Appendix

Overview

This is the Appendix for the paper "DeepAlign: Mitigating Modality Conflict through Modality-Specific Alignment". In this supplementary material we present:

- The specific details and configurations about our implementation is described in Section A.
- The detailed introduction of the settings for each of our experiments is provided in Section B.

A. Implementation Details.

DeepAlign is a model-agnostic post-training method and we apply it on LLaVA-v1.5-7B [9], Qwen2.5-VL-7B, and InternVL3-8B to explore the broad applicability of it.

Training Data. To collect the post-training data, we first leverage a set of text-image pairs selected from a high-quality subset of the CC3M dataset [5]. Following [22], each image corresponds to filtered synthetic captions prepared by BLIP2 [6]. In total, we screen about 1M text-image pairs. Moreover, we also incorporate MiniGPT-4-Instruction and LLaVA-v1.5-mix-665k to preserve the original instruction-following capabilities of the MLLM. Specifically, MiniGPT-4-Instruction [22], as the instruction-tuning data for [22], includes about $3.5K$ instances refined by ChatGPT from detailed descriptions. LLaVA-v1.5-mix-665k, as the instruction-tuning data for [9] is a wider collection with $665K$ instructions, which encompass a wide range of task categories, including dialogue-based Q&A pairs, multiple-choice short Q&A, detailed descriptions, and text-only reasoning tasks. All post-training data are derived from the original pre-training and instruction fine-tuning datasets of the backbone MLLM. And we have not introduced any additional training data.

Model Details. Prior to post-training, we first train the modality classifier f using our collected image-text pairs, which is composed of several MLP layers. It operates on cross-layer pooled visual/textual representations for modality discrimination (e.g., distinguishing visual from textual features). This design prioritizes "lightweight alignment" to avoid over-complicating the main model with excessive parameters. The modality shift module $\texttt{SHF}(\cdot)$ is a convolutional neural network containing 2 layers of 3×3 convolutional layers (with ReLU activation), designed to adjust spatial shift directions in visual feature maps. It is inserted into the middle layers of the visual encoder (e.g., Layers 4 and 6 in ViT), which capture mid-level semantic features balancing low-level spatial structures (e.g., object boundaries) and high-level semantics (e.g., part relationships). This choice ensures a balance between retaining modality-specific details and enabling cross-layer semantic alignment.

Training Setting. During post-training, the only trainable parameters are those of the adapter layers (*i.e.*, modality shift module, 200M params), with a peak learning rate of $3e-5$. We employ cosine decay as the learning rate scheduler, with a weight decay rate of 0.05. Additionally, we set the hyperparameters $\alpha, \beta, \gamma, \mu$ to 1.0. The entire post-training process is conducted on two A100 GPUs and lasts a total of 48 hours.

Computational Efficiency. During post-training, only the modality shift module is trainable, comprising 200M parameters. The entire training process is highly streamlined, requiring only 1M training samples and 48 training hours on two A100 GPUs to achieve convergence. Considering the significant performance enhancement, we conclude that DeepAlign is an effective approach.

B. Experimental Setting

B.1. Zero-shot Vision-Language Comprehension

We compare DeepAlign with several other baseline methods also dedicated to achieving better modality alignment through post-training on a range of academic benchmarks. Following POVID [21] and SIMA [16], our experimental settings are as follows. The evaluation benchmarks include MLLM Benchmarks (MMBench [10], MMStar [1], MMMU [19], HallusionBench [4], OCRBench [11], MMVet [18]) and VQA Benchmarks (ScienceQA [12], TextVQA [14], RealWorldQA, MTVQA [15]). The post-training methods compared with our model include RLHF [8], HADPO [20], DataTailor [17], POVID [21], and SIMA [16]. Below, we provide a brief introduction to these benchmarks, models, and methods.

