Hierarchical Approaches for Domain-Specific Image Captioning: Classification, Distillation, and Optimization

Anonymous ACL submission

Abstract

This paper presents a novel hierarchical approach to improve image captioning, focusing on enhancing the accuracy, contextual appropriateness, and linguistic diversity of generated captions. We address key limitations of existing multimodal models, including inaccurate object relationships, missed details, and poor domain-specific understanding. Our method incorporates three main components: a classification-guided prompting system that utilizes domain-specific knowledge, a knowledge distillation framework that transfers captioning capabilities from GPT-4o to the LLaVA model, and an iterative Direct Preference Optimization (DPO) approach that refines caption quality. Extensive experiments demonstrate that our approach outperforms existing methods, achieving near-GPT-4o performance while maintaining computational efficiency. Additionally, we release a high-quality dataset of 9,840 image-caption pairs across 18 categories, providing valuable resources for future research in domain-specific image captioning.

1 Introduction

Image captioning, the task of automatically generating natural language descriptions for images, has emerged as a crucial indicator of models' ability to understand visual and language information effectively (Mokady et al., 2021; Li et al., 2023). Despite significant advances in this field, current state-of-the-art models still face substantial challenges in generating accurate, contextually appropriate, and comprehensive descriptions.

Recent multimodal models, while demonstrating impressive capabilities across various tasks (Li et al., 2023; Liu et al., 2024), exhibit notable limitations in image captioning. These limitations manifest in several aspects: inaccurate object relationship descriptions, missed subtle but important details, and inadequate ability to express domain-specific concepts. Moreover, traditional evaluation metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) often fail to capture these nuanced aspects of performance, potentially leading to misleading assessments of model capabilities.

To address these limitations, we propose a novel hierarchical approach that combines classification-guided prompting, knowledge distillation, and iterative preference optimization. Our method consists of three key components:

- A classification-first strategy that categorizes images into semantic classes, enabling the use of class-specific prompting templates for more precise and contextually relevant caption generation.
- A knowledge distillation framework that leverages GPT-4o's superior captioning abilities to create high-quality training pairs, which are then used to fine-tune a more efficient LLaVA model[1].
- An iterative Direct Preference Optimization (DPO) approach that continuously refines the model's output quality by learning from automatically generated preference pairs using a Pixtral (Agrawal et al., 2024) model for evaluation.

Through comprehensive experiments, we demonstrate that this hierarchical approach significantly improves captioning quality across multiple dimensions. The classification-guided prompting ensures domain-specific accuracy, while the knowledge distillation from GPT-4o provides rich linguistic diversity. Furthermore, the iterative DPO refinement helps align the model's outputs with quality standards, leading to more natural and contextually appropriate descriptions.

[1] LLaVA serves as our base multimodal model due to its strong vision-language capabilities and computational efficiency.

Our extensive evaluation shows that the proposed approach achieves significant improvements over existing methods. Specifically, our model achieves a Pixtral score of 4.312 after DPO optimization, approaching the performance of GPT-4o (4.473) and GPT-4o-mini (4.293), while substantially outperforming the base LLaVA model (3.598). Human evaluation strongly correlates with our automatic metrics (Pearson correlation: 0.779), validating the effectiveness of our evaluation framework. These results demonstrate the success of combining semantic classification, knowledge distillation, and preference optimization in improving image captioning quality.

Our main contributions are summarized as follows:

- We propose a novel classification-guided prompting strategy that leverages domain-specific knowledge for more accurate and contextually appropriate caption generation.
- We present an effective knowledge distillation pipeline that efficiently transfers captioning capabilities from GPT-4o to a more practical LLaVA model.
- We develop an iterative DPO framework that uses automated evaluation for continuous model improvement, demonstrating consistent performance gains over multiple iterations.
- We release a comprehensive dataset of 9,840 high-quality image-caption pairs across 18 diverse categories, where each image is paired with a GPT-4o-generated caption. This curated dataset provides valuable training resources for future research in domain-specific image captioning.
- We introduce a fine-tuned DPO version of the LLaVA model, which demonstrates strong image captioning capabilities, further enhancing the practical application of the model.

2 Related Works

2.1 Image Captioning Models

Recent advances in image captioning have been largely driven by the development of large multimodal models. Traditional approaches relied on encoder-decoder architectures (Xu et al., 2016; Anderson et al., 2018), while more recent work has focused on end-to-end multimodal training. Models like BLIP (Li et al., 2023) and LLaVA (Liu et al., 2024) have demonstrated impressive capabilities in understanding and describing visual content. However, these models often struggle with domain-specific details and contextual accuracy, motivating our classification-guided approach.

2.2 Caption Evaluation Metrics

The evaluation of image captioning systems has evolved through several generations:

Reference-based Metrics Traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015) evaluate captions by comparing them with human-written references. While widely used, these metrics often fail to capture semantic accuracy and contextual appropriateness (Anderson et al., 2016).

Learning-based Metrics More recent approaches like BERTScore (Zhang* et al., 2020) and CLIP-Score (Hessel et al., 2022) leverage pre-trained models to assess caption quality. These metrics show better correlation with human judgments but still lack explainability and domain-specific understanding.

LLM-based Evaluation The emergence of large language models has introduced new possibilities for caption evaluation. Recent work has shown that LLMs can provide more nuanced and context-aware evaluations (Liu et al., 2023). Our work builds on this direction by using the Pixtral model for automated evaluation and preference learning.

