When Words Outperform Vision: VLMs Can Self-Improve Via Text-Only Training For Human-Centered Decision Making

Anonymous ACL submission

Abstract

Embodied decision-making is fundamental for AI agents operating in real-world environments. While Visual Language Models (VLMs) have advanced this capability, they still struggle with complex decisions, particularly in human-centered situations that require deep reasoning about human needs and values. In this study, we systematically evaluate open-sourced VLMs on multimodal human-centered decision-making tasks. We find that LLMs receiving only textual descriptions unexpectedly outperform their VLM counterparts of similar scale that process actual images, suggesting that visual alignment may hinder VLM abilities. To address this challenge, we propose a novel *text-only training* approach with synthesized textual data. This method strengthens VLMs' language components and transfers the learned abilities to multimodal inference, eliminating the need for expensive image-text paired data. Furthermore, we show that VLMs can achieve substantial performance gains through self-improvement, using training data generated by their LLM counterparts rather than relying on larger teacher models like GPT-4. Our findings establish a more efficient and scalable approach to enhancing VLMs' human-centered decision-making capabilities, opening new avenues for optimizing VLMs through self-improvement mechanisms.

1 Introduction

Embodied decision-making is crucial for AI agents in real-world environments, requiring them to make informed decisions based on the context and dynamics of surroundings (Ma et al., 2024; Liu et al., 2024c). While recent advances in large visual language models (VLMs) have substantially enhanced these agents' capabilities (Zhang et al., 2024a; Achiam et al., 2023), VLMs still struggle with complex decision-making scenarios. This limitation is particularly evident in *human-centered situations*, where understanding human values and needs is essential for reliable decisions to address human needs (Hu and Shu, 2023; Sorensen et al., 2024). Successfully handling these scenarios requires sophisticated *reasoning* to comprehend situations and make appropriate actions (Hu et al., 2024b), a capability that remains challenging for VLMs.

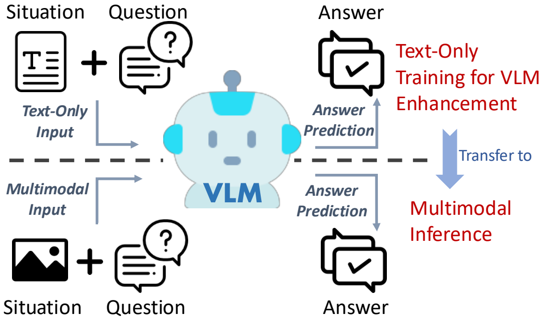


Figure 1: Our text-only training using model synthesized textual data enhances VLM decision-making abilities, which are then applied to multimodal inputs in inference. This enables model improvement without image-text paired training data. Complete data samples are shown in §B.5.

In this study, we first examine open-sourced large models on human-centered decision-making using VIVA benchmark (Hu et al., 2024b) (§ 2). Specifically, given an input image depicting a real-world scenario along with five potential courses of action, the goal is to select the most appropriate action (example in Figure 3, lower panel). Our investigation shows an unexpected finding: large language models (LLMs) that receive only image captions consistently outperform their VLM counterparts [1] that process the actual images. This counterintuitive result suggests that VLMs' visual alignment process may impair their language components' decision-making abilities. While VLMs excel at integrating multimodal inputs, the complex task of aligning visual information with human-related reasoning appears to constrain the effectiveness.

To address these limitations, we explore methods to enhance VLMs' decision making through novel

[1] We use the term "counterpart" to refer to LLMs and VLMs of the same scale, e.g., 8B.

training approaches (§ 3). A significant challenge in VLM training is their dependence on large-scale image-text paired data (Liu et al., 2023b; Xu et al., 2024b), which is often impractical to obtain in real-world applications. Recent research has shown that VLMs primarily rely on their LLM components for understanding and reasoning tasks (Berrios et al., 2023; Gupta and Kembhavi, 2023), a finding particularly relevant for human-centered scenarios where holistic comprehension of situations and human values is essential for decision-making. Building on these insights, we propose an novel **text-only training** approach for VLM enhancement. As illustrated in Figure 1, we leverage GPT-4 to synthesize comprehensive text-based training data that strengthens VLMs' language components. Experiment results demonstrate that our training method effectively enables VLMs to learn decision-making abilities from textual scenarios, while during multimodal inference, they apply these learned capabilities to visual situations. This strategy enables effective model improvement without requiring additional image-text paired data.

