (A)	**Ranking**
	Please exchange the color of the large blue rubber sphere and the small purple rubber cylinder.		No problem.	**A**
	Please exchange the color of the large purple metal cube and the large red metal cylinder.		No problem.	**A**
	Please exchange the color of the large cyan metal sphere and the small red rubber cube.		No problem.	**A**
	Please exchange the position of the small yellow rubber cube and the small purple rubber cube.		No problem.	**A**
	Please exchange the color of the large purple metal cube and the small blue metal cylinder.		No problem.	**A**
	Please put the large purple metal cube under the large gray metal sphere.		No problem.	**B**

5 Related works

In this section, we first review some representative datasets and methods on vision-and-language tasks whose inputs and outputs are always unimodal. Next, we introduce the prior multimodal dialogue models. Finally, we briefly summarize the recent controllable text-to-image generation methods and the main differences between them and our task.

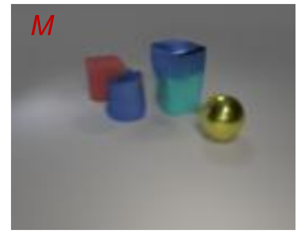
5.1 Vision-and-Language Tasks

Recent years have witnessed the rapid development of vision-and-language tasks. Their development trend can be observed through the construction of the datasets, which can be roughly categorized into two groups: *Vision to Language*, *Language to Vision*. We briefly summarize these relevant datasets and methods.



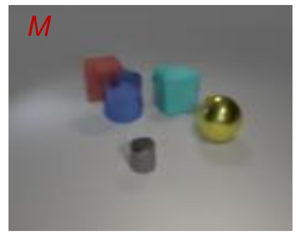
(a) Please put the big red cube on top of the big cylinder.

(b) Please put the big blue cube on top of the big blue cylinder.

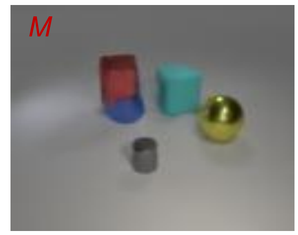


(c) Please put the big blue cube on top of the big cyan cylinder.

(d) Please put the big red on top of the cylinder.



(e) red cube on top of the big cylinder.



(f) Please put the amazing big red cube on top of the amazing big blue cylinder.

Figure 4: Image re-creation results on the CLEVR-ATVC sub-dataset using imperfect query. The red text represents the imperfect query, which may cause uncertainty.

Vision to Language. There are several sub-tasks for this category such as *Visual Description Generation* Ordonez et al. (2011), *Visual Storytelling* Lin et al. (2014), *Visual Dialog* Das et al. (2017); Kottur et al. (2019), *Visual Entailment* Vu et al. (2018); Liu et al. (2020), *Visual Reasoning* Johnson et al. (2017), *Visual Question Answering* Antol et al. (2015); Marino et al. (2019) and *Visual Referring Expression* Liu et al. (2019).

- Visual Description Generation, a.k.a., captioning, is the most well-known task, where many datasets in the past decade are constructed to address different scales, scenes, etc. For examples, image captioning datasets, SBU1M Ordonez et al. (2011), Flickr8k Hodosh et al. (2013), Flickr30k Young et al. (2014), MSCOCO Lin et al. (2014), Multi30k-CLID Elliott et al. (2016), Flickr30k-Entities Plummer et al. (2015), STAIR Captions Yoshikawa et al. (2017), Conceptual Captions Sharma et al. (2018), MSCOCO-Entities Cornia et al. (2019), Personality Captions Shuster et al. (2019) and video captioning datasets, MSVD Chen & Dolan (2011), MPII Cooking Rohrbach et al. (2012), YouCook Das et al. (2013), TACoS Regneri et al. (2013), TACoS-MultiLevel Rohrbach et al. (2014), M-VAD Torabi et al. (2015), MSR-VTT Xu et al. (2016), VTW Zeng et al. (2016), ANetCap Krishna et al. (2017), YouCook II Zhou et al. (2018), ANetEntities Zhou et al. (2019), COIN Tang et al. (2019), HowTo100M Miech et al. (2019), have played an important role in advancing visual captioning.

- Visual Storytelling usually requires the methods to generate a series of text to describe a set or a sequence of the visual inputs, as shown in several datasets: NYC-Storytelling Park & Kim (2015), Disneyland Storytelling Kim et al. (2015), SIND, and VIST Huang et al. (2016).

- Visual Dialog can be regarded as a natural conversation system based on the image or video. Examples can be found in the existing benchmarks such as VisDial Das et al. (2017), CLEVR-Dialog Kottur et al. (2019) and AVSD Alamri et al. (2018).

- Visual Entailment is introduced in V-SNLI Vu et al. (2018), SNLI-VE Xie et al. (2019), and VIOLIN Liu et al. (2020), which requires the method to learn to select the correct premise and hidden semantic information that can be inferred from the visual contents.