2.3 Knowledge Distillation and Model Alignment

Knowledge distillation has proven effective in transferring capabilities from larger to smaller models (Hinton et al., 2015). In the multimodal domain, recent work has shown success in distilling vision-language capabilities (Li et al., 2023). Our approach extends this to image captioning by distilling knowledge from GPT-4o to a more efficient LLaVA model.

Direct Preference Optimization DPO has emerged as a powerful technique for aligning language models with human preferences (Rafailov et al., 2024). Recent work has demonstrated its effectiveness in improving text generation quality (Liu et al., 2023). We adapt this approach to image

captioning by using automated evaluations to guide the optimization process.

2.4 Category-Specific Generation

The idea of leveraging category information for improved generation has been explored in various contexts. Previous work has shown that domain-specific prompting can improve text generation quality (Liu et al., 2024). Our work extends this concept to image captioning by introducing a classification-guided prompting strategy that adapts to different visual domains.

3 Method

Our approach consists of three main components: (1) a classification-guided prompting system, (2) a knowledge distillation pipeline from GPT-4o, and (3) an iterative DPO optimization process. Figure 1 illustrates the complete architecture of our proposed method. In this section, we describe each component in detail.

3.1 Classification-Guided Prompting

The effectiveness of image captioning heavily depends on the model's ability to recognize and describe domain-specific content accurately. To address this challenge, we propose a classification-first approach that leverages a fine-tuned large multimodal model (LMM) for image classification, followed by category-specific prompting for caption generation.

Classification Model Training We fine-tune a LLaVA model to serve as our classifier. The training process consists of several key steps:

1) **Category Definition**: We define 18 comprehensive categories $\mathcal{C}$ covering most common image scenarios:

- Scene types: Animal, Architecture, Art, City, Landscape
- Daily activities: Daily_Life, Food, Love, Medical, Sports
- Technical: Autonomous_Driving, Smart_Cities, Transportation
- Visual content: Cartoon, Chart, Person, Plant
- Others: Other

2)**Training Data Collection**: For each category $c \in \mathcal{C}$, we collect images from multiple sources, including Flickr, JAAD (Rasouli et al., 2017, 2018), ScienceQA (Lu et al., 2022), ChartQA (Masry et al., 2022), etc.:

$$D_{\text{train}} = (I_i, c_i)_{i=1}^{N} \tag{1}$$

where I_i represents an image and c_i its corresponding category label.

3) **Model Fine-tuning**: We fine-tune the LLaVA model using our collected dataset:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{N} \log p_\theta(c_i|I_i) \tag{2}$$

where θ represents the model parameters.

Classification Process Given a new image I, the classification is performed as:

$$c = f_{\text{LLaVA}}(I) \in \mathcal{C} \tag{3}$$

where f_{LLaVA} is our fine-tuned LLaVA classifier.

Structured Prompt Design For each category c, we design a specialized prompting template P_c to guide the caption generation process. Representative examples of our prompting templates are provided in Appendix A, and the complete set is available in our repository.

Caption Generation The final caption generation process involves:

$$C_{\text{final}} = g_{\text{LLaVA}}(I, P_c) \tag{4}$$

where g_{LLaVA} is another LLaVA model fine-tuned on GPT-4o distilled data (detailed in Section 3.2).

3.2 Knowledge Distillation from GPT-4o

As mentioned in Section 3.1, our caption generation model g_{LLaVA} is fine-tuned through knowledge distillation from GPT-4o. We leverage the same dataset used for classifier training to create high-quality caption pairs.

High-Quality Caption Generation For each image I in our training dataset, we use GPT-4o to generate captions:

$$C_{\text{GPT-4o}} = g_{\text{GPT-4o}}(I, P_c) \tag{5}$$

where $g_{\text{GPT-4o}}$ represents the GPT-4o model's caption generation function.