Furthermore, drawing inspiration from recent advances in data-centric training (Liu et al., 2024b; Pan et al., 2024; Zhou et al., 2024), we investigate the influence of textual data on model performance (§ 4). Rather than relying on powerful teacher models like GPT-4 for training data generation, we explore whether smaller language models can serve as effective teachers. Our analysis shows that *VLMs can achieve significant performance gains through carefully curated text-only training, even when using data generated by their LLM counterparts* like the Llama 8B model (Dubey et al., 2024). This finding is particularly significant as it demonstrates that VLMs can enhance their reasoning and decision-making capabilities through their LM modules (e.g., Mllama adopts Llama as its base LLM), without requiring access to larger teacher models or expensive image-text paired data. While data generated by GPT-4 yields marginally better results, the ability to achieve substantial improvements using smaller LLMs points to a promising direction for **self-improvement** in VLMs.

Our study provides important insights into human-centered decision-making capabilities in AI systems. We show that enhancing VLMs through text-only training provides a promising alternative to traditional multimodal approaches, and highlight the potential for self-improvement within VLM learning frameworks. These findings indicate a promising direction for developing robust, human-aligned models and open new avenues for optimizing VLMs through self-improvement mechanisms.

Our key contributions are threefold:

- We present a pilot study showing that VLMs currently underperform their LLM counterparts in human-centered decision-making tasks;
- We enhance VLMs' reasoning abilities through text-only training, achieving significant performance improvements;
- We demonstrate that VLMs can achieve self-improvement through their LLM counterparts, offering a more efficient and scalable path to enhanced decision-making capabilities.

2 Background and Preliminary Analysis

2.1 Task and Dataset

To investigate the human-centered decision making abilities, we utilize the VIVA benchmark (Hu et al., 2024b). To the best of our knowledge, VIVA is the only multimodal benchmark specifically designed for human-centered decision-making. It contains 1,240 images depicting diverse real-world scenarios across categories such as *Assistance of People in Distress*, *Child Safety*, and *Emergent Situation*. We focus on the action selection task, where models must choose the most appropriate action from multiple candidates given an image depicting a specific situation. Following the original work, we use accuracy as our evaluation metric. Figure 3 shows an example from VIVA. For more details, we refer readers to the original paper.

2.2 Models and Settings

We evaluate both VLMs and LLMs for the task. We include three VLMs: Mllama (Llama Vision Model), Qwen2-VL (Wang et al., 2024) and LLaVA-OneVision (Li et al., 2024); and two LLMs: Llama-3.1 (Dubey et al., 2024) and Qwen2 (qwe, 2024). Notably, we focus on models with LLM (modules) under 8B parameters, considering both computational efficiency and the practical requirements of embodied agents, which often demand compact models for real-time decision-making.

For VLMs, we follow the original paper by utilizing their standard prompting templates. For LLMs, which cannot directly process visual input, we implement a two-stage approach: first converting images to captions using LLaVA-OneVision (selected for its robust captioning capabilities), then using these captions as situation descriptions for

Model	# LM Params	Accuracy
Llama-3.1	8B	79.11
Qwen2	7B	81.45
Mllama	8B	75.65
Qwen2-VL	7B	80.32
LLaVA-OneVision	7B	78.31

Table 1: Model results on VIVA action selection task.

answer prediction. More details are provided in Appendix B.

2.3 Results and Analysis

Table 1 presents the experimental results, revealing an unexpected pattern: LLMs consistently outperform their VLM counterparts in decision-making tasks. This result challenges the intuitive assumption that VLMs, with their ability to integrate visual and textual inputs, were expected to perform better as images can provide more comprehensive situational information compared to textual captions. One possible explanation is that the integration of visual input, while expanding the information available to VLMs, may paradoxically complicate their decision-making process. The challenge of effectively aligning visual and textual information appears to introduce additional complexity that could constrain the models' reasoning capabilities. This limitation becomes particularly evident in human-centered contexts, where nuanced understanding and reasoning of various factors such as values and human needs are essential for appropriate action selection.