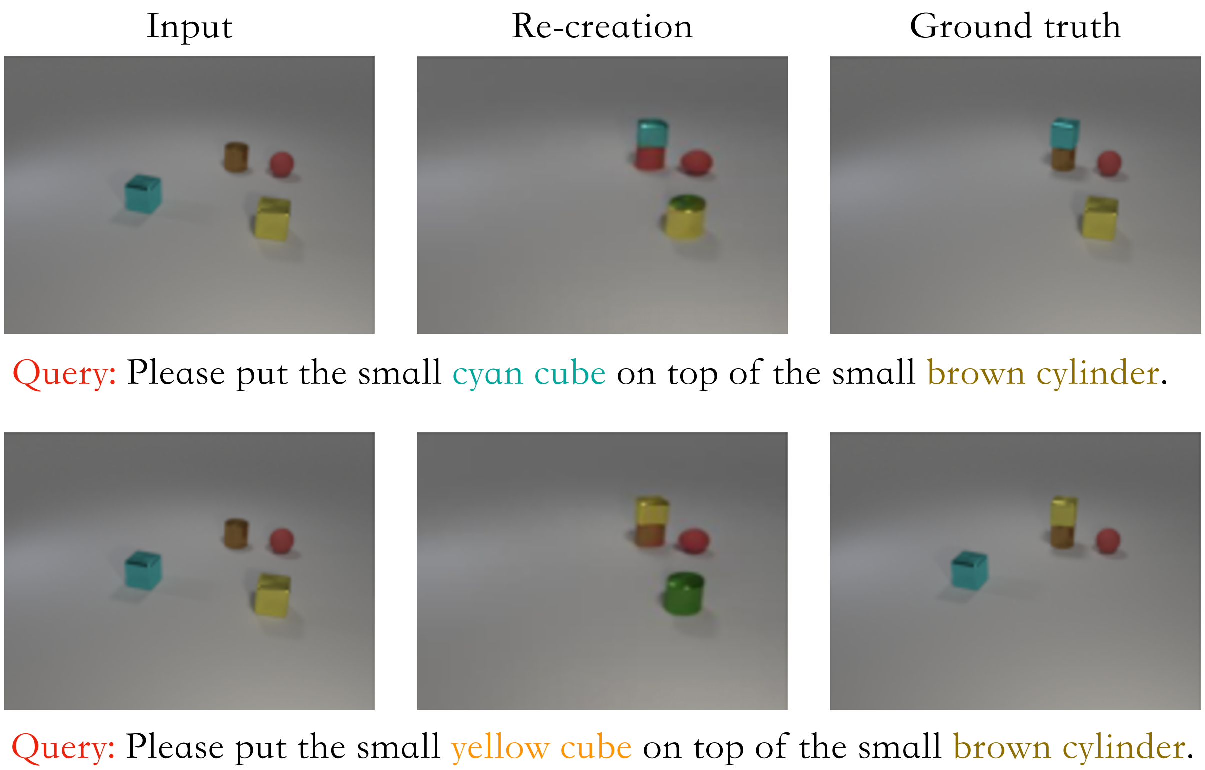


Figure 5: Analyse for the results of VQGAN-based image encoder. VQGAN frequently changes the colors of the objects, which downgrades the performance.

- Vision Reasoning is expected to provide the scene graphs between objects and then to answer questions based on the visual input and relationships. Many datasets such as CLEVR Johnson et al. (2017), NLVR Suhr et al. (2017), CLEVR-CoGenT Yang et al. (2018), NLVR2 Suhr et al. (2018), GQA Hudson & Manning (2019), VCR Zellers et al. (2019), and Visual COMET Park et al. (2020) have been constructed between 2017 and 2020, which demonstrate visual reasoning abilities can be effectively learned by the deep neural networks LeCun et al. (2015).
- Visual Question Answering is another representative vision-to-language task, in which the model is expected to generate text-based answer according to the question and the visual cues. The most influential dataset VQA v1.0 Antol et al. (2015) as well as MovieQA Tapaswi et al. (2016), TVQA Lei et al. (2018), OK-VQA Marino et al. (2019), KVQA Shah et al. (2019), and VQA v2.0. Antol et al. (2015) have greatly promoted the development.
- Visual Referring Expression is another task which requires the method to comprehend the referring expressions by showing evidence in the visual images, *i.e.*, using bounding boxes to locate the corresponding objects. There are several datasets targeting this task, such as RefCLEF Kazemzadeh et al. (2014), and CLEVR-Ref+4 Liu et al. (2019).