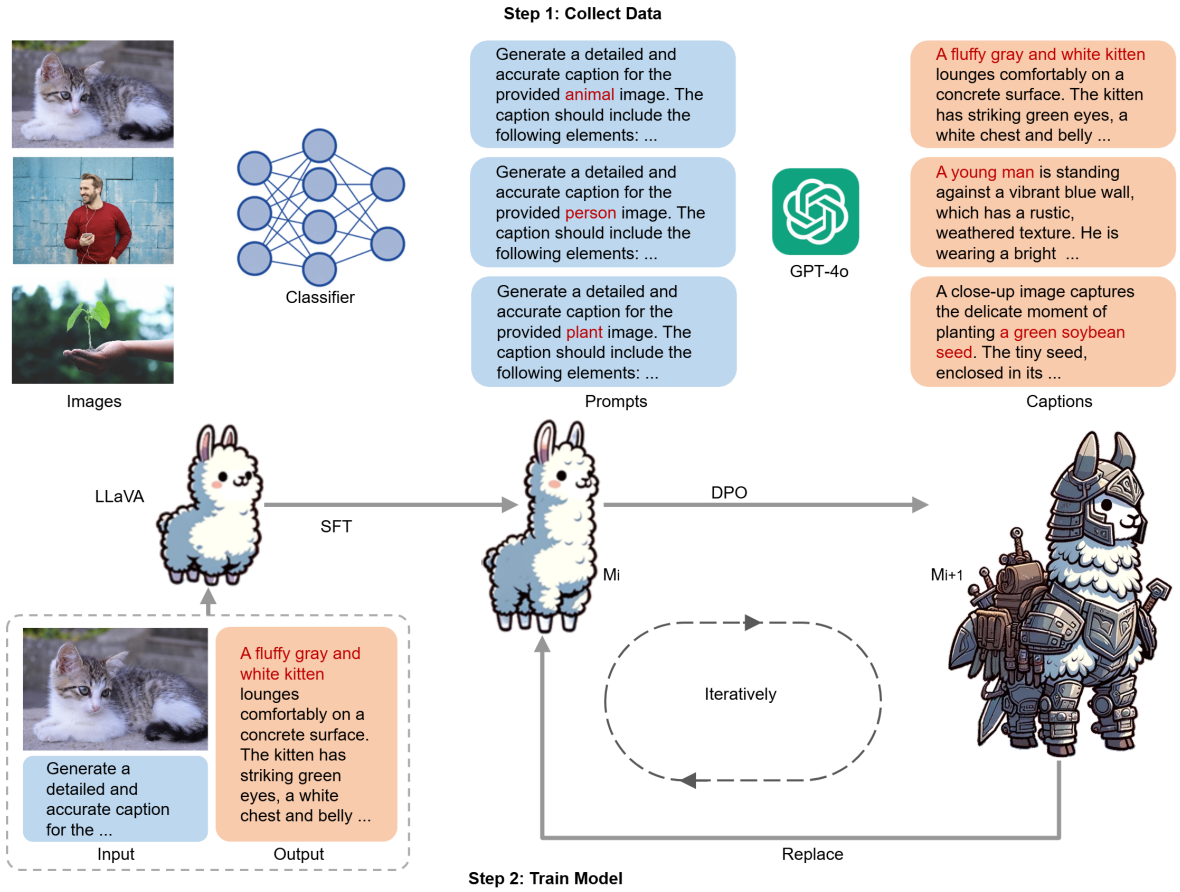

Figure 1: method

Training Data Creation We create training pairs by combining:

$$D_{\text{distill}} = \{(I_i, P_{c_i}, C_{\text{GPT-4o},i})\}_{i=1}^{N} \quad (6)$$

where N is the total number of images across all categories.

LLaVA Fine-tuning We fine-tune the LLaVA model using these high-quality training pairs with both full parameter fine-tuning and LoRA approaches. For both versions, the supervised learning objective is:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{N} \log p_\theta(C_{\text{GPT-4o},i} | I_i, P_{c_i}) \quad (7)$$

where θ represents the parameters of the LLaVA model. The detailed training configurations for both approaches are provided in Appendix B.

3.3 Iterative DPO Optimization

To further improve the quality of generated captions, we implement an iterative Direct Preference Optimization (DPO) process using a Pixtral model for preference scoring. We leverage both full-parameter and LoRA fine-tuned models to generate diverse candidate captions.

Caption Generation and Scoring For each image, we use both the full-parameter and LoRA fine-tuned models to generate five captions each, resulting in ten candidate captions per image. These captions are then evaluated using the Pixtral model, which assigns a score between 1 and 5 to each caption following a structured evaluation prompt (detailed in Appendix C).

Preference Pair Selection From the ten candidates, we identify the highest and lowest scoring captions. If their score difference exceeds 2 points (on the 1-5 scale), we use this pair for DPO training. This threshold ensures we only learn from pairs with significant quality differences.

DPO Training The DPO objective is defined as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(I,C_w,C_l)} \left[\log \sigma \left(\beta(r_\theta(C_w) - r_\theta(C_l))\right)\right] \quad (8)$$

where $r_\theta(C)$ is the reward score for caption C, β is temperature parameter, and σ is sigmoid function.

Iterative Refinement The DPO process consists of:

1. Generate ten captions per image
2. Score captions using Pixtral
3. Select pairs with significant score differences
4. Update model using DPO objective
5. Repeat until convergence

The detailed training configurations and implementation details are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

Dataset We collect and curate a new dataset from multiple sources including Flickr, JAAD, ScienceQA, ChartQA, etc., covering 18 diverse categories including animal, architecture, art, autonomous driving, cartoon, chart, city, daily life, food, landscape, love, medical, other, person, plant, smart cities, sports, and transportation. The dataset consists of 9,840 training images and 796 validation images, with approximately balanced distribution across categories. All images are manually verified to ensure quality and correct categorization. The validation set maintains a similar category distribution as the training set to ensure fair evaluation.

For each image in the training set, we generate a high-quality caption using GPT-4o following our category-specific prompting strategy. These captions are carefully crafted to capture domain-specific details and contextual information relevant to each category. This results in a rich dataset where each image is paired with a detailed, contextually appropriate caption that can serve as high-quality training data for future research.

The dataset is designed with several key features:

- **Category Balance**: Training and validation sets maintain approximately balanced distribution across all categories, ensuring comprehensive coverage across domains.
- **Quality Control**: All images undergo manual verification for visual quality and category correctness.
- **Domain Diversity**: The 18 categories cover a wide range of scenarios from daily life to specialized domains like medical imaging and autonomous driving.
- **High-Quality Captions**: Each training image is paired with a GPT-4o generated caption that follows category-specific guidelines.

We plan to release this dataset along with our code and models to facilitate future research in domain-specific image captioning.