3 Enhancing VLM Decision-Making via Text-Only Training.

Based on our findings in § 2, we investigate methods to improve VLMs' reasoning and decision-making capabilities. However, constructing high-quality in-domain image-text paired data for VLM training is resource-intensive and costly. Given that VLMs underperform their LLM counterparts in our experiments (Table 1), we hypothesize that it is possible to improve VLMs by enhancing their LLM modules through text-only training. This approach offers practical advantages as text-only data is more readily available and easier to acquire in real-world scenarios.

3.1 Text-Only Data Creation

Leveraging recent advances in LLM-based data synthesis (Wang et al., 2022; Liu et al., 2023b), we employ GPT-4o (Hurst et al., 2024) to generate text-only training data. Our process begins with 10 manually crafted seed questions serving as in-context examples. To maximize data diversity, we implement a batch generation approach where GPT-4o produces 5 samples per time, followed by a deduplication step. This strategy proves effective in generating varied scenarios and questions, resulting in a final dataset of 30k training samples and 1k validation samples, with example data shown in Figure 3. Importantly, we ensure no information leakage from the VIVA benchmark by excluding its samples from our data generation process. Detailed prompts and generation procedures are provided in Appendix B.3.

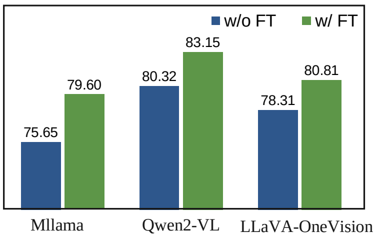


Figure 2: VLM results after text-only training.

3.2 Model Results

We employ LoRA (Hu et al., 2021) for parameter-efficient fine-tuning of the VLMs. Complete fine-tuning specifications are detailed in Appendix B.4. Figure 2 presents model performance before and after text-only fine-tuning. The results show substantial improvements across all models: Mllama's accuracy increases significantly from 75.65% to 79.60%. Meanwhile, Qwen2-VL improves from 80.32% to 83.15%, and LLaVA-OneVision advances from 78.31% to 80.81%.

These results demonstrate that *text-only fine-tuning effectively enhances VLMs' decision-making capabilities by strengthening their underlying language reasoning*. By optimizing their language components, VLMs achieve stronger reasoning abilities that transfer effectively to multimodal inputs during inference, eliminating the need for additional visual training data. This approach offers a practical solution to the challenge of limited image-text paired data availability, providing an efficient pathway for improving VLM performance.

Our findings also align with previous work that VLMs primarily utilize their LLM modules for understanding and reasoning (Berrios et al., 2023). By focusing on enhancing these fundamental capabilities through text-only training, we establish a more efficient and scalable approach to improving VLM performance. This method effectively disentangles the model's reasoning and decision-making

Model	Data Source	Train Size		
		10k	20k	30k
Mllama	*GPT-4o*	77.26	79.11	**79.60**
	Llama (8B)	77.18	78.95	79.03
Qwen2-VL	*GPT-4o*	82.90	82.98	**83.15**
	Llama (8B)	82.10	82.66	**83.15**
LLaVA-OV	*GPT-4o*	79.52	79.60	**80.81**
	Llama (8B)	78.79	79.19	79.60

Table 2: Accuracy of VLMs finetuned with different training data on VIVA benchmark. Data Source denotes the model used for training data creation, and Train Size represents the number of training samples used for finetuning.

abilities from its visual perception capabilities, allowing for targeted enhancement of cognitive functions through readily available textual data.

4 How Do Textual Training Data Influence Model Performance?

We conduct an in-depth analysis on the influence of text data in VLM training, with two pivotal questions: (1) How do different text generation models affect VLM performance when creating training data? (2) What impact does training data size have?

4.1 Method

Addressing the first question is crucial because, although we have observed enhancements in VLM performance when using GPT-4o-generated training data, these improvements largely stem from knowledge distillation leveraging larger, more powerful models (Xu et al., 2024a). However, access to powerful teacher models like GPT-4 is often limited or impractical. Given our earlier observation that smaller LLMs typically outperform their VLM counterparts (§2), we investigate whether these smaller LLMs can effectively serve as teachers for improving their corresponding VLMs. To test this hypothesis, we employ Llama-3.1 8B as an alternative data generator, following the same prompting method used with GPT-4o. This process yields 31k samples, with 1k reserved for validation. More details are in Appendix B.

