Comparison Visual Instruction Tuning

Anonymous Author(s)

Abstract

Comparing two images in terms of Commonalities and Differences (CaD) is a fundamental human capability that forms the basis of advanced visual reasoning and interpretation. It is essential for the generation of detailed and contextually relevant descriptions, performing comparative analysis, novelty detection, and making informed decisions based on visual data. However, surprisingly, little attention has been given to these fundamental concepts in the best current mimic of human visual intelligence - Large Multimodal Models (LMMs). We develop and contribute a new two-phase approach CaD-VI for collecting synthetic visual instructions, together with an instruction-following dataset CaD-Inst containing 349K image pairs with CaD instructions collected using CaD-VI . Our approach significantly improves the CaD spotting capabilities in LMMs, advancing the SOTA on a diverse set of related tasks by up to 17.5%. It is also complementary to existing difference-only instruction datasets, allowing automatic targeted refinement of those resources increasing their effectiveness for CaD tuning by up to 10%. Additionally, we propose an evaluation benchmark with 7.5K open-ended QAs to assess the CaD understanding abilities of LMMs.

1 Introduction

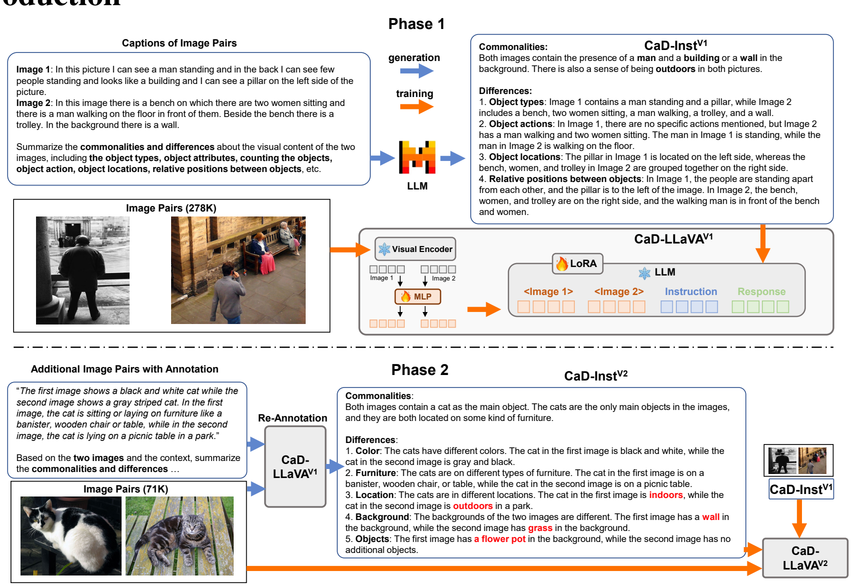

Figure 1: Pipeline of our two-phase CaD-VI : In Phase-1, we leverage captions for image pairs and an LLM to generate CaD VI data - CaD-Inst^{V1} (278K), and perform visual instruction tuning on it to arrive at the Phase-1 model CaD-LLaVA^{V1} . In Phase-2, we leverage CaD-LLaVA^{V1} to generate CaD VI data on additional image pairs and collect CaD-Inst^{V2} (71K). Visual instruction tuning with CaD-Inst^{V1} and CaD-Inst^{V2} leads to our final model CaD-LLaVA^{V2} .

Submitted to Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Understanding the Commonalities and Differences (CaD) between two signals (e.g., images) is a basic capability innate to humans [1]. Spotting change and difference alerts us to interesting events happening in our surroundings, warns us of hazard, and drives us toward learning new concepts exposed after the change or relative movement. Understanding what is common helps structure visual information and allows differences to emerge by elimination. Together, these form powerful tools for human learning and acquiring world knowledge.