Evaluation Metrics We employ two types of evaluation:

- **Automated Evaluation**: The Pixtral model serves as our primary metric, assigning scores from 1 to 5 based on caption quality. We evaluate all generated captions on the entire validation set of 796 images.
- **Human Evaluation**: Three expert annotators evaluate 54 randomly sampled image-caption pairs using the same 1-5 scale. The sample size was chosen to ensure thorough evaluation while maintaining feasible annotation effort.

Baselines We compare our method with several strong baselines:

- **GPT-4o**: A large-scale language model developed by OpenAI, known for its advanced natural language understanding and generation capabilities. It serves as a benchmark for state-of-the-art performance in various language tasks.
- **GPT-4o-mini**: A more compact version of GPT-4o, designed to offer competitive performance with reduced computational requirements. This model maintains a balance between efficiency and effectiveness, making it suitable for applications with limited resources.
- **Claude-3-Haiku**: A large language model developed by Anthropic, optimized for speed and affordability. It is designed to process 21,000 tokens per second for prompts under 32,000 tokens, making it suitable for enterprise workloads that require quick analysis of large datasets.
- **LLaVA-v1.5-7B**: An open-source multimodal model fine-tuned from LLaMA and Vicuna on GPT-generated multimodal instruction-following data. It is an auto-regressive language model based on the transformer architecture, trained to understand and generate

both text and images. The model was trained in September 2023 and is licensed under the LLAMA 2 Community License.

- **LLaVA-FT**: A version of LLaVA model that has undergone full-parameter fine-tuning on our curated dataset of 9,840 image-caption pairs, trained without Direct Preference Optimization (DPO). This model represents the effectiveness of traditional fine-tuning approaches on our domain-specific dataset.

- **LLaVA-LoRA**: A variant of the LLaVA model fine-tuned on our dataset using Low-Rank Adaptation (LoRA) techniques, without DPO. This approach demonstrates the efficacy of parameter-efficient fine-tuning methods on our carefully curated training data while maintaining model performance.

- **Qwen-VL-Chat** (Bai et al., 2023): A multimodal chatbot model that integrates vision and language understanding, enabling it to process and generate responses based on both textual and visual inputs. This model is designed to handle a wide range of multimodal tasks, including image captioning and visual question answering.

- **Qwen2-VL-7B-Instruct** (Wang et al., 2024): An advanced multimodal model that builds upon the Qwen-VL-Chat architecture, incorporating instruction-following capabilities to improve its performance on tasks requiring specific guidance. The 7B refers to the model's parameter size, indicating a balance between computational efficiency and performance.

4.2 Main Results

The experimental results reveal several comprehensive findings:

Base Model Comparisons: Among the models, we observe a clear performance hierarchy. While the closed-source GPT-4o maintains the highest performance (4.473), setting a strong upper bound, other closed-source models like GPT-4o-mini (4.293) and Claude-3-Haiku (4.268) also show strong performance. For open-source models, Qwen2-VL-7B-Instruct achieves 4.224 and Qwen-VL-Chat reaches 3.833. LLaVA-v1.5-7B serves as our baseline with a score of 3.598.

Fine-tuning Effectiveness: Our fine-tuning approaches show significant improvements over the base model. Full parameter fine-tuning achieves a score of 4.078 (+0.480 over base), demonstrating the effectiveness of comprehensive model adaptation. LoRA fine-tuning reaches 4.034 (+0.436 over base), nearly matching full fine-tuning while being parameter-efficient. Both approaches outperform Qwen-VL-Chat but still trail behind more advanced models like Claude-3-Haiku and Qwen2-VL-7B-Instruct, suggesting room for further improvement.

Model	Pixtral Score	Δ	#Params
GPT-4o	**4.473**	-	-
GPT-4o-mini	4.293	-	-
Claude-3-Haiku	4.268	-	-
LLaVA-v1.5-7B (Base)	3.598	-	7B
Qwen-VL-Chat	3.833	+0.235	7B
Qwen2-VL-7B-Instruct	4.224	+0.626	7B
LLaVA-FT (Full)	4.078	+0.480	7B
LLaVA-LoRA	4.034	+0.436	7B
DPO (1 iteration)	4.113	+0.515	7B
DPO (2 iterations)	4.168	+0.570	7B
DPO (3 iterations)	4.290	+0.692	7B
DPO (4 iterations)	**4.312**	+0.714	7B

Table 1: Performance comparison on the validation set (796 images). Scores range from 1 to 5. Δ shows absolute improvement over the base model. #Params shows the total number of parameters.

DPO Improvement: The iterative DPO process shows consistent and meaningful improvements. First iteration achieves +0.515 over base model, surpassing both fine-tuning approaches. Subsequent iterations show continuous improvement, with the second iteration adding +0.055 (reaching 4.168), the third iteration showing the largest incremental gain of +0.122 (achieving 4.290), and the fourth iteration providing a modest +0.022 improvement (reaching 4.312). The diminishing returns after the third iteration suggest convergence of the optimization process. Notably, our final DPO model outperforms both Qwen2-VL-7B-Instruct by +0.088 points and Claude-3-Haiku by +0.044 points, demonstrating the effectiveness of our approach.

Gap Analysis with Leading Models: Our final model (DPO 4 iterations) achieves a score of 4.312, which is particularly noteworthy as it surpasses both closed-source models (GPT-4o-mini and Claude-3-Haiku) and open-source models (Qwen2-VL-7B-Instruct and Qwen-VL-Chat). While there remains a small gap of 0.161 points from GPT-4o, our model outperforms GPT-4o-mini by +0.019 points, Claude-3-Haiku by +0.044

points, and the best open-source model Qwen2-VL-7B-Instruct by +0.088 points. This achievement is especially significant considering we maintain the computational efficiency of a 7B parameter model, demonstrating that our approach effectively narrows the performance gap between open-source and closed-source models in the field of image captioning.