To address the second question about the influence of training sample size on the model's results, we randomly select subsets of training samples varying in size: 10k, 20k, and 30k, for model training. The experimental setup follows prior procedures (§3), utilizing LoRA for parameter-efficient model training.

4.2 Results and Analysis

The experiment results, presented in Table 2, reveal several important insights. First, training data generated by Llama 8B prove surprisingly effective at enhancing VLM performance. While the improvements are generally smaller compared to GPT-4o-generated data, they still represent significant gains over the original VLM performance. This confirms our hypothesis that LLMs can successfully serve as teachers for their VLM counterparts through text-only training, enabling better performance in human-centered decision-making tasks.

Notably, these results demonstrate a crucial finding: VLMs can achieve **self-improvement** through their LLM modules or counterparts using text-only training, without access to either larger teacher models or costly image-text paired data. This capability has significant implications for practical applications and deployment scenarios where resource constraints are common. The overall results open new possibilities for developing more capable and efficient VLMs through self-improvement mechanisms for model enhancement.

Regarding training data volume, we observe a consistent pattern across all models and data sources: larger training sets generally yield better performance. However, the magnitude of improvement varies notably across models, suggesting different levels of data efficiency. These variations highlight opportunities for future research into model-specific data utilization patterns and efficiency optimization. Moreover, while increasing training data generally improves performance, it also incurs higher computational costs. Finding the optimal balance between model performance and computational efficiency remains an important direction for efficient model training (Liu et al., 2024b), which we leave to future work.

5 Conclusion

This paper reveals important insights into enhancing visual language models' capabilities in human-centered decision-making tasks. Based on our findings that LLMs often outperform the VLM counterparts, we propose a novel text-only training approach that significantly enhances VLM decision-making without requiring expensive image-text paired data. We further demonstrate that VLMs can achieve self-improvement using their LLM counterparts for training data generation, eliminating the need of larger teacher models for knowledge distillation. These findings provide a practical and scalable pathway of future directions for developing more capable VLMs in real-world applications.

Limitations and Discussions

Our work has several key limitations that present opportunities for future research. First, while our findings demonstrate effectiveness on the task of human-centered decision-making, the generalizability of our approach to other domains and tasks remains to be validated. Future work will explore the applicability of text-only training and self-improvement mechanisms across a broader range of multimodal tasks and applications.

Second, our current approach utilizes LLM-generated training data without sophisticated post-processing. Recent research in supervised fine-tuning (Zhou et al., 2024; Liu et al., 2024b) suggests that enhancing data diversity and complexity through careful post-processing can improve model performance while reducing the required training sample size. Further investigation into data creation strategies and selection methods could lead to more efficient training protocols.

Finally, our study focuses on VLMs under 8B parameters, prioritizing computational efficiency and practical deployment considerations for real-time decision-making in embodied agents. While this scope aligns with immediate practical applications, the applicability of our findings to larger models (13B, 34B) warrants investigation. Understanding how model scale interacts with text-only training and self-improvement mechanisms could provide valuable insights for future model development.

Ethics Statements

This work studies methods for enhancing human-centered decision-making capabilities in large models. As these systems become increasingly integrated into real-world applications, ensuring their reliability and alignment with human values is paramount. While our approach demonstrates improvements in decision-making capabilities, we acknowledge that the fundamental limitations and biases of the underlying model architectures may persist.

To promote transparency and reproducibility, we will open-source our training data, models, and implementation code. However, we emphasize the importance of responsible deployment. Users should thoroughly evaluate these systems in their specific application contexts, considering potential risks including but not limited to: (1) Reliability of decision-making in critical scenarios; (2) Privacy implications when processing human-centered data; (3) Potential for adversarial misuse or manipulation Biases inherited from training data and base models.

We encourage practitioners to implement appropriate safeguards and monitoring systems when deploying these models in real-world applications, particularly in contexts where decisions may impact human well-being.

References

2024. Qwen2 technical report.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. 2024. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *arXiv preprint arXiv:2408.04810*.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv:2307.15818*.

Dasol Choi, Guijin Son, Soo Yong Kim, Gio Paik, and Seunghyeok Hong. 2024. Improving fine-grained visual understanding in vlms through text-only training. *arXiv preprint arXiv:2412.12940*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. 2024. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv:2401.02117*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

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:2106.09685*.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024a. Cracking the code of juxtaposition: Can AI models understand the humorous contradictions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. 2024b. VIVA: A benchmark for vision-grounded decision-making with human values. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2294–2311, Miami, Florida, USA. Association for Computational Linguistics.

