F2RVLM: Boosting Fine-grained Fragment Retrieval for Multi-Modal Long-form Dialogue with Vision Language Model

Hanbo Bi[1,2*], Zhiqiang Yuan[1*], Zexi Jia[1], Jiapei Zhang[1], Chongyang Li[1,2], Peixiang Luo[1], Ying Deng[1], Xiaoyue Duan[1], Jinchao Zhang[1†]

[1]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[2]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China
yuanzhiqang19@mails.ucas.ac.cn

Abstract

Traditional dialogue retrieval aims to select the most appropriate utterance or image from recent dialogue history. However, they often fail to meet users' actual needs for revisiting semantically coherent content scattered across long-form conversations. To fill this gap, we define the Fine-grained Fragment Retrieval (FFR) task, requiring models to locate query-relevant fragments, comprising both utterances and images, from multimodal long-form dialogues. As a foundation for FFR, we construct MLDR, the longest-turn multimodal dialogue retrieval dataset to date, averaging 25.45 turns per dialogue, with each naturally spanning three distinct topics. To evaluate generalization in real-world scenarios, we curate and annotate a WeChat-based test set comprising real-world multimodal dialogues with an average of 75.38 turns. Building on these resources, we explore existing generation-based Vision-Language Models (VLMs) on FFR and observe that they often retrieve incoherent utterance-image fragments. While optimized for generating responses from visual-textual inputs, these models lack explicit supervision to ensure semantic coherence within retrieved fragments. To address this, we propose F2RVLM, a generative retrieval model trained in a two-stage paradigm: (1) supervised fine-tuning to inject fragment-level retrieval knowledge, and (2) GRPO-based reinforcement learning with multi-objective rewards to encourage outputs with semantic precision, relevance, and contextual coherence. In addition, to account for difficulty variations arising from differences in intra-fragment element distribution, ranging from locally dense to sparsely scattered, we introduce a difficulty-aware curriculum sampling that ranks training instances by predicted difficulty and gradually incorporates harder examples. This strategy enhances the model's reasoning ability in long, multi-turn dialogue contexts. Experiments on both in-domain and real-domain sets demonstrate that F2RVLM substantially outperforms popular VLMs, achieving superior retrieval performance.

Code & Dataset — https://f2rvlm.github.io

Introduction

With the widespread adoption of messaging platforms and AI assistants, users frequently need to revisit earlier utterances or referenced images for information confirmation or context tracking. Efficiently retrieving specific content from ongoing multimodal dialogues has become a crucial capability for intelligent systems. Such dialogues often span dozens of turns, intertwining multiple topics, user intents, and visual-textual cues. As illustrated in Fig.1, unlike conventional retrieval tasks that select the most appropriate element from the dialogue history, this scenario demands models to directly locate semantically coherent fragments, including both text and images, enabling users to efficiently access relevant historical content. To address this need, we define the **Fine-grained Fragment Retrieval (FFR)** task, which aims to accurately extract query-relevant fragments

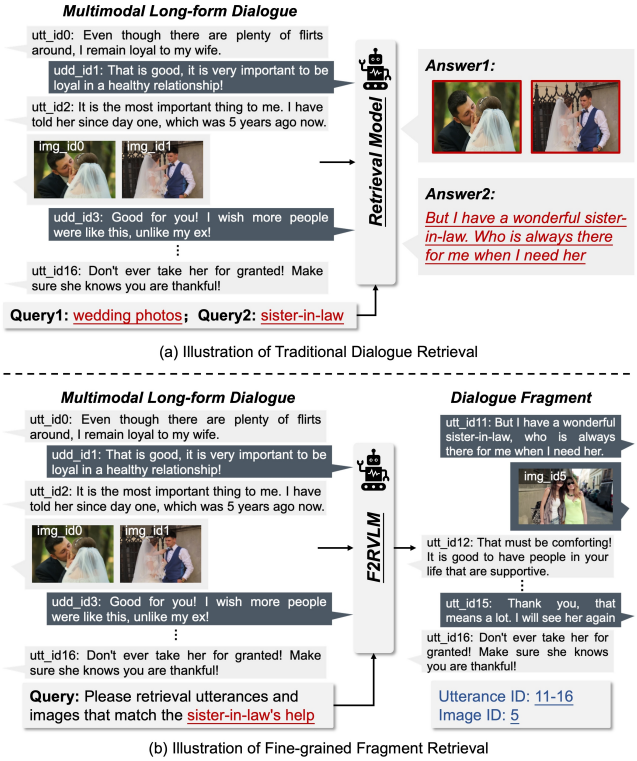

Figure 1: Comparison between the traditional Dialogue Retrieval and our Fine-grained Fragment Retrieval task.

*These authors contributed equally.
†Corresponding author: Jinchao Zhang

from long-form, multi-party, multimodal dialogues.

As a foundation for studying FFR, we construct MLDR, a large-scale multimodal long-form dialogue retrieval dataset. Each dialogue covers three distinct topics with an average of 25.45 turns, making it the longest-turn multimodal dialogue dataset to date. To support diverse retrieval intents, we provide multi-granularity annotations, including domain-level and event-level tags for each dialogue unit. In addition, we curate a WeChat-based test set comprising real-world dialogues with an average of 75.38 turns, aiming to evaluate generalization in practical settings. This set is collected from real conversations contributed by multiple volunteers and annotated through a multi-stage process, serving as a realistic and challenging benchmark for fragment retrieval.

Building on our constructed dataset resources, we investigate existing retrieval models on the proposed FFR task. In retrieval scenarios, current models predominantly follow two paradigms: embedding-based and generation-based. Embedding-based Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021) and BLIP2 (Li et al. 2023), encode images and texts into a shared embedding space and retrieve via similarity ranking. In contrast, generation-based models are typically pre-trained on large-scale image-text corpora to align cross-modal semantics, and subsequently instruction-tuned to generate responses conditioned on visual-textual inputs. Models like GPT-4o (Jaech et al. 2024) and Gemini (Comanici et al. 2025) demonstrate strong performance in prompt-driven retrieval settings, leveraging unified architectures with promising contextual understanding in open-ended multimodal dialogues.