Computational Efficiency: All our models maintain the original 7B parameter count, offering practical deployment capabilities compared to larger models like GPT-4o and Claude-3-Haiku. This enables efficient inference time while achieving competitive performance, providing a scalable solution for real-world applications.

4.3 Human Validation Study

To validate the reliability of Pixtral as an automated evaluation metric, we conducted a comprehensive human evaluation study. The study included 54 image-caption pairs randomly sampled from the validation set, evaluated by three experts with ML/CV backgrounds and experience in image captioning. Each annotator independently scored all pairs following detailed evaluation criteria, using a 1-5 scale with 1 point increments, which aligns with the Pixtral scoring scale.

Correlation Analysis The comparison between human and Pixtral scores shows strong alignment with a Pearson correlation of 0.762 (p-value: 0.000), Kendall's Tau of 0.545, and Spearman correlation of 0.595. These strong correlations across different statistical measures validate Pixtral's effectiveness as an automated evaluation metric.

Error Analysis The difference between Pixtral and human scores demonstrates reasonable agreement, with a Mean Absolute Error (MAE) of 0.414 and Root Mean Square Error (RMSE) of 0.572. These values indicate that Pixtral's scores typically deviate from human consensus by less than half a point on the 5-point scale, suggesting strong practical reliability.

Inter-annotator Agreement The Cohen's Kappa coefficients between annotators are 0.507, 0.617, and 0.466 for pairs 1-2, 1-3, and 2-3, respectively. These values reflect moderate to good inter-annotator agreement, highlighting that while there are some subjective differences in caption evaluation, the overall consistency is still acceptable. These results further underscore the advantage of using a consistent automated metric like Pixtral, and the importance of averaging multiple human judgments to mitigate individual biases.

4.4 Qualitative Analysis

We demonstrate our model's capability through several test cases on images not present in our training set. As shown in Fig. 1, our classification-guided prompting effectively handles various domain-specific characteristics. For animal images, our model accurately captures physical attributes (e.g., "fluffy gray and white kitten") and behavioral aspects (e.g., "lounges comfortably"). For person descriptions, it successfully combines human attributes with environmental context (e.g., "standing against a vibrant blue wall"). In plant-related cases, it shows capability in describing fine-grained details (e.g., "green soybean seed"). These examples demonstrate our model's robust performance on novel images. However, challenges remain in several aspects. The model sometimes struggles with complex scenes involving multiple objects or actions, may have difficulty with rare or unusual visual concepts, and occasionally misses subtle contextual details. These limitations suggest directions for future improvements in our approach.

5 Discussion and Conclusion

5.1 Key Findings

Our experimental results support several key conclusions:

1) The effectiveness of our approach in improving caption quality, evidenced by both automated and human evaluation. Through classification-guided prompting, knowledge distillation, and DPO refinement, our method demonstrates consistent improvements over the base model.

2) The validity of Pixtral as an automated evaluation metric, supported by strong correlation with human judgments. This validation enables efficient automatic evaluation of large-scale image captioning systems.

3) The efficiency of our pipeline, with our final model (DPO 4 iterations) achieving strong performance (4.312) while maintaining practical deployment requirements. This demonstrates the effectiveness of our knowledge distillation approach in transferring capabilities from larger models.

4) The complementary benefits of different components: classification-guided prompting for lan-

guage quality improvement, knowledge distillation for model capability transfer, and DPO for continuous optimization. Each component contributes to the overall performance gain, leading to consistent improvements over the base model.

5) We introduce a fine-tuned DPO version of the LLaVA model, which demonstrates strong image captioning capabilities, further enhancing the practical application of the model.

5.2 Dataset Contribution

A significant outcome of this work is our curated dataset, which includes 9,840 training images and 796 validation images, spanning 18 diverse categories. This dataset offers several unique advantages:

- **High-quality GPT-4o-generated captions**: Each image is paired with a detailed caption generated by the GPT-4o model, which accurately captures domain-specific details. We have carefully designed category-specific prompting strategies to ensure that the descriptions reflect the unique context and information of each category.

- **Balanced category distribution**: We ensured that both the training and validation sets have a balanced distribution of categories, ensuring comprehensive coverage across all categories. This balance helps the model perform well across different types of images, avoiding bias towards any specific category and improving generalizability.

- **Manual verification ensuring data quality and category accuracy**: All images have undergone manual verification to ensure their quality and the accuracy of their category labels. Expert reviewers manually checked each image to ensure it is correctly categorized and meets the required standards, eliminating the potential impact of incorrect labels.

- **Diverse domain coverage**: The dataset spans a wide range of domains, from common scenarios to specialized fields, including animal, architecture, art, daily life, food, medical, smart cities, and more. These categories include everyday scenes as well as specialized fields such as smart transportation, healthcare, and autonomous driving.

- **Rich Contextual Information and Details**: The captions provided with each image extend beyond mere visual descriptions, incorporating comprehensive contextual information. These captions transcend simple visual representation, integrating broader background and nuanced details to provide profound insights into the scene, environment, and wider context.