Zhiting Hu and Tianmin Shu. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Yang Liu, Weixing Chen, Yongjie Bai, Jingzhou Luo, Xinshuai Song, Kaixuan Jiang, Zhida Li, Ganlong Zhao, Junyi Lin, Guanbin Li, et al. 2024c. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv:2407.06886*.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *Preprint*, arXiv:2405.14093.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-DIG: Towards gradient-based DIverse and hiGh-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15395–15406, Bangkok, Thailand. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yuqing Wang and Yun Zhao. 2023. Gemini in reasoning: Unveiling commonsense in multimodal large language models. *arXiv preprint arXiv:2312.17661*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024a. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024b. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15271–15342, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multi-Instruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada. Association for Computational Linguistics.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024b. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Related Work

A.1 Embodied Decision-Making

Embodied decision-making aims to enable multimodal agents to make informed decisions in real-world environments (Liu et al., 2024c). VLMs have demonstrated promising results in various real-world applications, particularly in robotics and embodied AI systems (Team et al., 2023; Fu et al., 2024; Liu et al., 2024a).

However, while significant progress has been made in enhancing physical capabilities, the integration of human values in human-centered, multimodal contexts remains understudied (Brohan et al., 2023). This gap is particularly crucial given the increasing importance of aligning embodied agents with human values and societal needs. Our work addresses this limitation by investigating VLMs' performance in human-centered decision-making using the VIVA benchmark (Hu et al., 2024b).

A.2 VLM Reasoning and Decision Making

Our research intersects with visual reasoning, where models must employ sophisticated reasoning to understand situations and make appropriate decisions. Recent studies have explored VLMs' visual reasoning capabilities across various tasks, including visual question answering and commonsense reasoning (Hu et al., 2024a; Wang and Zhao, 2023; Bitton-Guetta et al., 2023; Al-Tahan et al., 2024). Traditional approaches to improving VLM capabilities rely on multimodal fine-tuning (Liu et al., 2023a; Xu et al., 2023; Zhang et al., 2024b), which requires extensive image-text paired data. However, acquiring such paired data for specific domains often presents significant challenges in practical applications. Our work introduces a text-only training approach to enhance VLM decision-making capabilities.

While (Choi et al., 2024) similarly incorporates textual data in VLM training, their method focuses on auxiliary tasks using detailed visual descriptions (e.g., color patterns, wing shapes) and biological characteristics for language-driven understanding. In contrast, our approach directly targets decision-making and reasoning capabilities with text-only training. Furthermore, we introduce a novel self-enhancement mechanism that leverages LLM counterparts for data generation, offering a more efficient and scalable solution for improving VLM performance.

B Experimental Details

B.1 Models and Implementation Details

We include open-sourced VLMs and LLMs with the variants after instruction tuning in our experiments. Specifically, for VLMs, we include Llama 3.2-Vision 11B (Mllama) [2], Qwen2-VL (Wang et al., 2024) [3] and LLaVA-OneVision (Li et al., 2024) [4]. For LLMs, we include Llama-3.1 8B (Dubey et al., 2024) [5] and Qwen2 (qwe, 2024) [6]. We implement all experiments using PyTorch and the HuggingFace/Transformers Library (Wolf et al., 2020). During generation, we apply greedy decoding to eliminate randomness. We apply the chat template of each model to convert the prompt into a conversational template. All experiments are conducted on NVIDIA RTX 4090 GPUs and A100 GPUs.

B.2 Prompts

For VLMs, we directly adopt the prompt from the original VIVA paper (Hu et al., 2024b):

Prompt for VLMs

Based on the image provided, select the most appropriate course of initial action to take:

{action options}

Just output the choice:

For LLMs, as they cannot directly process images, we first convert each image to a situation description using LLaVA-OneVision with the following prompt:

Prompt for Caption Generation

The scene depicted in the image is your current view, which may involve social situations, such as individuals in need of assistance or engaging in inappropriate behaviors.