Despite their general success, our evaluation reveals that even leading VLMs struggle to localize relevant fragments accurately in long-form multimodal dialogues. Models such as Qwen2.5-VL-72B (Bai et al. 2025) and Doubao-Seed-1.6 (Guo et al. 2025a) frequently retrieve incoherent utterance-image pairs, e.g., mismatched dialogue turns or irrelevant visual content, resulting in suboptimal F1 scores under real-world conditions. This limitation primarily stems from the gap between the models' learning objectives and the demands of fragment retrieval: while optimized for generating responses from visual-textual inputs, these models lack explicit supervision to ensure that the retrieved or generated fragments (both utterances and images) are semantically coherent and contextually aligned with the user query.

To address these limitations, we propose **F2RVLM**, a generation-based retrieval framework for long-form multimodal dialogues, tailored for **F**ine-grained **F**ragment **R**etrieval with **V**ision-**L**anguage **M**odel. F2RVLM follows a two-stage training paradigm: fragment-level retrieval knowledge is first injected via supervised fine-tuning, followed by GRPO-based reinforcement learning to align retrieval behavior with human preferences. We design a multi-objective reward scheme that encourages the generation of fragments with semantic precision and contextual coherence. Specifically: (1) an F1-based alignment reward encourages accurate matching with ground-truth fragments, penalizing both over- and under-retrieval; (2) a fragment order consistency enhances semantic alignment between selected utterances and images, guiding the model to organize content in a coherent, human-preferred manner. Moreover, we observe that fragment structures in long-form dialogues differ in their internal distribution, from locally dense to sparsely scattered, which directly reflects differences in retrieval difficulty. To exploit this inherent hierarchy of difficulty, we introduce a curriculum sampling that ranks training samples by predicted F1 and confidence, progressively exposing the model to harder instances with greater contextual complexity. This promotes robust reasoning in diverse, noisy, and long-range scenarios. Extensive experiments on both the in-domain MLDR and real-domain WeChat-based sets demonstrate that F2RVLM significantly outperforms mainstream VLMs in retrieval accuracy and contextual understanding.

Our main contributions are summarized as follows:

- We introduce Fine-grained Fragment Retrieval, a novel retrieval task that aims to directly locate semantically coherent utterance-image fragments from long-form dialogues, differing from traditional dialogue retrieval that selects the most appropriate individual elements.
- We construct MLDR, the longest-turn multimodal dialogue retrieval dataset to date, averaging 25.45 turns per dialogue, with each covering three distinct topics. Additionally, we curate a real-world WeChat-based test set averaging 75.38 turns per dialogue to evaluate retrieval generalization in practice.
- We propose F2RVLM, a generation-based retrieval model for the proposed task. It integrates GRPO-based reinforcement learning with multi-objective rewards and difficulty-aware curriculum sampling to progressively enhance the semantic consistency and completeness of retrieved fragments.
- Experimental results on both the in-domain MLDR validation set and the real-domain WeChat-based test set demonstrate that F2RVLM consistently outperforms popular VLMs in fragment retrieval accuracy.

Related Work

Multimodal Dialogue Datasets. Recent advances in vision-language modeling have accelerated progress in multimodal dialogue understanding (Meng et al. 2020). Existing dialogue datasets fall into two main types: (1) Image-grounded datasets (Shuster et al. 2020; Zheng et al. 2022; Lin et al. 2023) consist of dialogues explicitly constructed around a given image, often collected via crowdsourcing. While well-aligned, they lack the natural heterogeneity of real conversations, where not all utterances refer to images. (2) Image-sharing datasets(Zang et al. 2021; Lee et al. 2021; Feng et al. 2023; Lee et al. 2024) address this by capturing more spontaneous visual usage. For instance, MMDialog (Feng et al. 2023) collects over 1M social media conversations with images dispersed across turns, while DialogCC (Lee et al. 2024) enriches textual dialogues by inserting suitable images. However, these corpora still consist mostly of short, single-topic dialogues and lack the long-range, multi-topic structure of real-world interactions.

VLMs for Dialogue Retrieval. VLMs have achieved notable success in image-text retrieval, typically following two paradigms: (1) embedding-based VLMs (Radford et al.

