Going beyond Imagination! Enhancing Multi-modal Dialogue Agents with Synthetic Visual Descriptions

Haolan Zhan, **Sameen Maruf** [*], **Ingrid Zukerman** and **Gholamreza Haffari**
Department of Data Science & AI, Monash University, Australia
{haolan.zhan, ingrid.zukerman, gholamreza.haffari}@monash.edu

Abstract

Building a dialogue agent that can seamlessly interact with humans, in multi-modal regimes, requires two fundamental abilities: (1) understanding emotion and dialogue acts within situated user scenarios, and (2) grounding perceived visual cues to dialogue contexts. However, recent works have uncovered shortcomings of existing dialogue agents in understanding emotions and dialogue acts, and in grounding visual cues effectively. In this work, we investigate whether additional dialogue data with only visual descriptions can help dialogue agents effectively align visual and textual features, and enhance the ability of dialogue agents to ground perceived visual cues to dialogue contexts. To this end, in the absence of a suitable dataset, we propose a synthetic visual description generation pipeline, and contribute a large-scale synthetic visual description dataset. In addition, we propose a general training procedure for effectively leveraging these synthetic data. We conduct comprehensive analyses to evaluate the impact of synthetic data on two benchmarks: MELD and IEMOCAP. Our findings suggest that synthetic visual descriptions can serve as an effective way to enhance a dialogue agents' grounding ability, and that the training scheme affects the extent to which these descriptions improve the agent's performance.

1 Introduction

There have been impressive advances in large-scale vision and language models (VLMs) in performing multi-modal tasks, such as visual question answering (VQA) and image captioning (Guo et al., 2023; Chen et al., 2022; Liu and Chen, 2024). While VLMs are powerful general-purpose models for a wide range of tasks, most state-of-the-art VLMs still struggle with providing real-world, situated multi-modal assistance (Wu et al., 2023, 2024).

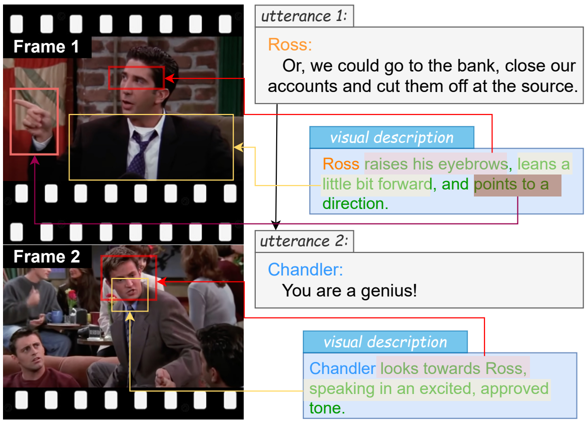

Frame 1
utterance 1:
Ross:
Or, we could go to the bank, close our accounts and cut them off at the source.
visual description
Ross raises his eyebrows, leans a little bit forward, and points to a direction.
Frame 2
utterance 2:
Chandler:
You are a genius!
visual description
Chandler looks towards Ross, speaking in an excited, approved tone.

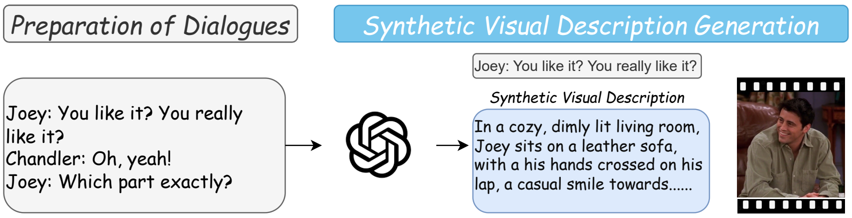
Figure 1: Visual descriptions can be an effective way to help dialogue agents interpret the visual cues from images, further enhancing the understanding ability towards human emotion and dialogue acts.

Building a situated dialogue agent that can seamlessly interact with humans in a multi-modal scenario requires two essential abilities: (1) understanding the interlocutor's emotion and dialogue acts within situated user scenarios, and (2) grounding perceived visual cues to dialogue contexts.