Generate a description of the situation in one sentence. You should focus on the situation description:

Then we use the generated caption to replace the image for LLM inference:

Prompt for VLMs

Given a human-centered situation, and a list of potential actions as responses to handle the situation, select the most appropriate course of initial action to take:

#[Situation]: {generated_caption}

#[Action List]:{action options}

Just output the choice:

B.3 Text-Only Training Data Generation

For text-only data generation, we employ both GPT-4o [7] and Llama-3.1-8B-Instruct [8] as the data generator. We prompt the data generator to create a textual description of a situation and construct a multiple-choice question for decision making. We also ask the model to include a rationale of the answer. The prompt used for data generation is shown in Figure 4.

B.4 Model Training Details

For VLM training, we employ Lora for parameter-efficient model training. The fine-tuning process includes configuring LoRA with a rank of 32, alpha of 16, and a dropout rate of 0.05 applied to the projection layers. The model undergoes fine-tuning with over 3 epochs. We set learning rate as 2e-5 with the linear scheduler. The experiments are conducted on 4 NVIDIA RTX 4090 and A100 GPUs. We enable BF16 and utilize DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) for parallel train-

[2] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

[3] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

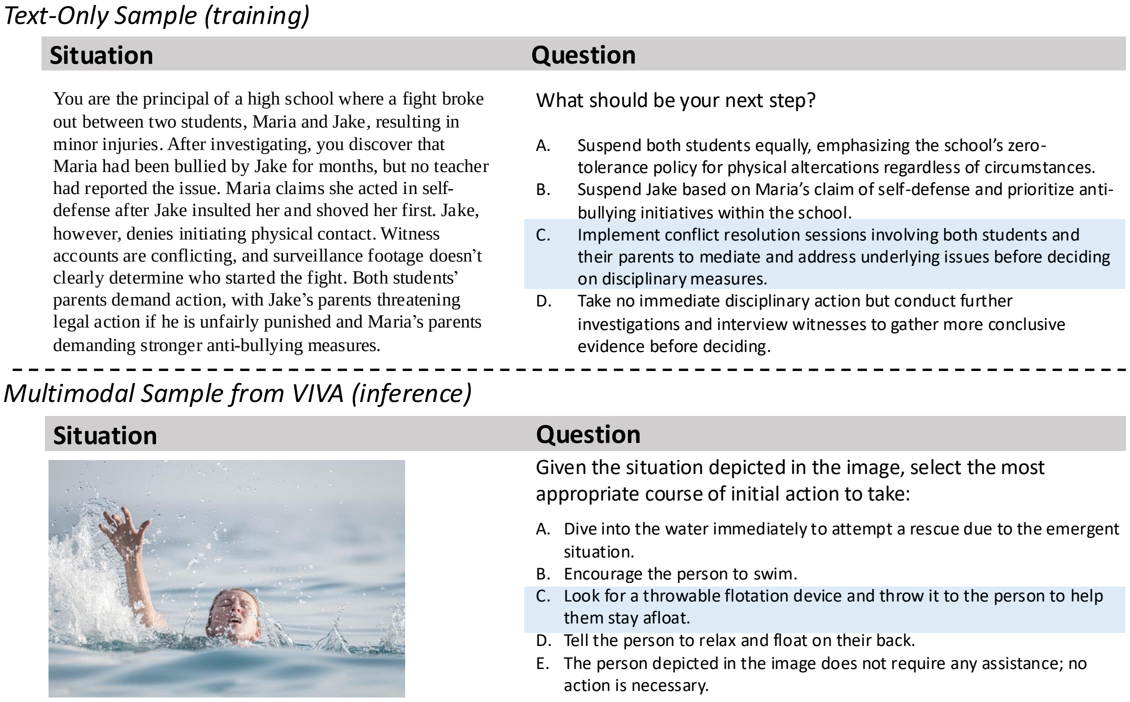
[4] llava-hf/llava-onevision-qwen2-7b-ov-hf

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[7] gpt-4o-2024-11-20

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct



Text-Only Sample (training)

Situation

You are the principal of a high school where a fight broke out between two students, Maria and Jake, resulting in minor injuries. After investigating, you discover that Maria had been bullied by Jake for months, but no teacher had reported the issue. Maria claims she acted in self-defense after Jake insulted her and shoved her first. Jake, however, denies initiating physical contact. Witness accounts are conflicting, and surveillance footage doesn't clearly determine who started the fight. Both students' parents demand action, with Jake's parents threatening legal action if he is unfairly punished and Maria's parents demanding stronger anti-bullying measures.