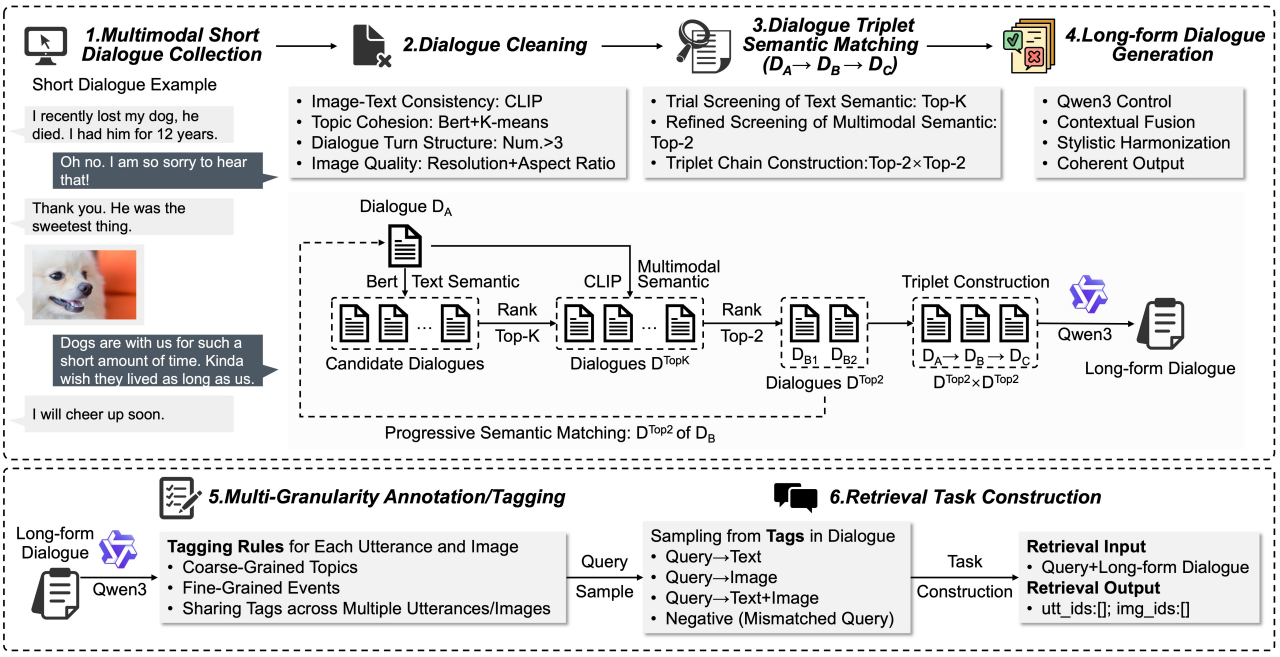

Figure 2: Overview of the MLDR construction pipeline, which integrates multimodal short dialogue processing, Qwen3-driven long-form dialogue generation, multi-granularity annotation, and retrieval-oriented task design.

2021; Li et al. 2022; Liu et al. 2022; Zhang et al. 2024; Wei et al. 2024a) that learn joint visual-text representations for efficient matching, and (2) generation-based VLMs (Liu et al. 2023a; Achiam et al. 2023; Wang et al. 2024; Jiang et al. 2024b; Liu et al. 2025a) that combine LLMs with visual encoders to generate context-aware responses via multimodal prompts. Recently, multimodal dialogue retrieval has gained traction, focusing on selecting the most appropriate sentence or image given the dialogue history (Yin et al. 2024; Choe, Oh, and Yang 2025). For example, DRIBER (Lee et al. 2023) and VCU (Wei et al. 2024b) generate intermediate image descriptions for dialogue-to-image retrieval, while IGSR (Wang et al. 2025) retrieves stickers by identifying user intent. However, these tasks are constrained to short-context response selection. In this work, we investigate a more challenging yet underexplored setting, fine-grained fragment retrieval, which aims to retrieve query-relevant fragments scattered across long-form dialogues.

VLMs with Reinforcement Learning. Recent advances like OpenAI's o1 (Jaech et al. 2024) and DeepSeek-R1 (Guo et al. 2025b) have shifted LLM research toward enhancing reasoning via reinforcement learning (RL), leading to paradigms such as GRPO (Shao et al. 2024), DAPO (Yu et al. 2025), and VAPO (Yue et al. 2025). This trend is expanding into the multimodal domain (Yuan et al. 2024; Li et al. 2025), with methods like LMM-R1 (Peng et al. 2025) introducing rule-based two-stage RL, and Reason-RFT (Tan et al. 2025) using SFT with CoT data for RL initialization. Vision-R1 (Huang et al. 2025) employs progressive reflection suppression in GRPO, while Visual-RFT (Liu et al. 2025b) enhances visual reasoning via verifiable rewards. Video-R1 (Feng et al. 2025) further extends GRPO with temporal modeling for video-language tasks. In this work, we explore RL for multimodal dialogue content retrieval, aiming to improve fine-grained reasoning and fragment localization in long-form image-text contexts.

#	Datasets	Avg. Turn	Avg. Topic
ENG	ImageChat (Shuster et al. 2020)	1.98	1.00
	OpenViDial (Meng et al. 2020)	1.00	1.00
	PhotoChat (Zang et al. 2021)	12.74	1.00
	MMDD (Lee et al. 2021)	11.56	1.00
	MMDialog (Feng et al. 2023)	4.56	1.00
	DialogCC (Lee et al. 2024)	8.20	1.00
CN	MMChat (Zheng et al. 2022)	2.59	1.00
	M3ED (Zhao et al. 2022)	9.17	1.00
	CPED (Chen et al. 2022)	11.08	1.00
	CMMA (Zhang et al. 2023)	7.27	1.00
	TikTalk (Lin et al. 2023)	2.25	1.00
	Our MLDR Dataset	**25.45**	**3.00**

Table 1: Summary of main multimodal dialogue datasets.

MLDR Dataset

This section outlines the construction of MLDR and the WeChat test set, along with their statistical analysis.

MLDR Construction

As highlighted in Tab.1, existing multimodal dialogue datasets primarily consist of short, single-topic conversations, limiting their utility in modeling the multi-topic nature of real-world dialogues. To fill this gap, we construct a multimodal long-form dialogue corpus and further develop the MLDR dataset (Fig.2). The pipeline is summarized below:

Short Dialogue Collection & Clearning. We begin with topic-rich short dialogues from DialogCC, providing a semantically diverse foundation for coherent long-form gener-

ation. To ensure data quality, we apply a multi-criteria cleaning pipeline to remove samples with poor image-text alignment, topic drift, unstructured turns, or low image quality.

Dialogue Triplet Semantic Matching. To construct coherent long-form dialogues, we design a triplet-based matching strategy that ensures topic continuity and multimodal alignment. For each short dialogue D_A, we first identify the Top-K semantically similar candidates using BERT-based sentence embeddings. We then refine this set by computing a weighted multimodal similarity score with CLIP (0.7 for text, 0.3 for image), and select the Top-2 most aligned dialogues D_B. This process is repeated for each D_B to retrieve two additional candidates D_C, forming four unique triplets of the form $D_A \rightarrow D_B \rightarrow D_C$, which serve as structural units for long-form synthesis.

Long-form Dialogue Generation. Each matched triplet is converted into a coherent long-form dialogue utilizing the Qwen3-235B (Yang et al. 2025) model with structured prompts. The generation process preserves multimodal semantics, ensures topical coherence through smooth transitions, and yields fluent, contextually grounded long-form dialogues, forming high-quality samples for downstream retrieval and reasoning tasks.

Multi-Granularity Annotation. To support fine-grained fragment retrieval, we adopt a two-level shared tagging scheme automatically generated by the Qwen3 under structured prompts. Each sentence and image caption is assigned a coarse-grained tag (e.g., domain) and a fine-grained tag (e.g., event), with each tag shared by at least two elements to form semantically consistent fragments.