However, recent work (Wu et al., 2023; Liu et al., 2023; Xenos et al., 2024) has unveiled shortcomings of existing VLM-based dialogue agents with respect to these abilities. We hypothesize that current limitations can be attributed to the gap between different modalities, also known as the misalignment between visual and textual features. We argue that visual descriptions can serve as a potential way to bridge this gap by interpreting visual cues from images. To verify our hypothesis, we propose to investigate whether additional dialogue data with only visual descriptions can help dialogue agents effectively align visual and textual features, and enhance their ability to ground perceived visual cues to dialogue contexts. For instance, looking at the images in Figure 1, *visual descriptions* are capable of conveying subtle but important visual

[*] Work was done when Sameen was at Monash.

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 420–427
September 18–20, 2024. ©2024 Association for Computational Linguistics


Preparation of Dialogues
Synthetic Visual Description Generation
Synthetic Visual Description

Joey: You like it? You really like it?
Chandler: Oh, yeah!
Joey: Which part exactly?

Joey: You like it? You really like it?

In a cozy, dimly lit living room, Joey sits on a leather sofa, with a his hands crossed on his lap, a casual smile towards......

Figure 2: Synthetic data generation pipeline. Please note that the image on the right is provided for reference only, to aid in understanding the generated visual description, and is not produced by ChatGPT.

cues (e.g., *facial expression*, *human position*) about the people in these images.

Given the absence of datasets that offer annotations for visual descriptions, we devise a novel synthetic visual description generation pipeline using ChatGPT and contribute a large-scale synthetic visual description dataset by extending existing multi-dialogue corpora with additional visual descriptions. Furthermore, to effectively utilize these synthetic data, we explore several training schemes based on *knowledge distillation* (*KD*) (Hinton et al., 2015). Those training schemes aim to instruct dialogue agents to align the features in different modalities by distilling the ability to interpret visual cues learnt from the synthetic data.

We conduct a comprehensive analysis where we evaluate the effectiveness of synthetic data on two benchmarks: *MELD* (Poria et al., 2018) and IEMOCAP (Busso et al., 2008). Our results show that synthetic visual descriptions play an important role in helping dialogue agents understand and ground visual cues from images to dialogue contexts. Specifically, our method outperforms the baseline VLMs (e.g., LLaVa-1.5) by at least 6% on both emotion detection and dialogue act classification tasks. Moreover, the remarkable performance of our training framework based on knowledge distillation demonstrates that the training scheme affects the extent to which these descriptions improve a dialogue agent's performance.

2 Synthetic Data Generation

To appropriately understand an interlocutor's emotions and dialogue acts, VLM-based dialogue agents must ground perceived visual cues within dialogue contexts. We hypothesize that current VLMs are limited by a large gap between different modalities, which affects their ability to *ground* visual cues to dialogue contexts. However, the annotation that links visual cues and dialogue contexts is missing from existing widely-used datasets (e.g., MELD (Poria et al., 2018)). In this section, we investigate whether additional *visual descriptions* can help dialogue agents, and to what extent synthetic data can be used to bridge the gap between visual cues and dialogue contexts.

We can easily have access to large amounts of dialogue contexts, but it is hard to obtain the corresponding images or videos. In the absence of suitable multimodal datasets with the grounding annotations, we propose a visual description generation approach in the rest of this section. We then propose a training procedure (§ 3) for leveraging this synthetic data to improve the performance of multi-modal dialogue agents on the tasks of dialogue act and emotion prediction.

Problem Formulation and Notation. Given a multi-turn dialogue $d = \{u_1, \ldots, u_m\}$ consisting of a sequence of utterances u_i in plain text, our goal is to prompt ChatGPT to generate synthetic visual description $v_i^{'}$ for each utterance. We will get a synthetic dataset in which each of the utterances is paired with a synthetic visual description instead of a image. This synthetic dataset, augmented with the visual descriptions, will be used in training for reducing the modality gap, as explained below; see Figure 2 for an example.