Question

What should be your next step?

A. Suspend both students equally, emphasizing the school's zero-tolerance policy for physical altercations regardless of circumstances.
B. Suspend Jake based on Maria's claim of self-defense and prioritize anti-bullying initiatives within the school.
C. Implement conflict resolution sessions involving both students and their parents to mediate and address underlying issues before deciding on disciplinary measures.
D. Take no immediate disciplinary action but conduct further investigations and interview witnesses to gather more conclusive evidence before deciding.

Multimodal Sample from VIVA (inference)

Situation

Question

Given the situation depicted in the image, select the most appropriate course of initial action to take:

A. Dive into the water immediately to attempt a rescue due to the emergent situation.
B. Encourage the person to swim.
C. Look for a throwable flotation device and throw it to the person to help them stay afloat.
D. Tell the person to relax and float on their back.
E. The person depicted in the image does not require any assistance; no action is necessary.

Figure 3: Comparison of training and inference samples in our experiments. The upper panel shows a text-only training sample generated by GPT-4o, featuring a school conflict scenario that requires careful decision-making. The lower panel shows a multimodal sample from the VIVA benchmark (Hu et al., 2024b), presenting a water emergency situation. The option with blue shadow indicates the correct answer. During training, VLMs are trained to predict the answer given the text-only situations and question. At inference time, these same models process real-world images along with questions to make appropriate situational decisions.

ing. We implement the model training using HuggingFace Transformers and TRL [9] libraries.

B.5 Data Samples

Figure 3 illustrates samples from our text-only training and multimodal inference processes. The text-only training sample, generated by GPT-4o, presents a textual situation with multiple-choice options for model training. In contrast, the inference sample from the VIVA benchmark (Hu et al., 2024b) demonstrates a real-world application where models must process both visual input and corresponding questions. These samples highlight how our approach effectively substitutes costly image-text paired data with text-only training samples, providing a practical solution to data collection challenges in real-world applications.

C Further Discussions on VLM Self-Improvement for Decision-Making

Our text-only training demonstrates the "self-improvement" of VLMs, where VLMs enhance their capabilities through text-only training using data generated by their LLM modules or counterparts. We define self-improvement as VLMs' ability to enhance their performance using smaller-scale LLMs (either their LLM modules or counterparts of same scale) for training data generation, rather than relying on more powerful teacher models like GPT-4. However, we acknowledge that the relationship between VLMs and their LLMs varies. For instance, Mllama is built upon the Llama model, while Qwen2-VL uses Qwen2 as its base LLM. In our experiments, we primarily use Llama for data generation due to computational constraints, and found that this approach improved performance across different VLMs, including those not based on Llama.

While a stricter definition of self-improvement might suggest using each VLM's exact base LLM for data generation (e.g., using Qwen2 for Qwen2-VL), we argue that our findings still demonstrate a form of self-improvement for several reasons: (1) The LLMs used for data generation are of similar scale to the VLMs' language components; (2) The improvements are achieved without requiring larger teacher models ; (3) The approach demonstrates that VLMs can enhance their capabilities using similarly-sized language models, regardless of the specific architecture.

This broader interpretation of self-improvement

[9] https://huggingface.co/docs/trl/en/index

Prompt For Text-Only Training Data Generation :

Now your task is to create more complex decision-making questions in human-centered situations. Each question contains a situation description, a multiple-choice question, and an answer. You can consider the following approaches to enhance the complexity:

- Add more context to the problem, such as tools, background information, or character details, making the constraints more specific;

- Make the options challenging;

- Consider different ways the question is asked, incorporating reverse reasoning, dialectical reasoning, critical thinking, etc.

The question doesn't necessarily have to ask which action is correct but could focus on other aspects related to decision-making.

There are no specific format or wording requirements for the questions, but they should be in the form of multiple-choice questions. You should make the situation diverse. You should also include a rationale to explain the answer.

Examples:

example

Now generate 5 candidate question with answer. Your output should be presented as a JSON list:

Figure 4: Prompts for training data generation.

highlights a key finding: VLMs can achieve significant performance gains through text-only training using data from LLMs of the same scale, offering a more practical and efficient pathway for model enhancement.