The forefront of modern AI shifted with the recent emergence of Large Language Models (LLMs) [2], where the top-performing ones [3–6] closely align to human reasoning and world-knowledge capabilities. LLMs' great performance and wide applicability quickly led to their wide adoption into most of the current ML pipelines. In the Vision community, this impacted the development of Large Multi-modal Models (LMMs) [4, 7–12] largely considered the best available mimic of human visual intelligence to date. While multiple methods for adding multi-modal support to LLMs have been proposed, currently the more popular and better performing open LMMs largely rely on tuning using Visual Instructions (VI) [7, 13]. These methods align image tokens produced by visual encoders to be 'understandable' by an LLM decoder, allowing images to be seamlessly integrated into the LLM decoder input context stream together with the query text during inference. In most recent methods [7, 9–11], VI takes the form of a multi-turn conversation: with 'human' turns providing image context and asking the questions, and LMM turns answering them [7]. However, the majority of VI data focused on providing merely a single image in the VI conversations [7], while only a few works included multi-image VI samples [12, 14], and surprisingly, very few included some form of CaD VI data [9, 10, 15] to enable CaD support in the resulting LMM.

Due to the fundamental importance of endowing LMMs with CaD capabilities, thus getting them closer to achieving human visual intelligence in all its diversity, we propose CaD-VI - a multi-phase CaD generation approach, for progressive dense and structured CaD VI data collection, which we employ to build CaD-Inst training curriculum and associated CaD-QA benchmark comprised of CaD-related open-ended questions, both contributed in this work. In essence, the final CaD-Inst curriculum associates diverse and large-scale (349K) image pair collection with highly detailed and structured CaD summaries. CaD summaries computed for an additional set of 7.6K image pairs, are used for extracting open CaD-related QA resulting in CaD-QA .

As shown in Fig. 1, the Phase-1 of CaD-VI is a 'cold start' where, in the absence of LMMs with substantial CaD capabilities, we leverage image captions and an LLM to hallucinate (coarse) CaD VI data - CaD-Inst^{V1} (278K), where we collect *structured* and *detailed* CaD summaries for our paired images sourced from a dense & large-scale image collection [16]. Training on the first phase CaD-Inst^{V1} data we arrive at CaD-LLaVA^{V1} - an LMM that has strong CaD capabilities compared to a large variety of leading LMMs including the very few trained with some CaD data (see Sec. 4). Next, leveraging our CaD-LLaVA^{V1} model to produce non-hallucinated, image-informed CaD data, we generate additional CaD instructions into the collection CaD-Inst^{V2} (71K). Combining CaD-Inst^{V1} and CaD-Inst^{V2} we form CaD-Inst and train our final CaD-LLaVA^{V2} 7B and 13B LMMs to achieve (1) significant (up to 17.5%) absolute improvement over a large variety of recent SOTA LMMs over a variety of 5 CaD-related existing closed-QA evaluation benchmarks (namely BISON[17], SVO Probes[18], NLVR2[19], EQBEN[20], and COLA[21]), and (2) strong (up to over 20%) relative improvements on our contributed open-QA CaD benchmark - CaD-QA . Additionally, as CaD-Inst can be safely mixed with the LLaVA VI data [22], we show in Tab. **??** that our CaD-LLaVA^{V2} models effectively avoid forgetting the general capabilities of the corresponding LLaVA LMMs.

Our contributions are as follows: (i) we contribute CaD-Inst - a large-scale visual instruction tuning dataset for enhancing CaD reasoning capabilities of LMMs; (ii) we contribute CaD-QA - an open QA evaluation benchmark for assessing CaD capabilities; (iii) we contribute and open source a CaD-VI methodology for collecting and enhancing CaD instruction tuning data; (iv) we demonstrate significant (up to 17.5%) improvements in CaD reasoning for LMMs trained using CaD-Inst as well as potential to scale CaD-Inst via self-improvement by CaD-Inst -trained models.

2 Two-Phase CaD Visual Instruction Tuning

As illustrated in Fig. 1, our CaD-VI consists of two phases: in Phase-1, we employ an LLM to generate summary of CaD for image pairs (Sec. 2.1) and perform visual instruction tuning on the collected data (Sec. 2.2); in Phase-2, we leverage the Phase-1 model to generate CaD on additional image pairs and perform training with combined instruction data from both phases (Sec. 2.3).