Synthetic Visual Description Generation. Multi-modal dialogue tasks utilize plain-text dialogue and visual cues simultaneously. The motivation for the synthetic visual description generation is to explore if we can leverage it instead of real images to improve the performance of multi-modal dialogue agents. The idea is that these descriptions will stimulate a VLM-based dialogue agent to ***imagine*** potential visual scenes. We select three main factors that can affect visual scenes on the task of dialogue act and emotion prediction, viz (1) facial expression, (2) human action, and (3) human position; and incorporate them into synthetic visual description. We then prompt ChatGPT via in-context learning (ICL) to generate a potential synthetic visual description for each utterance, as shown in Figure 2. We provide an example in the Appendix B to better understand the synthetic generation process.

From the MELD and IEMOCAP datasets, we have extracted and prepared 6,357 multi-turn dialogues, comprising a total of 22,126 utterances. Each utterance is associated with a synthetic visual description that depicts the potential visual scene associated with the dialogue context. The average length of each synthetic visual description is 15.6 words.

3 Fine-tuning a VLM-based Agent using Synthetic Data

In this section, we propose a methodology for leveraging the synthetic data produced as explained in Section 2. One intuitive way is to combine synthetic data with real data for training. However, as there is a large gap in both modalities and patterns between real images and synthetic visual descriptions, a straightforward concatenation of real and synthetic dataset would not be the best choice. We therefore propose a multi-stage training framework, which trains the dialogue agent with synthetic and real data separately, followed by a knowledge distillation training stage (Hinton et al., 2015). Specifically, we choose the state-of-the-art VLM model, LLaVa-v1.5 (Liu et al., 2023) as the backbone of our system, which integrates the visual encoder of CLIP (Radford et al., 2021) with the language decoder Vicuna (Chiang et al., 2023).

Fine-tuning with Synthetic Data. Suppose we have a synthetic training dataset of dialogues $\mathcal{D}_s$, where each dialogue $d' = \{(u'_1, v'_1, y'_1), ..., (u'_m, v'_m, y'_m)\}$ contains m utterances (u'), associated synthetic visual descriptions (v') and output labels (y'). We use the synthetic training dataset to fine-tune the LLaVa-v1.5 model with LoRA adapter (Hu et al., 2021), denoted by θ_s. As the dialogues and synthetic visual descriptions are both in text, instead of feeding images to the CLIP module, we only need to use the Vicuna module as the proxy to encode the synthetic descriptions for the visual encoding.

Fine-tuning with Real Data. The goal of fine-tuning with real data is to adapt the dialogue model to the real multi-modal situation. We have a real dataset $\mathcal{D}_r$ containing a set of multi-modal dialogues, where each dialogue $d = \{(u_1, v_1, y_1), ..., (u_m, v_m, y_m)\}$ has m utterances (u), corresponding images (v) and output labels (y). Unlike synthetic data fine-tuning, the CLIP and Vicuna module within the LLaVa-v1.5 will be used to process visual images and dialogue contexts collaboratively. This process will yield a fine-tuned adapter θ_r for the real data.

Knowledge Distillation. The distillation training procedure aims to transfer the "***imagination***" ability learnt from the synthetic data to enhance dialogue agents in grounding visual cues to dialogue contexts in multi-modal settings. We conduct the knowledge distillation procedure on the fine-tuned adapters θ_s and θ_r by applying the KL-divergence (Kullback and Leibler, 1951) regularization in three different settings, as follows.

- *Synthetic distillation (s → r)*: Knowledge is distilled from the synthetic adapter θ_s to the real adapter θ_r, based on the following training objective:

$$\max_{\theta_r} \sum_{d \in D_r} \sum_{(u,v,y) \in d} \log P_{\theta_r}(y|u,v) - \gamma KL(P_{\theta_r}(.|v,y)||P_{\theta_s}(.|v,y))$$

  where $\log P_{\theta_r}(y|u,v)$ refers to log-likelihood probability of generated label y from the model with real adapter θ_r. Besides, the distillation function $KL(\cdot||\cdot)$ aims to measure and minimize the difference between θ_r and θ_s. γ is the regularisation coefficient to control the trade-off between two objectives.