Task Construction for Retrieval. We formulate a unified retrieval task over annotated long-form dialogues, where sampled coarse- or fine-grained tags serve as natural language queries, and corresponding utterances/images sharing the same tag form the retrieval fragments. To enhance diversity and robustness, we adopt a query-driven sampling strategy that generates four types of query-fragment pairs: (a) multimodal fragments; (b) utterance-only fragments; (c) image-only fragments; and (d) negative samples constructed by replacing queries with unrelated tags. This formulation supports generalization across varied retrieval scenarios.

Real-domain Evaluation on WeChat Dialogues

To evaluate fine-grained retrieval in real-world scenarios, we construct a real-domain test set from naturally occurring multimodal WeChat conversations. Unlike the synthesized MLDR data, these dialogues reflect authentic user interactions, with noisy inputs, informal expressions, and frequent topic shifts, posing greater challenges for robust retrieval.

We collect multi-turn image-utterance chats from 12 volunteers and preprocess them by removing emojis, sensitive content, profanity, and semantically void utterances. Consecutive messages from the same speaker are merged, and only dialogues with at least two images are retained. This yields 270 long-form dialogues (avg. 145.38 turns), which are segmented if exceeding 100 turns, resulting in 580 coherent samples. All dialogues are manually annotated by a professionally trained team, producing 1,250 query-dialogue pairs. Each sample includes a natural language query, its multimodal dialogue (avg. 75.38 turns), and ground-truth utterance and image IDs as retrieval targets. The task format remains consistent with MLDR, enabling direct evaluation of model generalization in open-domain fragment retrieval.

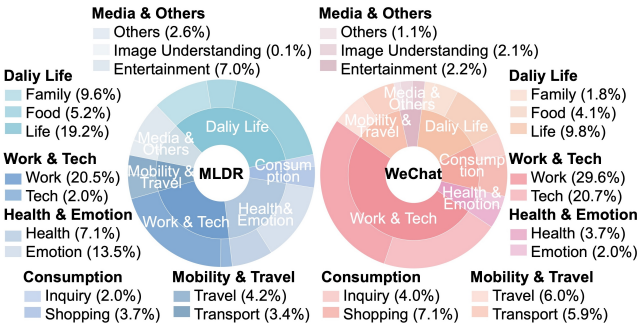

Figure 3: Comparison between the MLDR and WeChat-based datasets in terms of dialogue topic distributions.

Dataset Statistics and Analysis

To better understand the characteristics of MLDR and its real-world generalization benchmark, we compare topic diversity between MLDR and the WeChat-based test set (Fig.3). We organize dialogue content into six high-level domains (e.g., Daily Life, Work & Tech), further divided into 14 fine-grained topics. Each MLDR dialogue is deliberately constructed to span three distinct topics, promoting multi-topic contextual modeling and ensuring a balanced semantic distribution. In contrast, the WeChat-based test set exhibits natural domain bias, with over 50% of conversations centered on Work & Tech.

Methodology

Task Definition

Fine-grained fragment retrieval (FFR) can be formulated as a structured prediction task over long-form, multi-turn dialogues that include both textual utterances and images. Given a multimodal dialogue $D = \{(u_1, m_1), (u_2, m_2), \ldots, (u_T, m_T)\}$, where each turn t consists of a speaker u_t and a message m_t that may contain text or images, and a user-issued query q, the objective is to retrieve a subset of utterance and image IDs that are semantically relevant to the query. To enable the VLMs to perceive and understand the semantic and temporal structure of each turn, explicit structural markers are inserted into the dialogue: `<|utt_id_start|>...<|utt_id_end|>` to mark each utterance with a unique ID, while `<|img_id_start|>...<|img_id_end|>` marks each embedded image. The above process can be formalized as:

$$\hat{y} = \mathcal{F}(D, q) = (\hat{I}_{\text{utt}}, \hat{I}_{\text{img}}) \quad (1)$$

where $\hat{I}_{\text{utt}}$ and $\hat{I}_{\text{img}}$ denote the predicted sets of relevant utterance and image IDs, respectively.

Framework Overview

As illustrated in Fig.4, given a long-form dialogue D and a query q, F2RVLM generates structured outputs of relevant