2.1 Phase-1a: LLM Instruction Data Collection

In our first phase, we leverage an LLM to generate a summary of commonalities and differences for a pair of two images (Fig. 1 (top row)). Specifically, we construct image pairs and prompt an LLM, supplying it with two image captions (one per image) and an instruction prompt asking it to summarize all the commonalities and differences according to the provided captions, contributing to our first phase CaD instruction data collection denoted as CaD-Inst^{V1} .

We select the Localized Narratives dataset [16] which consists of 873K image-caption pairs. Inspired by LLaVA [7] who used an LLM for visual instruction collection, we leverage the Mixtral 8×7B open LLM [23] for generating detailed and structured summaries of commonalities and differences for pairs of images. As the LLM can only accept text as input, in Phase 1 we use image captions to represent visual content of images. This is a rather crude approximation, which is alleviated in Phase 2 of our CaD-VI approach. We specifically prompt the LLM *to structure* the commonalities and differences summaries according to the following 6 visual aspects: (i) object types; (ii) attributes; (iii) counts; (iv) actions; (v) locations; and (vi) relative positions; as illustrated in Fig. 1.

In CaD-Inst^{V1} we collected structured summaries of CaD for 278K image pairs, with average length of 157 words (40 for commonalities and 117 for differences). We construct a two-turn conversation for each image pair. In the first turn, we define the task of summarizing CaD by providing the encoded visual tokens of the two images and instructing the model to summarize the CaD , where the response part of the turn is the LLM-generated structured summary collected above. In this instruction, we do not provide the image captions, forcing the model to rely only on image tokens to complete the task. In the second turn, we reinforce the image-text alignment by employing a simple task of text-to-image retrieval to avoid forgetting the model's general capabilities. We randomly sample one of the two captions and request the model to select the image (from the current pair) to which the caption belongs.

2.2 Phase-1b: CaD Visual Instruction Tuning

Architecture. As illustrated in Fig. 1, we use our collected CaD-Inst^{V1} data to perform visual instruction tuning using the open-sourced code of LLaVA-1.5 [22] LMM. The LLaVA-1.5 model consists of $\phi_L(\cdot;\theta_L)$ - a pretrained Vicuna 1.5 [24] LLM (finetuned from LLama 2 [25]); $\phi_V(\cdot;\theta_V)$ - a pretrained visual encoder CLIP ViT-L/14@336px [26]; and $\phi_M(\cdot;\theta_M)$ - a two-layer MLP projector converting the visual encoder tokens to post-embedding layer LLM tokens. Given a pair of two images x_{V_1}, x_{V_2} and the instruction x_I, the MLP projects the visual features computed by the visual encoder into embedded language tokens, *i.e.* $v_k = \phi_M(\phi_V(x_{V_k};\theta_V);\theta_M), k \in \{1,2\}$. Then the projected visual features and instruction text tokens are concatenated and fed into the LLM, where the response text tokens are generated in an autoregressive manner, *i.e.* $\hat{x}_R^i = \phi_L([v_1, v_2, x_I, \hat{x}_R^{<i}];\theta_L)$, where $\hat{x}_R^i$ denotes the i-th token in the generated response.

Training. We finetune the LLaVA-1.5 model using the LLaVA [7] pipeline. Specifically, following LLaVA pre-training, we finetune only the pretrained projection MLP and the (frozen) LLM with LoRA adapters [27]. We minimize the CLM loss of the next token prediction in the responses, $\mathcal{L}_{CLM} = \sum_i -\log p(\hat{x}_R^i|V_1, V_2, x_I, x_R^{<i})$.

To preserve the general VL capabilities of the LMM, we merge our CaD-Inst^{V1} with the finetuning data of LLaVA-1.5 (665K samples). In Tab. **??** we show that CaD-VI indeed preserves the general LMM capabilities compared to LLaVA-1.5 as evaluated on the SEED benchmark [28]. Phase-1 CaD visual instruction tuning results in our cold-start model CaD-LLaVA^{V1} which is an LMM that can be used for annotating visual commonalities and differences.