- *Real distillation (r → s)*: Knowledge is distilledn from the real adapter θ_r to the synthetic adapter θ_s, based on the following training objective:

$$\max_{\theta_s} \sum_{d' \in D_s} \sum_{(u',v',y') \in d'} \log P_{\theta_s}(y'|u',v') - \gamma KL(P_{\theta_s}(.|v',y')||P_{\theta_r}(.|v',y'))$$

- *Mutual distillation (s ↔ r)*: This is a mutual KD between two adapters,

$$\max_{\theta_r} \sum_{d \in D_r} \sum_{(u,v,y) \in d} \log P_{\theta_r}(y|u,v) - \gamma_1 KL(P_{\theta_r}(.|v,y)||P_{\theta_s}(.|v,y)) - \gamma_2 KL(P_{\theta_s}(.|v,y)||P_{\theta_r}(.|v,y))$$

Dataset	MELD		IEMOCAP	
	Emo.	DA	Emo.	DA
UniVL	66.37	61.47	54.91	61.19
MiniGPT-4	78.00	70.33	69.00	68.93
Video-LLaMa	72.38	68.42	63.16	65.75
MultiModal-GPT	73.54	68.01	61.27	64.92
LLaVa-1.5	79.26	76.39	66.03	71.48
ours	**87.38***	**81.03***	**73.89***	**77.29***

Table 1: Accuracy (%) of VLM-based multi-modal dialogue agents on the emotion (**Emo.**) and dialogue act (**DA**) understanding tasks. "*" indicates a significance of p-value < 0.05 in the Chi-Square test after Benjamini-Hochberg (BH) correction for false discovery rate (Benjamini and Hochberg, 1995).

where γ_1 and γ_2 are regularisation coefficients to balance the effects of two types of distillation between θ_r and θ_s.

4 Experiments

In our experiments, we aim to investigate the following two research questions: (1) How effectively do existing VLM-based agents comprehend emotions and dialogue acts from humans?, and (2) To what extent can synthetic visual descriptions enhance agents' multi-modal capabilities in understanding emotions and dialogue acts.

Settings. Our experiments were conducted on two datasets: *MELD* (Poria et al., 2018) and IEMOCAP (Busso et al., 2008), both of which are rich in annotations of emotion and dialogue acts. We evaluate the performance of each model by reporting its accuracy in predicting emotion and dialogue acts. In terms of VLMs, we select several state-of-the-art baselines including **UniVL** (Luo et al., 2021), **MiniGPT-4** (Zhu et al., 2023), **Video-LLaMa** (Zhang et al., 2023), **MultiModal-GPT** (Gong et al., 2023) and **LLaVa** (Liu et al., 2023). The details of each baseline can be found in Appendix A.

Performance of Existing VLMs. Table 1 presents the accuracy (%) of existing VLM-based agents on the emotion and dialogue act understanding tasks. We observe that LLaVa-1.5 outperforms other VLMs to a large extent in the MELD dataset and maintains competitive performance with MiniGPT-4 on the IEMOCAP dataset We also note that existing VLMs mainly rely on their LLM module (e.g., Vicuna module in the LLaVa-1.5 agent), but they struggle to merge the information extracted from the CLIP (visual) module with the LLM (textual) module, mainly due to the modality misalignment. The results support our hypothesis that visual descriptions can help bridge the gap by interpreting visual cues from images. We further provide an in-depth analysis of the impact of visual descriptions.

Dataset	MELD		IEMOCAP	
	Emo.	DA	Emo.	DA
LLaVa-1.5 (vanilla)	79.26	76.39	66.03	71.48
(1) *synthetic data (s)*	75.84	68.16	59.98	66.92
(2) *real data (r)*	82.67	78.25	69.72	72.94
(3) *mixture (s then r)*	84.01	79.86	71.35	74.16
(4) *mixture (r then s)*	76.15	71.55	62.04	68.28
(5) *distillation (s $\rightarrow$ r)*	**87.38***	**81.03***	73.89	**77.29**
(6) *distillation (r $\rightarrow$ s)*	80.19	77.91	65.63	71.74
(7) *distillation (r $\leftrightarrow$ s)*	85.43	79.59	**74.52**	76.13

Table 2: Ablation studies of different types of training data and distillation settings. "*" indicates a significance of p-value < 0.05 in the Chi-Square test with BH correction.