Language to Vision. This task aims at image generation based on the pure natural language. Currently, there are only a few datasets for this task such as Oxford-102 Reed et al. (2016), Caltech-UCSD Birds (CUB) Reed et al. (2016), Multi-Modal-CelebA-HQ Xia et al. (2021), Text2Human Jiang et al. (2022), Laion-400M Schuhmann et al. (2021), Laion-5B Schuhmann et al. (2022) and Text2Video Xu et al. (2016); Li et al. (2018); Hong et al. (2022); Ho et al. (2022); Khachatryan et al. (2023); **?**. These datasets are similar to the vision-to-language task above, and the text-image pairs contain text descriptions of the image content. The most representative task is text-based image generation, which has recently seen large progress in multimodal learning Brooks et al. (2023); Li et al. (2023b). Many previous works train GANs with textconditioning on publicly available image captioning datasets Tao et al. (2020); Zhang et al. (2021). DALL·E Ramesh et al. (2021b) uses 250 million text-image pairs to successfully achieve promising results, which can even perform zero-shot visual reasoning by using some prompts. Another zero-shot model termed CLIP Radford et al. (2021) is used to rank and measure the similarity between image and text, as usually one text prompt can produce numerous plausible results. The recent DALL·E 2 Ramesh et al. (2022) uses a diffusion prior on CLIP text latents and cascaded diffusion models to generate high resolution 1024×1024 images. GLIDE Nichol et al. (2021) also applies guided diffusion to the problem of text-conditional image synthesis. Imagen Saharia et al. (2022) combines the power of transformer language models with high-fidelity diffusion models to deliver an unprecedented degree of photorealism in text-to-image synthesis.

Unlike the above datasets that usually focus on one-side, our proposed datasets require models to simultaneously generate both visual re-creations and textual feedbacks. The visual manipulation requires not only to perform the actions correctly but remain the visual plausible background, meanwhile the feedback must be aware of all doable, cannot do, and forbidden queries with language-based answer, which insures the value and challenge of the datasets.

5.2 Multimodal Dialogue Models

The vision to language task requires visual sample as input and text as output (e.g., Visual Dialog), while the language to vision does the opposite. In the multimodal dialogue methods, the model needs to reason about multimodal (unimodal) input and give multimodal (unimodal) output, which can be divided into the following two categories.

- *Multimodal Input and Unimodal Output (MIUO).* This conversational task is very similar to visual dialog. Its input usually includes visual input and text-based prompts (i.e., multimodal input). The visual language model will perform visual reasoning on the image and feed back the answer (i.e., unimodal output) Alayrac et al. (2022); Brooks et al. (2023); Gong et al. (2023); Bo Li (2023); Li et al. (2023a); OpenAI (2023b); Su et al. (2023); Yang et al. (2023b); Zhao et al. (2023); Zhu et al. (2023); Liu et al. (2023); Mu et al. (2023); Wang et al. (2023); Zhang et al. (2023). GPT-4 OpenAI (2023b) and MultiModal-GPT Gong et al. (2023) can accept prompts consisting of both arbitrarily interlaced visual inputs and text-based query, and the model generates text outputs. VideoChat Li et al. (2023a) proposes a video-centric multimodal dialogue dataset, and the trained model is capable of understanding and generating detailed conversations about videos.
- *Multimodal Input and Multimodal Output (MIMO).* This task requires the model to simultaneouly perform multimodal reasonning and generation Koh et al. (2023); Jing Yu Koh (2023); Chenfei Wu & Duan (2023); Yang et al. (2023a). Visual ChatGPT Chenfei Wu & Duan (2023) is a pioneering work that combines ChatGPT and a series of pre-trained visual foundation models to make them can accept and output text and image during textual-visual chat. GILL Koh et al. (2023) proposes a mapping network to efficiently map the output embedding space of a frozen text-only LLM to that of a frozen generation model (e.g., Stable Diffusion Rombach et al. (2022)), which only needs to be finetuned a small number of parameters on image-caption pairs for image retrieval, novel image generation, and multimodal dialogue. FROMAGe Jing Yu Koh (2023) is also involves image-text inputs and outputs to achieve multimodal dialogue, in which a few linear layers are finetuned and the pre-trained LLM is frozen. GPT4Tools Yang et al. (2023a) proposes an instruction dataset and extends Visual ChatGPT Chenfei Wu & Duan (2023) to more image understanding task.

5.3 Controllable Text-to-image Generation

The above text-based image generation methods mainly focus on generating a new high-quality image from a given text, and cannot allow the user to manipulate the generation of specific visual attributes using natural language descriptions Anderson et al. (2018); Nguyen et al. (2019); Thomason et al. (2020); Cheng et al. (2014); Scalise et al. (2018); Stepputtis et al. (2020); Nam et al. (2018); Yüksel et al. (2021); Shen et al. (2020); Li et al. (2020); Richardson et al. (2020); Esser et al. (2021). There are some works have investigated how the synthesis process can be controlled by representing scenes as compositional generative neural feature fields, which allows us to disentangle objects from the background as well as individual objects' shapes and appearances Li et al. (2019); Bodla et al. (2018); Nam et al. (2018); Chen et al. (2021); Shuster et al. (2019). ControlGAN Li et al. (2019) can effectively synthesise high-quality images and also control parts of the image generation according to natural language descriptions. FusedGAN Bodla et al. (2018) disentangles the different factors in the latent space to achieve more controllability in sampling. Text2Human Jiang et al. (2022) can synthesize Human images by specifying the clothes shapes and textures using solely natural language descriptions.

The prior works mainly control the image generation process by composition or disentanglement approaches. In this paper, we would like to control the results of re-created images, in which the system needs to learn to say "no" to commands that are prohibited or cannot be completed.