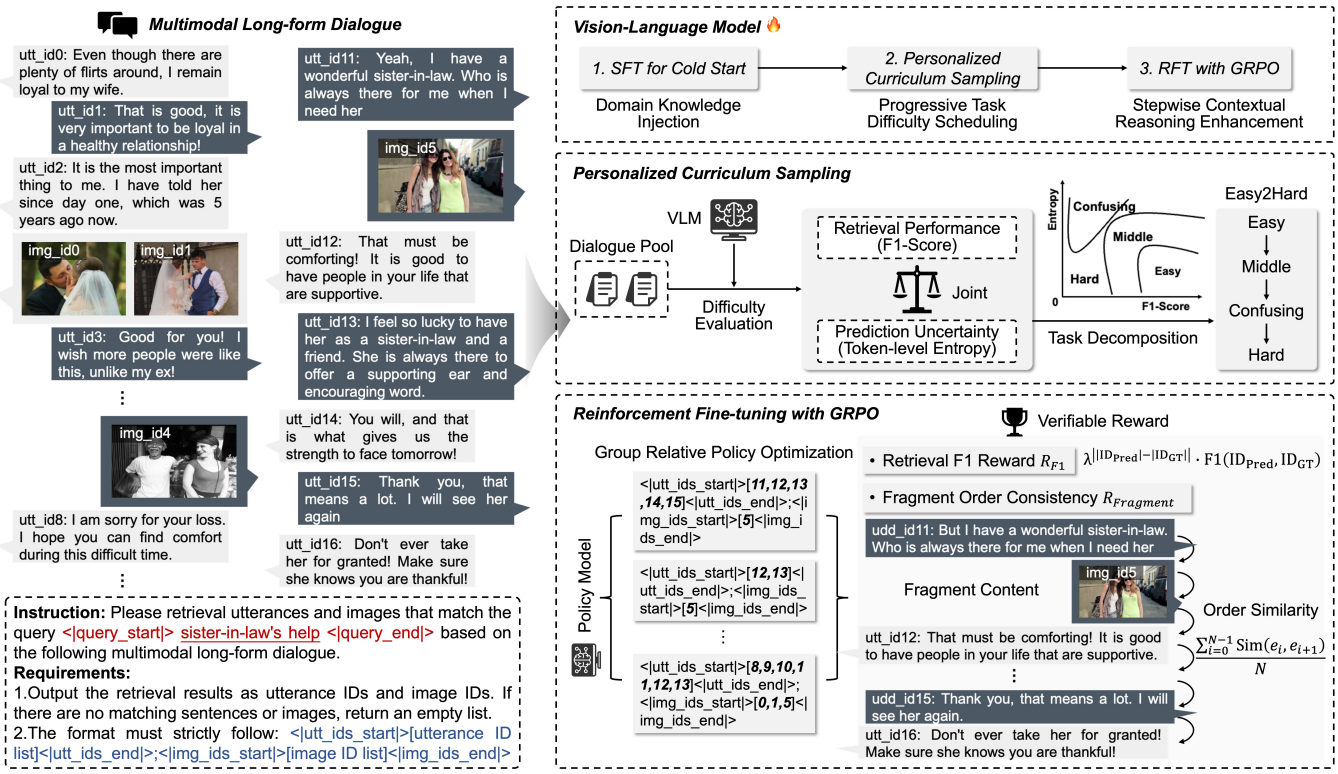

Figure 4: Overview of the F2RVLM framework for fine-grained fragment retrieval in multimodal long-form dialogue.

utterance and image IDs. To enable this behavior, the model is trained in two stages: first, supervised fine-tuning injects task-specific knowledge for fragment retrieval; then, GRPO-based reinforcement fine-tuning is conducted with a multi-objective reward that promotes semantic precision and contextual coherence. To stabilize training and enhance long-range reasoning, samples are organized into a difficulty-aware curriculum based on predicted F1 and uncertainty.

Optimize Fragment Semantic Coherence by GRPO

We adopt a rule-based variant of Group-Relative Policy Optimization (GRPO) (Shao et al. 2024) to align model behavior with human-preferred retrieval patterns. To this end, we design three verifiable reward functions targeting format compliance, retrieval accuracy, and fragment order consistency, jointly guiding the model to generate structured, precise, and contextually coherent outputs.

Format Reward (R_{Format}). R_{Format} encourages the model to strictly adhere to the expected output format:

- `<|utt_ids_start|>[...]<|utt_ids_end|>;`
- `<|img_ids_start|>[...]<|img_ids_end|>`

This binary reward returns 1 only when the output fully matches the required format. A uniqueness constraint is also enforced: if any duplicate IDs are present, the reward is set to 0, even if the overall format appears correct.

Retrieval F1 Reward (R_{F1}). To encourage precise yet well-scoped retrieval, we design a reward function based on the F1 score with an exponential penalty on length deviation. It considers both utterance and image ID overlaps between predictions and ground truth, while explicitly penalizing over- and under-retrieval:

$$R_{\text{F1}} = \sum_{m \in \{\text{utt,img}\}} \lambda_m \cdot \text{F1}(I_m^{\text{pred}}, I_m^{\text{gt}}) \cdot \gamma^{||I_m^{\text{pred}}| - |I_m^{\text{gt}}||} \quad (2)$$

where I_*^{pred} and I_*^{gt} denote the predicted and ground-truth ID sets, respectively. The weighting factors λ balances modality importance. The penalty base $\gamma \in (0, 1)$ modulates the severity of the length penalty, which increases exponentially with deviation in prediction length.

Fragment Order Consistency (R_{Fragment}). To encourage semantically coherent, contextually aligned fragment retrieval, we propose a reward based on cross-modal fragment order consistency. It evaluates whether the retrieved utterances and images maintain the natural progression of information, which is especially important in long conversations with intertwined modalities. Given the predicted sets of utterance and image IDs, we first locate the corresponding textual and visual elements in the original dialogue and encode them into a unified embedding space with CLIP. These embeddings are then interleaved and temporally ordered according to their original positions within the dialogue, forming a cross-modal sequence $\{e_1, e_2, ..., e_N\}$. The reward is defined as the average pairwise cosine similarity between adjacent embeddings in this sequence:

$$R_{\text{Fragment}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \cos(\mathbf{e}_i, \mathbf{e}_{i+1}) \quad (3)$$