We believe that this dataset will serve as a valuable resource for future research in domain-specific image captioning and multimodal understanding. It provides researchers with a high-quality, manually verified data foundation that can significantly advance research in image understanding, caption generation, and cross-modal learning. Additionally, the dataset's diversity and high-quality captions offer a rich source of experimental data for training and evaluating domain-specific generation models, contributing to the further development of more refined models for image captioning.

5.3 Conclusion

We have presented a hierarchical approach to image captioning that combines classification-guided prompting, knowledge distillation, and iterative preference optimization. Our method demonstrates significant improvements over baseline approaches, achieving performance comparable to much larger models while maintaining practical deployment requirements. We also introduce a fine-tuned DPO version of the LLaVA model, which further enhances the practical application of the model for image captioning. Together with our released dataset, this work contributes both methodological advances and valuable resources to the field of image captioning. We believe these contributions will facilitate future research in domain-specific visual understanding and caption generation.

5.4 Limitations and Future Work

Although our approach performs well, it has some limitations. The current classification system with 18 categories may not capture all image scenarios, and the model faces challenges with complex scenes containing multiple objects. Additionally, the DPO method shows slow convergence in later iterations. Future work should focus on improving classification systems and enhancing the model's ability to handle complex scenes.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *Preprint*, arXiv:1607.08822.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *Preprint*, arXiv:1707.07998.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training With Frozen Image Encoders and Large Language Models. *arXiv preprint*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2017. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *ICCVW*, pages 206–213.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2018. It's not all about size: On the role of data properties in pedestrian detection. In *ECCVW*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *Preprint*, arXiv:1411.5726.

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*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention. *Preprint*, arXiv:1502.03044.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Category-Specific Prompts

This appendix provides the complete set of category-specific prompts used in our research. Each prompt follows the structured design described in Section 3.1. Below we present the prompts for all 18 categories, illustrating our comprehensive approach to domain-specific image captioning.

A.1 Animal Category: Biological Entity Focus

System Prompt:

Generate a detailed and accurate caption for the provided animal image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Animal Species**:

- Mention the species of the animal if identifiable

3. **Physical Characteristics**:

- Describe the physical features of the animal (e.g., color, size, distinguishing marks)

4. **Behavior**:

- Describe the behavior or activity of the animal in the image

5. **Habitat**:

- Mention the environment or setting where the animal is located (e.g., forest, desert, ocean)

6. **Contextual Details**:

- Provide any additional relevant information (e.g., weather conditions, time of day)

User Text: "You are given an image of an animal. Please generate a detailed and accurate caption for the image without any additional information."

A.2 Architecture Category: Built Environment Focus

System Prompt:

Generate a detailed and accurate caption for the provided architecture image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Building Type**:

- Mention the type of building (e.g., residential, commercial, historical)

3. **Architectural Style**:

- Describe the architectural style (e.g., modern, Gothic, Baroque)

4. **Physical Characteristics**:

- Describe the physical features of the building (e.g., height, materials, distinctive elements)

5. **Surroundings**:

- Mention the building's surroundings (e.g., urban area, countryside)

6. **Contextual Details**:

- Provide any additional relevant information (e.g., weather conditions, time of day)

User Text: "You are given an image of an architectural structure. Please generate a detailed and accurate caption for the image without any additional information."

A.3 Art Category: Creative Work Focus

System Prompt:

Generate a detailed and accurate caption for the provided art image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Art Form**:

- Mention the form of art (e.g., painting, sculpture, digital art)

3. **Style**:

- Describe the style of the artwork (e.g., abstract, realism, impressionism)

4. **Elements and Composition**:

- Describe the elements and composition of the artwork (e.g., colors, shapes, subjects)

5. **Context**:

- Mention any relevant context (e.g., artist, period, cultural significance)

6. **Emotional and Interpretative Elements**:

- Provide interpretation or emotional elements conveyed by the artwork

User Text: "You are given an image of an artwork. Please generate a detailed and accurate caption for the image without any additional information."

A.4 Autonomous Driving Category: Vehicle Environment Focus

System Prompt:

Generate a detailed and accurate caption for the provided autonomous driving image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Scene Overview**:

- Briefly describe the overall scene (e.g., urban street, highway, suburban area)

- Mention weather conditions and time of day (e.g., daytime, night, rainy)

3. **Road Environment**:

- Describe road type (e.g., two-way four-lane, one-way street)

- Mention road surface conditions (e.g., dry, wet, snowy)

- Describe road markings and traffic signs

4. **Traffic Conditions**:

- Describe positions and behaviors of surrounding vehicles

- Mention pedestrians, cyclists, and other road users

- Describe traffic light status (if visible)

5. **Ego Vehicle Status**:

- Describe the autonomous vehicle's position and driving status

- Mention speed and relative position to other vehicles

6. **Potential Hazards**:

- Point out any potential dangers or situations requiring special attention

7. **Autonomous Driving Relevant Elements**:

- Describe elements particularly important for the autonomous driving system

- Mention situations that may require special handling

User Text: "You are given an image related to autonomous driving. Please generate a detailed and accurate caption for the image without any additional information."

A.5 Cartoon Category: Animated Content Focus

System Prompt:

Generate a detailed and accurate caption for the provided cartoon image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Character Description**:

- Mention the main characters and their characteristics

3. **Content Description**:

- Describe the content of the image

4. **Scene Description**:

- Describe the scene or setting (e.g., location, time of day)

5. **Actions and Interactions**:

- Describe the actions and interactions of the characters

6. **Visual Style**:

- Mention the visual style of the cartoon (e.g., color palette, drawing style)

7. **Contextual Details**:

- Provide any additional relevant information (e.g., mood, tone)

User Text: "You are given an image of a cartoon. Please generate a detailed and accurate caption for the image without any additional information."