6 Conclusion

In this paper, we raise an important but less studied issue, the accountability of multimodal generative models. To this end, we construct two new datasets, CLEVR-ATVC and Fruit-ATVC, for a novel task, termed Accountable Text-based Visual Re-creation, which aims to teach VLMs to reject human instructions. Our datasets include both visual and textual inputs and outputs, requiring the model to complete the visual re-creation while ensuring image quality if the answer is "yes." If the model could not complete the action or it was prohibited, it had to provide an explanation. The baseline models, experimental settings, evaluation metrics, and a thorough analysis are provided with some promising results. Our high-quality datasets can be applied to other similar tasks, and we hope this work inspires further research on the accountability problem, including model design, label annotation, and larger datasets. We believe that addressing the issue of accountability is critical for advancing the development and deployment of multimodal generative models.

References

Huda Alamri, Chiori Hori, Tim K Marks, Dhruv Batra, and Devi Parikh. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*, volume 2, 2018.

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

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

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

Liangyu Chen Jinghao Wang Fanyi Pu Jingkang Yang Chunyuan Li Ziwei Liu Bo Li, Yuanhan Zhang. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

Navaneeth Bodla, Gang Hua, and Rama Chellappa. Semi-supervised fusedgan for conditional image generation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 669–683, 2018.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 190–200, 2011.

Tianyi Chen, Yi Liu, Yunfei Zhang, Si Wu, Yong Xu, Feng Liangbing, and Hau San Wong. Semi-supervised single-stage controllable gans for conditional fine-grained image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9264–9273, 2021.

Weizhen Qi Xiaodong Wang Zecheng Tang Chenfei Wu, Shengming Yin and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Yu Cheng, Yunyi Jia, Rui Fang, Lanbo She, Ning Xi, and Joyce Chai. Modelling and analysis of natural language controlled robotic systems. *IFAC Proceedings Volumes*, 47(3):11767–11772, 2014.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8307–8316, 2019.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335, 2017.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2634–2641, 2013.

Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.

Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5706–5714, 2017.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *CVPR*, 2021.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, 2016.

Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 2(3):11, 2019.

Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. *arXiv preprint arXiv:2105.03761*, 2021.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

Gunhee Kim, Seungwhan Moon, and Leonid Sigal. Ranking and retrieval of image sequences from multiple paragraph queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1993–2001, 2015.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*, 2019.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549–5558, 2020.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7880–7889, 2020.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023a.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10900–10910, 2020.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4185–4194, 2019.

Michele Loi and Matthias Spielkamp. Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 757–766, 2021.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2630–2640, 2019.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12527–12537, 2019.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11453–11464, 2021.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

OpenAI. Chatgpt. *https://openai.com/blog/chatgpt/*, 2023a.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023b.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

Cesc C Park and Gunhee Kim. Expressing an image stream with a sequence of natural sentences. *Advances in neural information processing systems*, 28:73–81, 2015.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pp. 508–524. Springer, 2020.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021a.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021b.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pp. 1060–1069. PMLR, 2016.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pp. 184–195. Springer, 2014.

Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1194–1201. IEEE, 2012.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Rosario Scalise, Shen Li, Henny Admoni, Stephanie Rosenthal, and Siddhartha S Srinivasa. Natural language instructions for human–robot collaborative manipulation. *The International Journal of Robotics Research*, 37(6):558–565, 2018.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Haddow B Sennrich R and Birch A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909v5*, 2015.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8876–8884, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12516–12526, 2019.

Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Neurips*, 2020.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223, 2017.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.

Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4631–4640, 2016.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pp. 394–406. PMLR, 2020.

Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. *arXiv preprint arXiv:1806.05645*, 2018.

Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2256–2265, 2021.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Guangyu Robert Yang, Igor Ganichev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. A dataset and architecture for visual reasoning with a working memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 714–731, 2018.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023a.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023b.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*, 2017.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. *arXiv preprint arXiv:2104.00820*, 2021.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731, 2019.

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Generation for user generated videos. In *European conference on computer vision*, pp. 609–625. Springer, 2016.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023.

Zhiqiang Zhang, Wenxin Yu, Jinjia Zhou, Xuewen Zhang, Ning Jiang, Gang He, and Zhuo Yang. Customizable gan: A method for image synthesis of human controllable. *IEEE Access*, 8:108004–108017, 2020.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yimin Zhou, Yiwei Sun, and Vasant Honavar. Improving image captioning by leveraging knowledge graphs. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 283–293. IEEE, 2019.