Model	In-domain Val Set (MLDR)				Real-domain Test Set (WeChat)			
	Precision(%)	Recall(%)	F1(%)	MCC(%)	Precision(%)	Recall(%)	F1(%)	MCC(%)
CLIP-Embedding† (Radford et al. 2021)	48.74	30.80	42.73	21.14	20.44	49.91	31.56	17.62
BLIP2-Embedding† (Li et al. 2023)	31.88	2.96	23.47	0.00	15.22	52.74	25.05	4.94
E5-V-Embedding† (Jiang et al. 2024a)	53.85	48.83	51.82	28.92	30.33	47.13	36.99	24.40
GME-Embedding† (Zhang et al. 2024)	62.27	23.75	35.52	24.17	29.63	53.11	38.22	26.68
Qwen2.5-VL-7B† (Bai et al. 2025)	21.20	4.27	7.84	0.00	10.22	16.34	12.58	0.00
MiMo-7B-RL† (Xiaomi 2025)	67.68	55.19	61.30	45.52	41.74	24.48	30.94	23.68
Qwen2.5-VL-72B† (Bai et al. 2025)	61.11	67.14	64.09	44.61	32.55	36.95	36.60	25.26
Doubao-Seed-1.6† (Guo et al. 2025a)	73.83	42.19	54.67	42.57	55.22	28.47	38.63	34.06
Claude-Sonnet-4† (Anthropic 2025)	67.21	58.09	62.80	46.57	51.15	40.88	46.30	40.50
GPT-4o† (Jaech et al. 2024)	70.43	52.49	60.32	45.91	56.11	41.82	48.89	42.85
Gemini-2.5-Flash† (Comanici et al. 2025)	70.18	69.30	69.87	54.66	51.89	49.80	53.21	48.22
Ovis2-2B* (Lu et al. 2024)	76.97	42.48	54.82	43.98	40.83	54.51	46.69	33.16
mPLUG-Owl3-2B (Ye et al. 2024)	74.86	83.18	78.88	67.59	18.35	45.99	26.37	11.61
Qwen2-VL-2B (Wang et al. 2024)	74.97	92.99	83.07	74.18	23.20	76.03	36.65	26.55
Qwen2.5-VL-3B (Bai et al. 2025)	80.35	91.41	85.57	77.98	33.75	75.82	47.60	39.56
LLaVA-1.5-7B-hf* (Liu et al. 2023b)	68.15	92.06	78.43	66.95	20.91	**81.27**	33.61	22.99
MiMo-7B-RL (Xiaomi 2025)	80.90	**93.46**	86.71	79.78	49.46	48.22	49.05	41.84
Qwen2-VL-7B (Wang et al. 2024)	83.15	90.72	86.81	79.92	42.80	69.60	53.54	45.75
Qwen2.5-VL-7B (Bai et al. 2025)	81.52	91.42	86.23	79.00	39.99	71.30	51.71	43.65
InternVL3-8B* (Chen et al. 2024)	80.67	92.98	86.43	79.35	34.08	75.52	47.07	36.65
DeepSeek-VL2-Small-16B (Wu et al. 2024)	77.52	91.13	83.82	75.23	47.43	19.48	27.91	24.51
F2RVLM-Qwen2-VL-2B	79.64	89.55	84.35	76.07	26.97	72.00	40.46	30.90
F2RVLM-Qwen2.5-VL-3B	82.74	91.65	87.00	80.19	45.24	70.86	55.60	48.09
F2RVLM-Qwen2-VL-7B	**84.02**	90.67	**87.25**	**80.60**	**57.21**	67.46	**62.07**	**55.46**
F2RVLM-Qwen2.5-VL-7B	83.24	91.60	87.24	80.55	54.83	65.16	59.60	52.39

Table 2: Comparison with popular VLMs on the MLDR validation set and WeChat test set. "†" indicates zero-shot inference without MLDR fine-tuning. "∗" indicates models limited by context length, evaluated via sliding-window inference.

If the sequence length N is less than 2, a fallback reward (i.e., 0.5) is returned.

Difficulty-aware Curriculum Sampling

Fragment structures in long-form dialogues differ in the distribution of their internal elements: some comprise utterances or images that are tightly clustered within adjacent turns, while others span sparsely across distant positions. This variation significantly affects retrieval difficulty, clustered fragments are generally easier to locate, whereas dispersed ones are harder to handle reliably.

To leverage this inherent difficulty hierarchy, we adopt a difficulty-aware curriculum sampling that dynamically quantifies instance difficulty based on predicted F1 scores and confidence. Training begins with high-confidence, high-F1 samples (i.e., easy cases) and gradually incorporates harder ones with lower scores. This progressive schedule enables the model to first master reasoning in dense contexts, then adapt to long-range, complex scenarios, mirroring the incremental nature of human learning. Specifically, for each training sample x_i, we compute two indicators utilizing a cold-start model: (1) **Retrieval F1 Score** f_i: measures overlap between predicted and ground-truth utterance/image ID sets; and (2) **Prediction Entropy** h_i: quantifies uncertainty as the average entropy of predicted token distributions. Each instance is then assigned a difficulty level based on the following criteria:

$$d_i = \begin{cases} \text{Easy}, & f_i \geq Q_{0.75}^{(f)} \text{ and } h_i \leq Q_{0.25}^{(h)} \\ \text{Confusing}, & f_i \leq Q_{0.25}^{(f)} \text{ and } h_i \geq Q_{0.75}^{(h)} \\ \text{Hard}, & f_i \leq Q_{0.25}^{(f)} \text{ and } h_i \leq Q_{0.25}^{(h)} \\ \text{Medium}, & \text{otherwise} \end{cases} \quad (4)$$

Here, $Q_p^{(f)}$ and $Q_p^{(h)}$ denote the p-th percentiles of F1 and entropy distributions. "Easy" samples are confident and correct, "Confusing" are uncertain and incorrect, while "Hard" are incorrect yet overconfident.

Experiments

Experimental Settings

Models & Details. We implement F2RVLM based on the ms-swift (Zhao et al. 2024) framework, using Qwen-VL-series as the backbone with 2B, 3B, and 7B parameters. Parameter-efficient fine-tuning is conducted on the MLDR dataset via LoRA (Hu et al. 2022). For evaluation, we compare F2RVLM against a comprehensive set of VLMs, including both proprietary generation-based models (e.g., GPT-4o (Jaech et al. 2024), Gemini-2.5 (Comanici et al. 2025), Claude-Sonnet-4 (Anthropic 2025)) and open-source models (e.g., Qwen2.5-VL (Bai et al. 2025), DeepSeek-VL2 (Wu et al. 2024), and Mimo-VL (Xiaomi 2025)), as well as embedding-based models such as BLIP-2 (Li et al. 2023), E5-V (Jiang et al. 2024a), and GME (Zhang et al. 2024). Open-source models are fine-tuned on MLDR using SFT, while proprietary and embedding-based models are evaluated in inference-only mode.

Metrics. We evaluate fragment-level retrieval performance utilizing four metrics: Precision, Recall, F1 Score, and Matthews Correlation Coefficient (MCC), computed separately for utterance IDs and image IDs. To obtain unified scores, we average the F1 and MCC values across both modalities and calculate the harmonic mean of Precision and Recall to reflect joint retrieval performance.