A.6 Chart Category: Data Visualization Focus

System Prompt:

Generate a detailed and accurate caption for the provided chart image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Details**:

- Include relevant specifics such as the type of chart, data points, trends, and key observations.

3. **Consistency**:

- Maintain a consistent style and format.

4. **Readability**:

- Write in clear, concise language, avoiding unnecessary complexity.

5. **Insights or Recommendations**:

- If appropriate, provide simple insights or recommendations based on the chart data.

User Text: "You are given an image of a chart image. Please generate a detailed and accurate caption for the image without any additional information."

A.7 City Category: Urban Environment Focus

System Prompt:

Generate a detailed and accurate caption for the provided city image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Cityscape Overview**:

- Briefly describe the overall cityscape (e.g., skyline, street view)

3. **Landmarks and Buildings**:

- Mention notable landmarks or buildings

4. **Activity and Movement**:

- Describe any visible activity or movement (e.g., traffic, pedestrians)

5. **Weather and Time of Day**:

- Mention weather conditions and time of day

6. **Contextual Details**:

- Provide any additional relevant information (e.g., cultural or historical significance)

User Text: "You are given an image of a city. Please generate a detailed and accurate caption for the image without any additional information."

A.8 Daily Life Category: Everyday Scene Focus

System Prompt:

Generate a detailed and accurate caption for the provided daily life image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Scene Overview**:

- Briefly describe the overall scene (e.g., home, office, street)

3. **Activities**:

- Describe the activities of people in the scene

4. **Objects and Environment**:

- Mention notable objects and environmental details

5. **Emotional and Social Context**:

- Describe the emotional and social context of the scene

6. **Contextual Details**:

- Provide any additional relevant information (e.g., time of day, weather)

User Text: "You are given an image depicting daily life. Please generate a detailed and accurate caption for the image without any additional information."

A.9 Food Category: Culinary Focus

System Prompt:

Generate a detailed and accurate caption for the provided food image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Dish Name**:

- Mention the name of the dish if identifiable

3. **Ingredients**:

- Describe the main ingredients used

4. **Presentation**:

- Describe the presentation and plating of the dish

5. **Appearance**:

- Describe the appearance of the dish

6. **Context**:

- Mention any relevant context (e.g., cuisine type, occasion)

7. **Sensory Details**:

- Provide sensory details (e.g., aroma, taste, texture)

User Text: "You are given an image of food. Please generate a detailed and accurate caption for the image without any additional information."

A.10 Landscape Category: Natural Scene Focus

System Prompt:

Generate a detailed and accurate caption for the provided landscape image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Scene Overview**:

- Briefly describe the overall landscape (e.g., mountains, beach, forest)

3. **Natural Features**:

- Mention notable natural features (e.g., rivers, trees, cliffs)

4. **Weather and Time of Day**:

- Mention weather conditions and time of day

5. **Flora and Fauna**:

- Describe any visible flora and fauna

6. **Contextual Details**:

- Provide any additional relevant information (e.g., location, season)

User Text: "You are given an image of a landscape. Please generate a detailed and accurate caption for the image without any additional information."

A.11 Love Category: Emotional Focus

System Prompt:

Generate a detailed and accurate caption for the provided love-themed image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Scene Overview**:

- Briefly describe the overall scene

3. **Emotional Elements**:

- Describe the emotions depicted in the image

4. **Activities**:

- Mention any specific activities or interactions

5. **Visual Details**:

- Provide visual details that enhance the romantic theme

User Text: "You are given an image with a love theme. Please generate a detailed and accurate caption for the image without any additional information."

A.12 Medical Category: Healthcare Focus

System Prompt:

Generate a detailed and accurate caption for the provided medical image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Scene Overview**:

- Briefly describe the overall scene

3. **Medical Procedures**:

- Describe any visible medical procedures or activities

4. **Medical Equipment**:

- Mention notable medical equipment and instruments

5. **Contextual Details**:

- Provide any additional relevant information (e.g., condition being treated, emotional atmosphere)

User Text: "You are given an image depicting a medical scene. Please generate a detailed and accurate caption for the image without any additional information."

A.13 Person Category: Human Subject Focus

System Prompt:

Generate a detailed and accurate caption for the provided person image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Physical Appearance**:

- Describe the person's physical appearance (e.g., age, gender, clothing)

3. **Actions and Activities**:

- Mention any actions or activities the person is engaged in

4. **Emotional State**:

- Describe the person's emotional state or expression

5. **Context and Setting**:

- Provide context about the setting or background

6. **Contextual Details**:

- Provide any additional relevant information (e.g., weather, time of day)

User Text: "You are given an image of a person. Please generate a detailed and accurate caption for the image without any additional information."

A.14 Plant Category: Botanical Focus

System Prompt:

Generate a detailed and accurate caption for the provided plant image. The caption should include the following elements:

1. **Accuracy**:

- Ensure the description is precise and free of errors.

2. **Plant Species**:

- Mention the species of the plant if identifiable

3. **Physical Characteristics**:

- Describe the physical features of the plant (e.g., color, size, distinctive elements)

4. **Environment**:

- Mention the environment or setting where the plant is located (e.g., garden, forest)

5. **Contextual Details**:

- Provide any additional relevant information (e.g., season, time of day)

User Text: "You are given an image of a plant. Please generate a detailed and accurate caption for the image without any additional information."