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

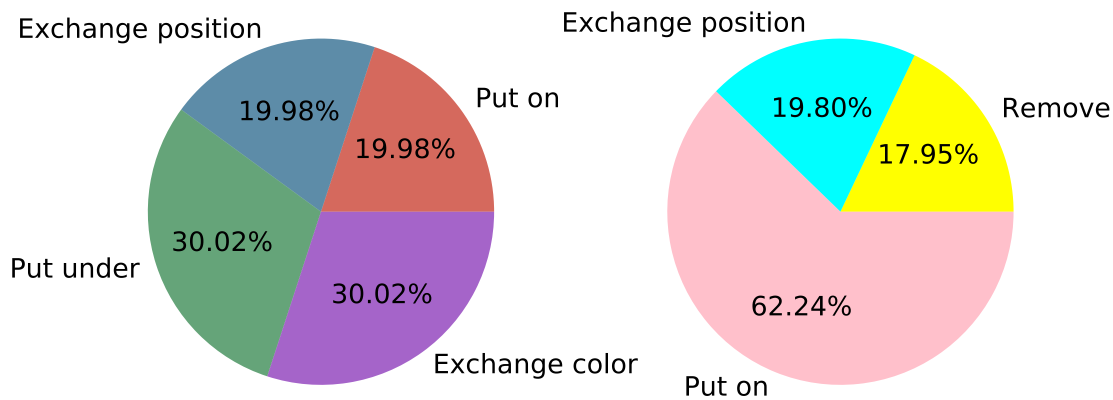


Figure A1: The distribution of different actions on CLEVR-ATVC and Fruit-ATVC datasets.

A More Details of Data

In this section, we first introduce the format of annotation file for our datasets. Next, we provide more example images of Fruit-ATVC dataset to show the diversity of collection scenes. We also provide the summary of statistics for the different actions on the datasets.

A.1 Annotation File

The Listing 1 and Listing 2 show the format of annotation files on the CLEVR-ATVC and Fruit-ATVC datasets. The annotation files follow a uniform format, with the only difference being that detailed 2D and 3D coordinate locations of objects are provided in the CLEVR-ATVC dataset. The Figure B2 illustrates the visualization of bounding boxes in the CLEVR-ATVC dataset. We hope that this dataset can be used for other similar tasks in the future, such as compositional visual reasoning. We have introduced the main structure of annotation files in the main paper. Here, we would like to present more details of them.

At the beginning of the annotation file, we summarize the contributors to the dataset, time, the project link, the version, and license information. Next are the categories of objects included in this dataset, the adjectives of the objects, and the actions, etc. All the information about visual inputs is included in the "images", including image index, image filename, image link, object number and objects. In the child list "objects", we also list in detail the corresponding properties of each object, including index, category index, coordinates of bounding box, etc. The text-based queries (Q) and textual feedbacks (A) are included in the "questions", in which the answer types are presented in numbers from 0 to 2. The numbers 0, 1, and 2 indicate that the instruction cannot be executed, can be executed, and the prohibited instruction, respectively. If the action type is 0 or 2, then "actions" in the next "recreations" will be empty. In particular, in the CLEVR-ATVC dataset, we also provide information about the new color or position of the object after a displacement or property change.

A.2 Diversity of Fruit-ATVC

The following Figure B3 shows the diversity of collection scenes on Fruit-ATVC dataset. The image resolution is average 2700×2700, and the images were taken with different categories of mobile phones, including iphone, huawei, xiaomi, etc. We would like to emphasize that the image quality of this dataset is very high, which results in the size of the original Fruit-ATVC is 168 GB. In addition, Moreover, the background of data collection, fruit categories and container types are also very diverse. We believe this dataset can also be used for other image generation tasks. The Figure A1 illustrates the distribution of different actions on our constructed datasets.

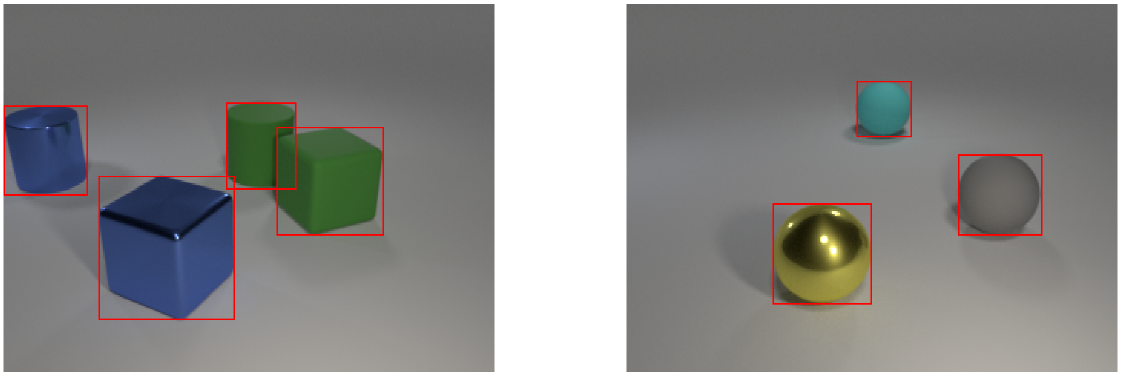


Figure B2: The visualization of bounding boxes in the CLEVR-ATVC dataset.

B Human Evaluation Experiments

In this section, we show some examples of interfaces used to assist human evaluation. During the evaluation phase, the visual input, text-based query, re-created image, answer, and ground truths for re-creation and textual feedback will be automatically imported to the corresponding location of the auxiliary tool, and the evaluator only needs to score the corresponding results and save them. After that, we will calculate the final score of the model based on the saved results. As introduced in the main paper, we assign the same results to different works, and they all provided the same score for the experimental results. The above experiments show that the auxiliary tools we design can not only greatly reduce the time required for evaluation, but also the widely used human evaluation process is accurate and reliable.