The Effectiveness of Synthetic Data. We conducted comprehensive ablation studies to investigate the effectiveness of using synthetic data to enhance the performance of our agent. We selected the top-performing VLM model, LLaVa-1.5, from Table 1 as our baseline. The results are presented in Table 2, which outlines seven different data configurations, including: (1) training only on *synthetic* data, (2) training only on *real* data, (3) mixed training involving initial training on *synthetic* data followed by *real* data, (4) mixed training involving the reverse sequence, and employing distillation techniques as discussed in Section 3, viz (5) synthetic distillation (synthetic→real), (6) real distillation (real→synthetic) and (7) mutual distillation (real↔synthetic).

The findings in Table 2 indicate that incorporating knowledge distillation into the training process enables LLaVa-1.5 to surpass the performance achieved through either naive mixed training or training solely on synthetic data or on real data. Notably, among the three distillation approaches ((5)-(7)), the strategy of distilling knowledge from synthetic to real data (*distillation (s $\rightarrow$ r)*) yielded the best results overall. In contrast, the performance of distillation from real to synthetic data was largely equivalent to that of LLaVa-1.5. This suggests that synthetic data must be utilized judiciously, as a significant discrepancy between real and synthetic data can adversely affect performance.

5 Related Work

Visual Dialogue. The visual dialogue task was proposed by Das et al. (2017). It requires an agent to answer multi-round questions about a given image, similarly to Visual Question Answering (VQA) (Das et al., 2017; Jiang et al., 2020b; Huber et al., 2018). Previous work (Wu et al., 2018; Kottur et al., 2018; Yang et al., 2019; Guo et al., 2019; Niu et al., 2019; Kang et al., 2019; Jiang et al., 2020a; Yang et al., 2021) focused on developing different attention mechanisms to model the interactions among image, question and dialogue history (Wang et al., 2020). With the rapid development of large-scale vision-language models (VLMs) (Chen et al., 2022; Dai et al., 2022; Wu et al., 2023; Zhu et al., 2023; Zhang et al., 2023), recent work focuses on building unified models that can handle multiple tasks. However, most models are still unable to support situated interaction with humans in real scenarios, especially capturing human emotions and dialogue acts, and grounding to their dialogue contexts.

Learning from synthetic data. There has been some work on learning from synthetic data for dialogue systems (Dai et al., 2022; Kim et al., 2022; Semnani et al., 2023; Bao et al., 2023; Abdullin et al., 2024; Zhan et al., 2024). Synthetic data are easy to generate, and are particularly useful for providing detailed labelling to reduce human labor, such as dialogue acts (Zhan et al., 2023), knowledge injection (Bao et al., 2023) or simulating dialogues in new scenarios, such as the rapid generation of a sequence of QA from documents (Dai et al., 2022). However, these works mainly focus on plain text dialogues, rather than multi-modal dialogues. We propose a novel framework to utilize synthetic data to address this gap, and thereby enhance the abilities of multi-modal dialogue agents on the task of emotion and dialogue act classification.

6 Conclusion

Our work demonstrates the potential of synthetic visual descriptions to improve the performance of dialogue agents, particularly in understanding emotions, dialogue acts and grounding visual cues to dialogue contexts. By introducing a novel synthetic visual description generation pipeline and a large-scale dataset, along with an effective training procedure, we have taken a crucial step towards overcoming the limitations of multi-modal dialogue agents. The positive outcomes observed in our experiments highlight the importance of appropriate training schemes to fully leverage synthetic data, pointing towards a promising direction for future research.

Limitations

As our work provides an initial step into incorporating synthetic visual descriptions into multimodal dialogue agents, we do not offer an exhaustive analysis of the synthetic data, nor do we identify the most suitable use cases for evaluating the effectiveness of synthetic data in such scenarios. Besides, we did not analyse why certain distillation schemes do better than others. Additionally, it is promising to conduct further evaluation to determine whether enhancing the agents' grounding capabilities could also improve their response abilities.

Acknowledgement

This material is based on research sponsored by DARPA under agreement number HR001122C0029. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon.

References

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. Synthetic dialogue dataset generation using llm agents. *arXiv preprint arXiv:2401.17461*.