B.1.1. Benchmarks

- **MLLM Benchmarks:**
  - **MMBench** [10]: A comprehensive fine-grained multimodal evaluation benchmark covering 20+ perception-reasoning core tasks across levels (e.g., object recognition, spatial reasoning, and text understanding) with 30K+ diverse samples for reproducible assessment, designed to assess MLLMs' full general capability.
  - **MMStar** [1]: A large-scale benchmark focusing on multimodal star tasks, including image-text retrieval, visual grounding, and cross-modal generation, with 100K+ high-quality image-text pairs.
  - **MMMU** [19]: A challenging multimodal benchmark simulating university entrance exams, covering 30+ subjects (e.g., math, physics) with 11.5K questions requiring complex reasoning over images and texts.
  - **HallusionBench** [4]: A specialized benchmark for evaluating hallucination in MLLMs, containing 4.5K adversarial examples (e.g., non-existent objects) to test the reliability of visual-text alignment.

  - **OCRBench** [11]: An OCR-centric benchmark with 25K images containing scene text, focusing on tasks like text detection, recognition, and understanding in real-world scenarios (e.g., menus, signs).
  - **MMVet** [18]: A real-world-oriented multimodal benchmark with 2.4K questions across 12 domains (e.g., healthcare, daily life), emphasizing practical utility and alignment with human needs.
- **VQA Benchmarks:**
  - **ScienceQA** [12]: A multimodal science question-answering dataset with 21K questions spanning physics, biology, and chemistry, integrating 10K chain-of-thought explanations and 5K scientific diagrams.
  - **TextVQA** [14]: An OCR-intensive visual QA benchmark with 28K images containing scene text, requiring models to read and comprehend text in visuals (e.g., "What is the expiration date on the bottle?").
  - **RealWorldQA**: A real-world scenario-based VQA dataset focusing on everyday visual understanding, with questions about common scenes (e.g., "Is this milk safe to drink?") to test practical reasoning.
  - **MTVQA** [15]: A multimodal task-oriented VQA benchmark covering cross-modal understanding, temporal reasoning, and multi-turn interaction, with 50K+ questions over images and short videos.

B.1.2. Baselines

- **Post-training Methods**
  - **RLHF** [8]: It integrates human feedback with reinforcement learning, training a reward model to align agent policies with human preferences. The framework involves three stages: policy initialization, reward modeling from pairwise comparisons, and Proximal Policy Optimization.
  - **HADPO** [20]: Its full name is Hallucination Aware Direct Preference Optimization, mitigating object hallucination in LVLMs by framing hallucination as a preference selection task, where models learn to favor non-hallucinating responses through contrastive training.
  - **DataTailor** [17]: It is designed to optimize multimodal data selection by focusing on three key criteria: *informativeness*, *uniqueness*, and *representativeness*. It aims to enhance the quality of selected data for downstream tasks by ensuring a balanced and diverse dataset that captures essential information across modalities.
  - **POVID** [21]: It fine-tunes vision large language models by aligning modalities through preference-based learning. It leverages human or model-generated preferences to improve the alignment between visual and textual representations, enhancing the model's performance in multi-modal tasks.
  - **SIMA** [16]: It is a framework designed to enhance the alignment between visual and language modalities in large vision-language models. By leveraging self-improvement mechanisms, it iteratively refines the interaction between these modalities, improving the model's ability to generate coherent and accurate multimodal representations.

B.2. Fine-grained Perceptive VQA Tasks of Mitigation of Modality Conflict

We evaluate DeepAlign on several perceptive-related tasks on BLINK [2], where the MLLMs must perceive the contents of the image to answer. BLINK is a diagnostic benchmark containing 5,120 adversarial image-text pairs designed to quantify modality conflicts through three perturbation strategies: color shift (RGB histogram displacement $\Delta > 50$), partial object occlusion ($30--50\%$ bounding box masking), and contextual contradiction (e.g., desert images labeled as arctic scenes). Next, we introduce the subtasks of Benchmark.