Figure B3: The example images of Fruit-ATVC dataset show the diversity of collection scenes.

Listing 1: The format of annotation file on CLEVR-ATVC dataset.

```
{
  "contributor": "Anonymous",
  "data_created": "2023",
  "url": "Anonymous",
  "description": "CLEVR-ATVC Dataset",
  "vertion": "1.0",
  "licenses": {
    "id": "1",
    "url": "Nnone",
    "name": "Creative Commons Attribution (CC-BY 4.0)",
  },
  "categories": ["cube, cylinder, sphere"],
  "sizes": ["small, large"],
  "colors": ["gray, red, blue, green, brown, purple, cyan, yellow"],
  "material": ["rubber, metal"],
  "actions": ["put on top, put under, exchange position, exchange color"],
  "images": [
        {
        "image_idx": "0000001",
        "image_filename": "521100_03_000011.png",
        "id": "521100_03_000011",
        "object_number": 3,
        "atvc_url": "http://images.atvcdata.org/clevratvc/train2023/V/521100_03_000011.png
        "data_created": "2023-04-27 00:17:14",
        "objects": [
            {
            "index": 0,
            "category_id": 0,
            "size_id": 1,
            "color_id": 3,
            "bbox": [29, 62, 81, 116],
            "3d_coords": [-2.96856045, -1.99158716, 0.6999999],
            "material_id": 1},
            {...},
            {...},
                    ]
        },
    ]
  "questions": [
        {
        "question_idx": "0000001",
        "question_id": "521100_03_000011",
        "question_number": 10,
        "ques": [
            {
            "ques_id": 0,
            "ques_idx": 1,
            "id": "521100_03_000011_0_01",
            "type": 1,
            "Q": ["Please put the cylinder on top of the cube."],
            "A": ["No problem."]
            },
        ]
```

```
    "recreations": [
        {
        "rec_idx": "0000001",
        "rec_id": "521100_03_000011",
        "rec_num": 10,
        "actions": [
            {
            "actions_id": 0,
            "actions_idx": 1,
            "rec_filename": "521100_03_000011_0_01.png",
            "object_number": 3,
            "atvc_url": "http://images.atvcdata.org/clevratvc/train2023/M/521100_03_000011_
            "date_created": "2023-04-27 00:17:59",
            "objects": [
            {},
            {},
            {
            "index": 2,
            "category_id": 1,
            "size_id": 0,
            "color_id": 1,
            "bbox": [39, 45, 65, 72],
            "3d_coords": [-2.96856045, -1.99158716, 1.75],
            "material_id": 0}
                        ]
            },
        ]
}
```

Listing 2: The format of annotation file on Fruit-ATVC dataset.


```
{
  "contributor": "Anonymous",
  "data_created": "2023",
  "url": "Anonymous",
  "description": "Fruit-ATVC Dataset",
  "vertion": "1.0",
  "licenses": {
    "id": "1",
    "url": "Nnone",
    "name": "Creative Commons Attribution (CC-BY 4.0)",
  },
  "fruits": ["apple, pear, cocount, orange, banana, kiwi, mango, avocado, lemon, pitaya"],
  "containers": ["plate, bowl, bottle"],
  "actions": ["put in, exchange position, remove"],
  "images": [
        {
        "image_idx": "0000001",
        "image_filename": "0000001.png",
        "id": "0000001",
        "fruit_number": 3,
        "container_number": 1,
        "atvc_url": "http://images.atvcdata.org/fruitatvc/train2023/V/0000001.png",
        "data_created": "2023-04-27 00:10:11",
        "fruits": [
                {
                    "index": 0,
                    "category_id": 0
                }
                {
                    "index": 1,
                    "category_id": 6
                }
                {
                    "index": 2,
                    "category_id": 8
                }
            ],
        "containers": [
                {
                    "index": 0,
                    "category_id": 2
                }
            ],
        },
    ]
  "questions": [
        {
        "question_idx": "0000001",
        "question_id": "0000001",
        "question_number": 3,
        "ques": [
            {
            "ques_id": 0,
```

```
            "ques_idx": 1,
            "id": "0000001_0_01",
            "type": 1,
            "Q": ["Please put the kiwi in the plate."],
            "A": ["No problem."]
            },
        ]
    "recreations": [
        {
        "rec_idx": "0000001",
        "rec_id": "0000001",
        "rec_num": 3,
        "actions": [
            {
            "actions_id": 0,
            "actions_idx": 1,
            "rec_filename": "0000001_0_01.png",
            "fruit_number": 3,
            "container_number": 1,
            "atvc_url": "http://images.atvcdata.org/fruitatvc/train2023/M/0000001_0_01.png
            "date_created": "2023-04-27 00:10:18",
            },
        ]
}
```