2.3 Phase-2: Data Collection and Tuning

Phase-2a: LMM-based CaD Instruction Collection. While in Phase 1 we used an LLM to extract a CaD summary based on human-generated captions, for Phase 2 data collection we leverage our Phase 1 model CaD-LLaVA^{V1} and additional image pairs to extract the CaD summaries informed by the images directly. Here we select the Scene-Difference [15] collection as an additional image source. It contains 71K pairs of similar images from COCO [29] and provides annotation of unstructured difference-only summaries (see Fig. 1 bottom left for an example). We feed both the image pairs and the original annotations into our CaD-LLaVA^{V1} model, and generate a *structured summary* of *both* commonalities and differences.

Phase-2b CaD Visual Instruction Tuning We follow the Phase-1b introduced in Sec. 2.2 for CaD visual instruction tuning. Here we finetune on a combination of LLaVA 1.5 [22] finetune data (665K), CaD-Inst^{V1} data (278K) and CaD-Inst^{V2} data (71K). This leads to the Phase 2 model, denoted as CaD-LLaVA^{V2} .

3 Benchmark of Open-Ended CaD QA

In order to evaluate LMMs on answering open-ended questions on commonalities and differences of a pair of two images, we construct and contribute the CaD-QA benchmark.

Data Collection. Similar to the data collection pipeline introduced in Sec. 2.1, we employ Visual Genome [30] and the detailed image captions from SVIT [31] as image & caption source. We collect 7.5K image pairs with 8 or more overlapping nouns in their captions. For each pair, we employ the Mixtral 8×7B LLM to produce the structured CaD summaries from the captions. Next, we prompt Mixtral with both the image captions and the CaD summary, instructing it to generate a multi-turn conversation with several rounds of Q&A, providing some in-context examples of the desired layout. Finally, we randomly select one Q&A per conversation. There are 7520 QA pairs with an average answer length of 26 words.

LLM-assisted Evaluation. Motivated by LLMs' ability to judge response consistently with human assessment [24], we employ the Mixtral 8×7B LLM to compare the generated responses to the collected open-ended QA responses. We feed the question, correct answer, and the predicted answer into the LLM and instruct it to provide a rating between 0 and 5 for the predicted answer quality.

4 Experiments

Evaluation Datasets

We evaluate on several VQA benchmarks of closed-ended and open-ended questions. For **closed-ended VQA on image pairs**, we include BISON [17], SVO Probes [18], EQBEN [20], COLA [21] and NLVR2 [19]. We also evaluate SEED-Bench Video [28] with two frames sampled from each video. For **open-ended tasks**, we use the LLM-as-a-judge metric (Sec. 3) and evaluate on our CaD-QA . Furthermore, we also directly evaluate the quality of LMM predicted CaD summaries for 210 image pairs in COLA with shorter summaries generated from brief captions, and for the 7.5K lengthy summaries from CaD-QA generated from detailed VG captions.

Dataset Random chance	#Instruct. Data	BISON 50%	SVO 50%	NLVR2 50%	EQBEN 25%	COLA 25%
SparklesChat	6.5K	56.70%	43.93%	58.00%	19.17%	20.00%
Otter	2.8M	40.67%	47.33%	52.00%	8.33%	8.10%
MMICL	5.8M	80.00%	88.13%	56.67%	20.83%	25.71%
EMU2-Chat	1.3M	46.00%	47.93%	60.00%	7.50%	13.33%
InternLM-XComposer2-VL	>600K	80.67%	82.07%	66.67%	25.00%	32.38%
LLaVA 1.6 7B	<1M	66.00%	70.40%	58.67%	20.83%	11.90%
LLaVA 1.6 13B	<1M	81.33%	82.13%	60.00%	17.50%	24.76%
LLaVA 1.5 7B	665K	54.00%	46.80%	61.33%	17.50%	7.62%
LLaVA 1.5 13B	665K	59.33%	56.27%	66.00%	16.67%	12.38%
CaD-VI 7B	1M	95.33%	92.73%	66.67%	39.17%	40.95%
CaD-VI 13B	1M	**96.67%**	**93.00%**	**69.33%**	**42.50%**	**43.33%**

Table 1: Performance on closed-ended VQA tasks with image pairs in accuracy. Here the method CaD-VI denotes our Phase-2 model CaD-LLaVA^{V2} .