Comparison with Popular VLMs

Overall Retrieval Performance. Tab.2 summarizes fragment-level retrieval results on the in-domain MLDR validation set and the real-domain WeChat test set. Key observations include: (1) F2RVLM achieves SOTA performance in both domains. The 7B variant tops MLDR

Model	#Params	Fragment Consis.	Query Sim.
Doubao-Seed-1.6†	Closed -Source	27.89	31.06
Gemini-2.5-Flash†		37.67	41.44
Claude-Sonnet-4†		42.29	46.64
GPT-4o†		43.17	48.21
Qwen2.5-VL†	72B	45.98	48.49
DeepSeek-VL2-Small	16B	12.52	13.67
MiMo-VL-RL	7B	34.69	36.69
Qwen2-VL	7B	55.28	59.49
Qwen2.5-VL	7B	54.81	58.39
F2RVLM	3B	**60.53**	**61.18**

Table 3: Comparison with popular VLMs on the real-domain WeChat test set, evaluated by Fragment Order Consistency and Query-Fragment Similarity.

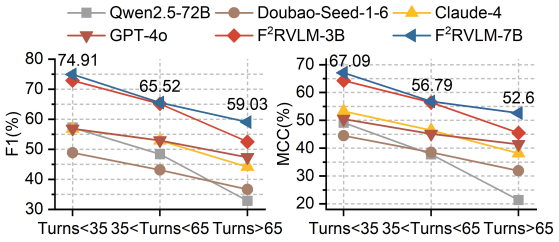

Figure 5: Performance comparison across dialogue turn groups on the real-domain WeChat test set.

with 87.25% F1, and leads on WeChat with 62.07% F1. Its 3B version also outperforms larger competitors such as MiMo-7B-RL, and GPT-4o. (2) MLDR Fine-tuning significantly boosts cross-domain retrieval. Qwen2.5-VL-7B improves from 12.58% (zero-shot) to 51.71% F1 on WeChat after MLDR tuning, demonstrating MLDR's value as a supervised resource for fragment retrieval. Moreover, F2RVLM offers less performance degradation from MLDR to WeChat compared to other models, indicating stronger generalization in real-world long-form dialogues.

Discussion about Recall Metric. While models like LLaVA achieve higher recall, they often sacrifice precision by retrieving many irrelevant fragments. In contrast, our model incorporates a penalty term in the R_{F1} to suppress over-retrieval, encouraging the selection of fewer but more semantically consistent fragments, better aligned with human preferences and fragment-level retrieval objectives.

Fragment-level Consistency and Alignment. To assess VLMs' ability to capture dialogue structure and semantic alignment, we introduce two metrics on the WeChat test set: Fragment Order Consistency (average cosine similarity between adjacent retrieved elements) and Query-Fragment Similarity (average similarity between the query and each retrieved element), as reported in Tab.3. F2RVLM achieves the highest scores on both, outperforming larger open- and closed-source models. This demonstrates its superior ability to retrieve coherent, semantically aligned fragments, enabled by the order-consistency reward in GRPO training.

Discussion about Dialogue Turns. Fig.5 reports model performance on the WeChat test set across short (<35 turns), medium (35–65 turns), and long (>65 turns) dialogues. F2RVLM-7B (Qwen2) consistently achieves the highest F1 and MCC scores as dialogue turn increases, with minimal

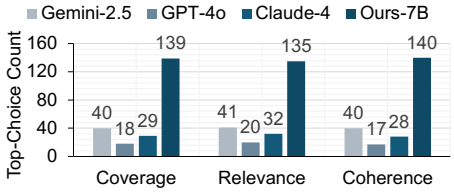

Figure 6: Human subjective evaluation results on the WeChat test set. We randomly sample 200 dialogues and ask expert annotators to select the top-performing model across three criteria: Coverage, Relevance, and Coherence.

R_{F1}	R_{Fragment}	Curriculum	In-domain		Real-domain	
			F1(%)	MCC(%)	F1(%)	MCC(%)
✗	✗	✓	86.02	78.69	53.39	46.07
✓	✗	✓	86.39	79.24	54.68	47.41
✓	✓	✗	84.69	77.53	50.65	43.71
✓	✓	✓	**87.00**	**80.19**	**55.60**	**48.09**

Table 4: Ablation study on reward components and curriculum sampling on MLDR validation and WeChat test sets.

performance degradation compared to other models, which demonstrates its robustness in long-context reasoning.

Subjective Results. We conduct a human evaluation on 200 dialogues from the WeChat test set, comparing our 7B model against Gemini-2.5, GPT-4o, and Claude-4 across three criteria: coverage, relevance, and fragment coherence. As depicted in Fig.6, our model is consistently preferred by annotators, receiving the highest top-choice counts in all aspects, demonstrating superior semantic completeness, alignment, and fluency in real-world dialogue retrieval.

Ablation Results

This section presents ablation studies to evaluate the effectiveness of the proposed reward functions and personalized curriculum learning, focusing on joint F1 and MCC metrics (Tab.4). (1) Replacing standard accuracy with R_{F1} significantly improves performance, while incorporating fragment order consistency R_{Fragment} further enhances cross-domain F1 and MCC, highlighting the role of intra-fragment semantic consistency. (2) Integrating difficulty-aware curriculum sampling into GRPO yields consistent gains: in-domain F1 improves from 84.69% to 87.00%, and real-domain F1 from 50.65% to 55.60%, validating the effectiveness of progressive learning from easier to harder samples.

Conclusion

This study defines Fine-grained Fragment Retrieval (FFR) to locate semantically coherent utterance-image fragments from long-form multimodal dialogues. To support this task, we construct MLDR, the longest-turn multimodal dialogue dataset to date, along with a real-world WeChat test set. Based on these resources, we propose F2RVLM, a generation-based retrieval model trained with GRPO-based reinforcement learning to encourage semantically coherent fragment prediction. F2RVLM surpasses popular VLMs, achieving superior accuracy in both in-domain and real-world evaluations.

Acknowledgments

This work was part of Hanbo Bi's research during his internship at Tencent WXG, under the guidance of Zhiqiang Yuan, and both made equivalent contributions. We also acknowledge the use of the publicly available multimodal dialogue dataset DialogCC (Lee et al. 2024).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Anthropic. 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, Y.; Fan, W.; Xing, X.; Pang, J.; Huang, M.; Han, W.; Tie, Q.; and Xu, X. 2022. CPED: A Large-Scale Chinese Personalized and Emotional Dialogue Dataset for Conversational AI. *arXiv:2205.14727*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Choe, S.; Oh, J.; and Yang, J. 2025. Multimodal Contrastive Learning for Dialogue Embeddings with Global and Local Views. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 155–166. Springer.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*.