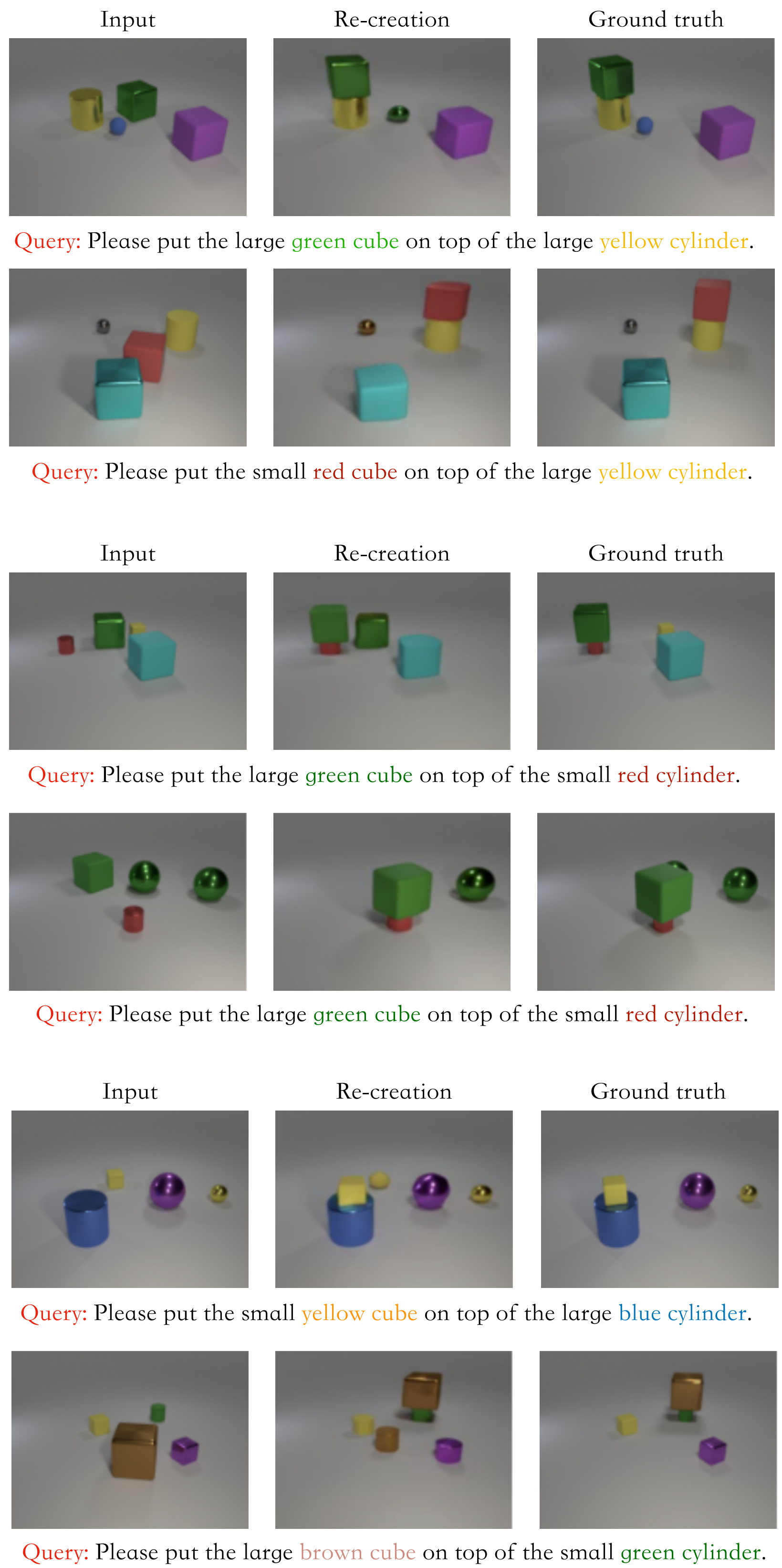
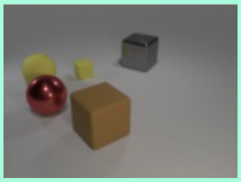
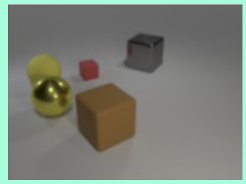


Figure B4: More results are to show why VQGAN-based image encoder performs worse on image re-creation.

Input	Re-creation	Ground truth
		

Query: Please exchange the color of the small yellow rubber cube and the large red metal sphere.

Answer: No problem.

Ground truth: No problem.

Human Evaluation

1. Please rate the visual re-creation?

● A: 1 ○ B: 0.5 ○ C: 0

2. Please rate the type of answer?

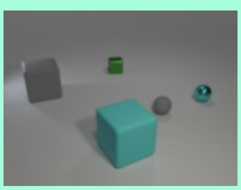
● 1 ○ 0

3. Please rate the explanation of answer?

● 1 ○ 0

4. What is the score for textual feedback? [1.0]

5. What is the full-match score? [1.0]

Input	Re-creation	Ground truth
		

Query: Please put the large gray metal cylinder under the small yellow rubber cylinder.

Answer: This action cannot be done. Because there is no large gray metal cylinder and no small yellow rubber cylinder.