Comparison to State-of-the-Art LMMs We first compare CaD-LLaVA^{V2} (denoted by CaD-VI in Table) to state-of-the-art LMMs on closed-ended VQA in Table 1. SparklesChat [9], Otter [10], MMICL [32], EMU2-Chat [12], InternLM-Xcomposer2-VL [33] all include samples with multi-image inputs in the visual instruction tuning while LLaVA 1.5 [22] and LLaVA 1.6 [34] are tuned with only single image instructions. The evaluated benchmarks are challenging due to the visually very similar image pairs with subtle compositional differences where the LMMs could easily make an incorrect decision leading to performance below random chance. Our CaD-VI 7B model already outperforms all the other baselines on the five benchmarks and our 13B finetuned model further boosts the performance.

References

[1] I. D. F. .-. IxDF, *What are the gestalt principles?* Aug. 2016. [Online]. Available: https://www.interaction-design.org/literature/topics/gestalt-principles.

[2] R. Bommasani, D. A. Hudson, E. Adeli, P. Liang, and et al., *On the opportunities and risks of foundation models*, 2022. arXiv: 2108.07258 [cs.LG].

[3] O. et al., *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL].

[4] G. T. et al., *Gemini: A family of highly capable multimodal models*, 2024. arXiv: 2312.11805 [cs.CL].

[5] Anthropic, *The claude 3 model family: Opus, sonnet, haiku*, 2024. arXiv: 2312.11805 [cs.CL].

[6] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2023.

[8] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, *The dawn of lmms: Preliminary explorations with gpt-4v(ision)*, 2023. arXiv: 2309.17421 [cs.CV].

[9] Y. Huang, Z. Meng, F. Liu, Y. Su, C. Nigel, and Y. Lu, "Sparkles: Unlocking chats across multiple images for multimodal instruction-following models," *arXiv preprint arXiv:2308.16463*, 2023.

[10] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[11] X. Dong *et al.*, "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[12] Q. Sun *et al.*, "Generative multimodal models are in-context learners," 2023.

[13] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[14] A. Awadalla *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[15] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[16] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, Springer, 2020, pp. 647–664.

[17] H. Hu, I. Misra, and L. Van Der Maaten, "Evaluating text-to-image matching using binary image selection (bison)," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[18] L. A. Hendricks and A. Nematzadeh, "Probing image-language transformers for verb understanding," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3635–3644.

[19] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6418–6428.

[20] T. Wang, K. Lin, L. Li, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Equivariant similarity for vision-language foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 998–12 008.

[21] A. Ray, F. Radenovic, A. Dubey, B. Plummer, R. Krishna, and K. Saenko, "Cola: A benchmark for compositional text-to-image retrieval," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[22] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023.

[23] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[24] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[28] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[31] B. Zhao, B. Wu, and T. Huang, "Svit: Scaling up visual instruction tuning," *arXiv preprint arXiv:2307.04087*, 2023.

[32] H. Zhao, Z. Cai, S. Si, X. Ma, K. An, L. Chen, Z. Liu, S. Wang, W. Han, and B. Chang, "Mmicl: Empowering vision-language model with multi-modal in-context learning," in *The Twelfth International Conference on Learning Representations*, 2024.

[33] P. Zhang, X. D. B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, H. Yan, *et al.*, "Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition," *arXiv preprint arXiv:2309.15112*, 2023.

[34] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, *Llava-next: Improved reasoning, ocr, and world knowledge*, Jan. 2024. [Online]. Available: `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.