Feng, J.; Sun, Q.; Xu, C.; Zhao, P.; Yang, Y.; Tao, C.; Zhao, D.; and Lin, Q. 2023. MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7348–7363.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv:2503.21776*.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025a. Seed1. 5-vl technical report. *arXiv:2505.07062*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv:2503.06749*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv:2412.16720*.

Jiang, T.; Song, M.; Zhang, Z.; Huang, H.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; and Zhuang, F. 2024a. E5-v: Universal embeddings with multimodal large language models. *arXiv:2407.12580*.

Jiang, Z.; Meng, R.; Yang, X.; Yavuz, S.; Zhou, Y.; and Chen, W. 2024b. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. *arXiv:2410.05160*.

Lee, N.; Shin, S.; Choo, J.; Choi, H.-J.; and Myaeng, S.-H. 2021. Constructing Multi-Modal Dialogue Dataset by Replacing Text with Semantically Relevant Images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 897–906.

Lee, Y.-J.; Ko, B.; Kim, H.-G.; Hyeon, J.; and Choi, H.-J. 2024. DialogCC: An Automated Pipeline for Creating High-Quality Multi-Modal Dialogue Dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1938–1963.

Lee, Y.-J.; Lee, D.; Sung, J. W.; Hyeon, J.; and Choi, H.-J. 2023. Large Language Models can Share Images, Too! *arXiv:2310.14804*.

Li, C.; Yuan, Z.; Zhang, J.; Deng, Y.; Bi, H.; Jia, Z.; Duan, X.; Luo, P.; and Zhang, J. 2025. Less Redundancy: Boosting Practicality of Vision Language Model in Walking Assistants. *arXiv preprint arXiv:2508.16070*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *LCML*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *LCML*, 12888–12900. PMLR.

Lin, H.; Ruan, L.; Xia, W.; Liu, P.; Wen, J.; Xu, Y.; Hu, D.; Song, R.; Zhao, W. X.; Jin, Q.; et al. 2023. TikTalk: a video-based dialogue dataset for multi-modal chitchat in real world. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1303–1313.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. In *NeurIPS*.

Liu, Y.; Zhang, Y.; Cai, J.; Jiang, X.; Hu, Y.; Yao, J.; Wang, Y.; and Xie, W. 2025a. Lamra: Large multimodal model

as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4015–4025.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025b. Visual-rft: Visual reinforcement fine-tuning. *arXiv:2503.01785*.

Liu, Z.; Xiong, C.; Lv, Y.; Liu, Z.; and Yu, G. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *arXiv:2209.00179*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv:2405.20797*.

Meng, Y.; Wang, S.; Han, Q.; Sun, X.; Wu, F.; Yan, R.; and Li, J. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv:2012.15015*.

Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv:2503.07536*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *LCML*, 8748–8763. PmLR.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*.

Shuster, K.; Humeau, S.; Bordes, A.; and Weston, J. 2020. Image-Chat: Engaging Grounded Conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2414–2429.

Tan, H.; Ji, Y.; Hao, X.; Lin, M.; Wang, P.; Wang, Z.; and Zhang, S. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv:2503.20752*.

Wang, B.; Du, Y.; Liang, B.; Bai, Z.; Yang, M.; Wang, B.; Wong, K.-F.; and Xu, R. 2025. A new formula for sticker retrieval: Reply with stickers in multi-modal and multi-session conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25327–25335.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*.

Wei, C.; Chen, Y.; Chen, H.; Hu, H.; Zhang, G.; Fu, J.; Ritter, A.; and Chen, W. 2024a. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 387–404. Springer.

Wei, Z.; Liao, L.; Du, X.; and Xiang, X. 2024b. Balancing Visual Context Understanding in Dialogue for Image Retrieval. In *EMNLP 2024-2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, 7929–7942. Association for Computational Linguistics (ACL).

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv:2412.10302*.

Xiaomi, L.-C.-T. 2025. MiMo-VL Technical Report. arXiv:2506.03569.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv:2505.09388*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. arXiv:2408.04840.

Yin, Z.; Hui, B.; Yang, M.; Huang, F.; and Li, Y. 2024. Dialclip: Empowering clip as multi-modal dialog retriever. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12421–12425. IEEE.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv:2503.14476*.

Yuan, Z.; Zhang, T.; Deng, Y.; Zhang, J.; Zhu, Y.; Jia, Z.; Zhou, J.; and Zhang, J. 2024. Walkvlm: Aid visually impaired people walking by vision language model. *arXiv preprint arXiv:2412.20903*.

Yue, Y.; Yuan, Y.; Yu, Q.; Zuo, X.; Zhu, R.; Xu, W.; Chen, J.; Wang, C.; Fan, T.; Du, Z.; et al. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv:2504.05118*.

Zang, X.; Liu, L.; Wang, M.; Song, Y.; Zhang, H.; and Chen, J. 2021. PhotoChat: A Human-Human Dialogue Dataset With Photo Sharing Behavior For Joint Image-Text Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6142–6152.

Zhang, X.; Zhang, Y.; Xie, W.; Li, M.; Dai, Z.; Long, D.; Xie, P.; Zhang, M.; Li, W.; and Zhang, M. 2024. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv:2412.16855*.

Zhang, Y.; Yu, Y.; Guo, Q.; Wang, B.; Zhao, D.; Uprety, S.; Song, D.; Li, Q.; and Qin, J. 2023. CMMA: benchmarking multi-affection detection in chinese multi-modal conversations. *Advances in Neural Information Processing Systems*, 36: 18794–18805.

Zhao, J.; Zhang, T.; Hu, J.; Liu, Y.; Jin, Q.; Wang, X.; and Li, H. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5699–5710.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zheng, Y.; Chen, G.; Liu, X.; and Sun, J. 2022. MMChat: Multi-Modal Chat Dataset on Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5778–5786.