Ground truth: This action cannot be done. Because there is no large gray metal cylinder and no small yellow rubber cylinder.

Human Evaluation

1. Please rate the visual re-creation?

● A: 1 ○ B: 0.5 ○ C: 0

2. Please rate the type of answer?

● 1 ○ 0

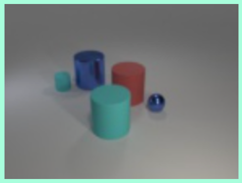
3. Please rate the explanation of answer?

● 1 ○ 0

4. What is the score for textual feedback? [1.0]

5. What is the full-match score? [1.0]

Input	Re-creation	Ground truth



Query: Please put the large tray metal cube on top of the large blue metal sphere.

Answer: This action is forbidden. Because you cannot put an object on top of the sphere, and there is no large gray metal cube and no large blue metal sphere.

Ground truth: This action is forbidden. Because you cannot put an object on top of the sphere, and there is no large gray metal cube and no large blue metal sphere.

Human Evaluation

1. Please rate the visual re-creation?

   ● A: 1 ○ B: 0.5 ○ C: 0

2. Please rate the type of answer?

   ● 1 ○ 0

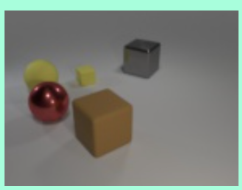
3. Please rate the explanation of answer?

   ● 1 ○ 0

4. What is the score for for textual feedback? [1.0]

5. What is the full-match score? [1.0]

Input	Re-creation	Ground truth



Query: Please put the large red metal sphere under the small yellow rubber cube.

Answer: This action is forbidden. Because you cannot put the sphere under an object, and there is no large red metal sphere and no small yellow rubber cube.

Ground truth: This action is forbidden. Because you cannot put the sphere under an object.

Human Evaluation

1. Please rate the visual re-creation?

   ● A: 1 ○ B: 0.5 ○ C: 0

2. Please rate the type of answer?

   ● 1 ○ 0

3. Please rate the explanation of answer?

   ○ 1 ● 0

4. What is the score for for textual feedback? [0.5]

5. What is the full-match score? [0.75]

Input | Re-creation | Ground truth

Query: Please put the small yellow metal sphere under the large red metal cube.

Answer: This action is forbidden. Because you cannot put the sphere under an object, and there is no small yellow metal sphere and no small red metal cube.

Ground truth: This action is forbidden. Because you cannot put the sphere under an object, and there is no small yellow metal sphere and no large red metal cube.

Human Evaluation

1. Please rate the visual re-creation?

   ● A: 1 ○ B: 0.5 ○ C: 0

2. Please rate the type of answer?

   ● 1 ○ 0

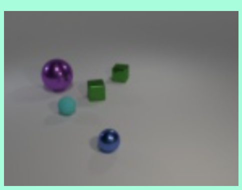
3. Please rate the explanation of answer?

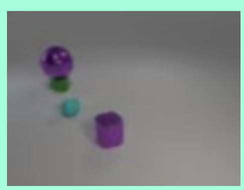
   ○ 1 ● 0

4. What is the score for for textual feedback? [0.5]

5. What is the full-match score? [0.75]

Input | Re-creation | Ground truth





Query: Please put the small green metal cube under the large purple metal sphere.

Answer: No problem.

Ground truth: No problem

Human Evaluation

1. Please rate the visual re-creation?

   ○ A: 1 ● B: 0.5 ○ C: 0

2. Please rate the type of answer?

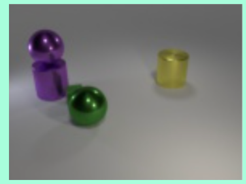
   ● 1 ○ 0

3. Please rate the explanation of answer?

   ● 1 ○ 0

4. What is the score for for textual feedback? [1.0]

5. What is the full-match score? [0.75]

Input	Re-creation	Ground truth
		

Query: Please put the large purple metal sphere on top of the large purple metal cylinder.

Answer: This action is forbidden. Because you cannot put the sphere under an object.

Ground truth: No problem.

Human Evaluation

1. Please rate the visual re-creation?

   ○ A: 1 ○ B: 0.5 ● C: 0

2. Please rate the type of answer?

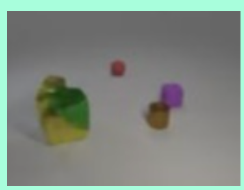
   ○ 1 ● 0

3. Please rate the explanation of answer?

   ○ 1 ● 0

4. What is the score for for textual feedback? [0]

5. What is the full-match score? [0]

Input	Re-creation	Ground truth
		

Query: Please exchange the position of the large yellow metal cube and the large green metal cylinder.

Answer: No problem.

Ground truth: This action cannot be done. Because there is no large yellow metal cube and no large metal cylinder.

Human Evaluation

1. Please rate the visual re-creation?

   ○ A: 1 ○ B: 0.5 ● C: 0

2. Please rate the type of answer?

   ○ 1 ● 0

3. Please rate the explanation of answer?

   ○ 1 ● 0

4. What is the score for for textual feedback? [0]

5. What is the full-match score? [0]