A.15 Smart Cities Category: Urban Technology Focus

System Prompt: Generate a detailed and accurate caption for the provided smart city image. The caption should include the following elements: 1. **Accuracy**: - Ensure the description is precise and free of errors. 2. **Cityscape Overview**: - Briefly describe the overall cityscape (e.g., skyline, street view) 3. **Smart Infrastructure**: - Mention notable smart infrastructure elements (e.g., smart buildings, IoT devices) 4. **Technological Elements**: - Describe visible technological

elements (e.g., sensors, automated systems) 5. **Traffic and Transportation**: - Mention traffic and transportation systems (e.g., autonomous vehicles, smart traffic lights) 6. **Contextual Details**: - Provide any additional relevant information (e.g., time of day, weather)

User Text: "You are given an image of a smart city. Please generate a detailed and accurate caption for the image without any additional information."

A.16 Sports Category: Athletic Activity Focus

System Prompt: Generate a detailed and accurate caption for the provided sports image. The caption should include the following elements: 1. **Accuracy**: - Ensure the description is precise and free of errors. 2. **Sport Type**: - Mention the type of sport being played 3. **Players and Actions**: - Describe the players and their actions 4. **Equipment**: - Mention any visible sports equipment 5. **Setting**: - Describe the setting (e.g., stadium, field) 6. **Contextual Details**: - Provide any additional relevant information (e.g., score, spectators, weather)

User Text: "You are given an image of a sports scene. Please generate a detailed and accurate caption for the image without any additional information."

A.17 Transportation Category: Mobility Focus

System Prompt: Generate a detailed and accurate caption for the provided transportation image. The caption should include the following elements: 1. **Accuracy**: - Ensure the description is precise and free of errors. 2. **Vehicle Type**: - Mention the type of vehicle(s) shown 3. **Scene Overview**: - Briefly describe the overall scene (e.g., road, railway, airport) 4. **Activity and Movement**: - Describe any visible activity or movement of the vehicles 5. **Environment**: - Mention the environment or setting (e.g., urban, rural) 6. **Contextual Details**: - Provide any additional relevant information (e.g., weather conditions, time of day)

User Text: "You are given an image of a transportation scene. Please generate a detailed and accurate caption for the image without any additional information."

A.18 Other Category: General Content Focus

System Prompt: Generate a detailed and accurate caption for the provided image. The caption should include the following elements: 1. **Accuracy**: - Ensure the description is precise and free of errors. 2. **Content Description**: - Provide a clear and relevant description of the content in the image. 3. **In-depth Understanding**: - Go beyond surface-level details by interpreting the underlying meaning or significance of the scene, including any implied relationships, emotions, or actions. 4. **Contextual Details**: - Include any necessary contextual details (e.g., cultural, historical, social significance) to enhance the understanding of the image. 5. **Visual and Environmental Elements**: - Describe notable visual and environmental elements present in the image, paying attention to subtle features that may influence the overall interpretation.

User Text: "You are given an image. Please generate a detailed and accurate caption for the image without any additional information, focusing on both surface details and deeper interpretations."

We plan to make these prompts and associated implementations available to facilitate future research in domain-specific image captioning.

B Training Details

B.1 Model Configurations

Classification Model

- Base model: LLaVA-v1.5-7B
- Training strategy: Full parameter fine-tuning
- Learning rate: 2e-5
- Batch size: 16 per device
- Number of epochs: 3

Caption Model (Full Parameter)

- Base model: LLaVA-v1.5-7B
- Training data: GPT-4o generated captions
- Learning rate: 2e-5
- Batch size: 16 per device
- Training epochs: 3

Caption Model (LoRA)

- LoRA rank: 128
- LoRA alpha: 256
- LoRA dropout: 0
- Learning rate: 2e-4
- Target modules: all

DPO Training

- LoRA rank: 8
- LoRA alpha: 16
- Learning rate: 2e-6
- Batch size: 2 per device
- Gradient accumulation steps: 8

B.2 Computing Infrastructure

- Hardware: 8 NVIDIA A100 GPUs (80GB)
- Framework: LlamaFactory with DeepSpeed
- Mixed precision: BF16

For detailed training scripts and configurations, we will make these publicly available in our upcoming repository to facilitate future research in this area.

C Pixtral Evaluation Implementation

The Pixtral evaluation system uses a standardized prompt structure for evaluating captions:

> You are a helpful and precise assistant for evaluating the quality of captions. A caption will be provided for a particular image. Below are the criteria for evaluation:
>
> <Criteria Begin> {category-specific criteria} <Criteria End>
>
> Based on the criteria provided, does the caption meet the requirements?
>
> Please rate the caption on a scale of 1 to 5, where:
>
> - 1 - Poor: Does not meet the criteria at all.
> - 2 - Fair: Meets a few criteria but has significant issues.
> - 3 - Good: Meets most criteria with minor issues.
> - 4 - Very Good: Meets all criteria with very few issues.
> - 5 - Excellent: Perfectly meets all criteria.
>
> Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 5 by strictly following this format: "Rating: [[rating]]", for example: "Rating: [[5]]".
>
> Here is the image and the corresponding caption.
> Caption: {caption to be evaluated}

For detailed evaluation implementations and criteria, we will make these publicly available in our upcoming repository to facilitate future research in this area.