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.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *ACL*, pages 10866–10882, Toronto, Canada. ACL.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

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

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10434–10443.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *CHI*.

Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. 2020a. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11125–11132.

Xiaoze Jiang, Jing Yu, Yajing Sun, Zengchang Qin, Zihao Zhu, Yue Hu, and Qi Wu. 2020b. Dam: Deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue. In *IJCAI*.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. 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 2024–2033.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2378.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

Mengchen Liu and Chongyan Chen. 2024. An evaluation of gpt-4v and gemini in online vqa. *arXiv preprint arXiv:2312.10637*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2021. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Zheng-Yu Niu, Hua Wu, Haifeng Wang, et al. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. 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 1782–1792.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. *EMNLP*.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. 2024. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*.

Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. 2023. Simmc-vr: A task-oriented multimodal dialog dataset with situated and immersive vr streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6273–6291.

Wei Wu, Xu Sun, and Houfeng Wang. 2018. Question condensing networks for answer selection in community question answering. In *ACL*.

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2024. Vllms provide better context for emotion understanding through common sense reasoning. *arXiv preprint arXiv:2404.07078*.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2561–2569.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, et al. 2024. Renovi: A benchmark towards remediating norm violations in socio-cultural conversations. *arXiv preprint arXiv:2402.11178*.

Haolan Zhan, Sameen Maruf, Lizhen Qu, Ingrid Zukerman, and Gholamreza Haffari. 2023. Turning flowchart into dialog: Plan-based data augmentation for low-resource flowchart-grounded troubleshooting dialogs. *arXiv preprint arXiv:2305.01323*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

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

A Baseline Models

UniVL (Luo et al., 2021) is a unified video and language pre-training model for multi-modal understanding and generation. UniVL model adpts Transformer as the backbone and has individual language and video encoder, following with a cross-encoder and decoder module.

MiniGPT-4 (Zhu et al., 2023) contains a vision encoder with a pre-trained ViT and Q-Former model, a single linear projection layer, and an advanced Vicuna large language model (LLM). MiniGPT-4 freezes the vision part and only requires training the linear projection layer to align the visual features with the Vicuna.

Video-LLaMa (Zhang et al., 2023) maintains a similar architecture with the MiniGPT-4, including the ViT and Q-Former for the visual and audo encoder. On the top of the architecture, a LLM (LLaMa or Vicuna) is followed to align multi-modal features with contextual features.

MultiModal-GPT (Gong et al., 2023) is based on the open-flamingo (Alayrac et al., 2022) model. MultiModal-GPT consists of a vision encoder from CLIP, a sampler to receive the spatial features from the vision encoder, and a language decoder LLaMA.

LLaVa-1.5 (Liu et al., 2023) considers the pre-trained CLIP visual encoder ViT-L/14, then the visual feature will be sent into a simple linear layer to connect image features into the word embedding space. Finally, LLaVa-1.5 choose the Vicuna as the LLM backbone.

B An Example of Synthetic Data Generation

We provide an example for better understanding the synthetic data generation process. Suppose we have prepared a dialogue shown in the Figure 2. Then, we prompt ChatGPT to generate a synthetic visual description. We present an example in Figure 3 to illustrate this process.

Synthetic Data Generation Prompt

Can you imagine a potential scene for the context of the following dialogue? Try to incorporate several elements: Facial expression, Human Action and Human Position.

Here is an example:
`<Dialogue Context>`:
Ross. Or! Or, we could go to the bank, close our accounts and cut them off at the source.

`<Visual Description>`:
Ross raises his eyebrows, leans forward, and speaks with a hint of frustration and determination.

Now, please generate a reasonable visual description for the last utterance of the following dialogue:

`<Dialogue Context>`:
Joey: You like it? You really like it?
Chandler: Oh, yeah!
Joey: Which part exactly?

Please try to generate a concise description within 30 words.

`<Output - Synthetic Visual Description>`:
In a cozy, dimly lit living room, Joey sits on a leather sofa, with a his hands crossed on his lap, leans forward, a casual smile towards Chandler.

Figure 3: An example of synthetic data generation.