- Spatial Relation (**Spatial**): It evaluates understanding of spatial relationships (e.g., left/right) using the Visual Spatial Reasoning dataset, reformatted into binary questions.
- Object Localization (**Local**): It focuses on fine-grained object localization using the LVIS dataset, with Gaussian noise added to ground-truth bounding boxes.
- Visual Correspondence (**Vis.Corr.**): It evaluates MLLMs' ability to identify the same scene point across different viewpoints or lighting conditions using the HPatches dataset. Ground-truth homography is used to compute correspondences.
- Visual Similarity (**Similarity**): It evaluates nuanced understanding of visual features using the DreamSim dataset, comparing a reference image with two alternatives.
- Counting (**Counting**): It assesses detection, recognition, and compositional reasoning using the TallyQA dataset, featuring challenging counting questions.
- Relative Depth (**Depth**): It tests geometric understanding by evaluating relative depth perception, a skill humans excel at.

B.3. Multimodal In-context Learning of Emergent Abilities

We measure the few-shot in-context learning capabilities of DeepAlign on OKVQA [13] and VQAv2 [3]. When provided with 4-shot, 8-shot and 16-shot examples, respectively. Below, we will introduce the two benchmarks involved.

- **VQA-v2** [3]: It is an extended version of the Visual Question Answering (VQA) dataset, containing over 1.1 million question-answer pairs based on $204,721$ images from the COCO dataset. It evaluates models on their ability to answer questions about images, with accuracy as the primary metric, and includes separate evaluations for different question types (e.g., yes/no, number, other).

- **OKVQA** [13]: It is a challenging dataset for Visual Question Answering (VQA) that requires models to leverage external knowledge to answer questions. It contains $14,055$ question-answer pairs based on images from the COCO dataset, with accuracy as the primary evaluation metric. Unlike traditional VQA datasets, OK-VQA emphasizes the integration of external knowledge sources to address questions that cannot be answered solely from visual information.

B.4. Demonstrative Instruction Following of Emergent Abilities

We conduct experiments on the DEMON benchmark [7], which is specifically designed to evaluate a model's capability of following demonstrative instructions on interleaved image-text sequences. The DEMON benchmark focuses on evaluating multimodal large language models (LLMs) in zero-shot settings by fine-tuning them to follow demonstrative instructions. It includes a diverse set of tasks that test the model's ability to generalize across modalities (e.g., text and images) without task-specific training. Performance is measured using task-specific metrics such as accuracy and F1 score, depending on the nature of the task. Below, we introduce each subtask of DEMON benchmark.

- Multimodal Dialogue (**MMD**): It involves integrating multiple modalities (e.g., text, speech, vision) to enable context-aware and natural interactions between humans and machines.
- Visual Storytelling (**VST**): It is the task of generating coherent and descriptive narratives from a sequence of images, combining visual understanding and natural language generation.
- Visual Relation Inference (**VRI**): It involves identifying and reasoning about the relationships between objects in an image, such as spatial or functional interactions.
- Multimodal Cloze (**MMC**): It leverages both visual and textual contexts to predict missing elements, enabling joint learning across modalities for improved contextual understanding.
- Knowledge Grounded QA (**KGQA**): It involves answering questions by grounding responses in both multimodal inputs (e.g., images, text) and external knowledge sources, requiring models to integrate diverse information for accurate reasoning.
- Text-Rich Images QA (**TRQA**): It focuses on answering questions about images containing embedded textual information, requiring models to jointly analyze visual content and extract relevant text for accurate comprehension.
- Multi-Image Reasoning (**MMR**): It involves analyzing and reasoning across multiple images to infer higher-level relationships, often leveraging techniques from computer vision and natural language processing to model complex interactions between visual and textual data.

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1

[2] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 2

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[4] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. 1

[5] Sanghyun Jo, Soohyun Ryu, Sungyub Kim, Eunho Yang, and Kyungsu Kim. Ttd: Text-tag self-distillation enhancing image-text alignment in clip to alleviate single tag bias, 2024. 1

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[7] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[8] Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism, 2023. 1, 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[10] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[11] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 1, 2

[12] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[13] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[14] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1, 2

[15] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024. 1, 2

[16] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, and Cao Xiao. Enhancing visual-language modality alignment in large vision language models via self-improvement, 2024. 1, 2

[17] Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness, 2024. 1, 2

[18] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 1, 2

[19] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 1

[20] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2023. 1, 2

[21] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. 1, 2

[22] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1