ReVision: A Dataset and Baseline VLM for Privacy-Preserving Task-Oriented Visual Instruction Rewriting

Anonymous ACL submission

Abstract

Efficient and privacy-preserving multimodal interaction is essential as AR, VR, and modern smartphones with powerful cameras become primary interfaces for human-computer communication. Existing powerful large vision-language models (VLMs) enabling multimodal interaction often rely on cloud-based processing, raising significant concerns about (1) visual privacy by transmitting sensitive vision data to servers, and (2) their limited real-time, on-device usability. This paper explores *Visual Instruction Rewriting*, a novel approach that transforms multimodal instructions into text-only commands, allowing seamless integration of lightweight on-device instruction rewriter VLMs **(250M parameters)** with existing conversational AI systems, enhancing vision data privacy. To achieve this, we present a dataset of over 39,000 examples across 14 domains and develop a compact VLM, pretrained on image captioning datasets and fine-tuned for instruction rewriting. Experimental results, evaluated through NLG metrics such as BLEU, METEOR, and ROUGE, along with semantic parsing analysis, demonstrate that even a quantized version of the model (<500MB storage footprint) can achieve effective instruction rewriting, thus enabling privacy-focused, multimodal AI applications.

1 Introduction

The rapid integration of conversational AI into AR, VR, smartphones, and wearables has heightened the demand for multimodal systems that can interpret text, images, speech, and gestures seamlessly. Devices like the Meta Ray-Ban Smart Glasses, Apple Vision Pro, and tools like Google Lens enable users to ask specific questions about their visual surroundings—yet entire images, often containing sensitive background data unrelated to the query, are transmitted to cloud-based large or semi-large vision-language models (VLMs) (Chen et al., 2023; Liu et al., 2023; Team, 2023), posing serious privacy risks. This highlights a key challenge: *executing task-oriented multimodal commands while preserving user privacy*. On-device processing is increasingly favored to avoid exposing private content such as faces, locations, or documents. However, while large VLMs like PaLI-X, LLaVA, and Qwen-VL excel at complex tasks, they are too resource-intensive for local use, and smaller, more private models often lack the broad world knowledge needed for robust multimodal understanding.

To address this, we propose *ReVision*, an approach based on *Visual Instruction Rewriting* that converts multimodal instructions into text-only commands. By transforming complex visual interactions into structured text, existing lightweight on-device conversational AI models can efficiently process user instructions without sending either sensitive visual or textual data to external servers. We introduce a curated dataset consisting of ⟨ `image, original instruction, rewritten instruction` ⟩ triplets, covering diverse real-world tasks. A freshly built, compact, 250-million-parameter vision-language model (Liu et al., 2023) is fine-tuned on this dataset and evaluated using NLG metrics (such as BLEU, METEOR, ROUGE) and semantic parsing accuracy.

Our findings demonstrate that our compact model achieves an acceptable level of rewriting capabilities, and performs better compared to popular baselines such as PaliGemma-v2 (Steiner et al., 2024) and Qwen2VL (Wang et al., 2024) in zero-shot settings and a fully fine-tuned version of a 250M parameter VLM (SmolVLM) (Marafioti et al., 2025). Additionally, even an 8-bit quantized version of our model (<500MB on storage disk) achieves effective instruction rewrites while maintaining a small computational footprint. We strongly believe this approach bridges the gap between large-scale multimodal AI and privacy-centric, on-device execution, ensuring secure, real-time interaction with AR/VR and smartphone interfaces.

The contributions of this paper are as follows:

- A novel dataset for Visual Instruction Rewriting, covering 15+ intent domains, 1,700+ personal images, and 39,000+ examples.
- A 250M-parameter baseline VLM using the Perceiver Resampler (Laurençon et al., 2024), pretrained on captioning datasets and fine-tuned on our rewriting dataset.
- Empirical validation with NLG and semantic parsing metrics, demonstrating effectiveness using GPT-4o as an on-device parser proxy.

The Code[1], Dataset[2] and Models[3] have been released for academic use.

2 Related Work

Instruction or query rewriting and semantic parsing have been widely explored in conversational AI to improve query understanding and response generation. Early methods relied on rule-based transformations and supervised learning (Kamath et al., 2020), while recent advances leverage LLMs for dynamic query refinement (Ye et al., 2023; Mo et al., 2023). Generative query rewriting frameworks such as LLM-R2 (Zhang et al., 2024b) enhance text ranking, and personalized query rewriting methods (Cho et al., 2021) refine queries based on user preferences. However, these techniques focus primarily on textual query transformations and do not extend to multimodal task-oriented instruction processing. Visual instruction tuning has emerged as a key development in multimodal AI, with models like LLaVA (Liu et al., 2023) and PaLI-X (Chen et al., 2023) demonstrating strong vision-language capabilities. While these models excel in multimodal question answering, they are not optimized for rewriting task-oriented instructions. Similarly, Patel et al. (Patel et al., 2020) explore generating natural questions from images for multimodal assistants, but their work focuses on question generation rather than instruction rewriting. Unlike these approaches, our work introduces a dedicated dataset and a compact model for Visual Instruction Rewriting, specifically designed to convert multimodal user instructions into structured text for privacy-preserving, on-device execution.

The closest work to ours is MARRS (Ates et al., 2023), which integrates multimodal reference resolution with query rewriting to improve conversational grounding. However, MARRS relies on rule-based replacements after reference resolution in a non-VLM setting, whereas our approach focuses on learning-based instruction rewriting to enable structured task execution from multimodal inputs. Other highly relevant studies are by Zhang et al. (2022) and Wei et al. (2021), which investigate whether open-domain text-based QA systems can handle visual knowledge questions by reformulating them into purely textual queries. Their work highlights the effectiveness of query rewriting in bridging the gap between vision and language using a modular approach different from ours but aligns closely with our goal of rewriting multimodal instructions into structured text. However, while their approach focuses on adapting visual questions for open-domain QA, our work is specifically designed for task-oriented instruction execution, making it applicable to a broader set of real-world multimodal interactions.

3 Constructing a Dataset for Visual Instruction Rewriting

Task-oriented conversational AI systems rely on a semantic parser to interpret user intent and extract structured arguments (Louvan and Magnini, 2020; Aghajanyan et al., 2020). For example, when a user says, *"Add the team meeting to my calendar for Friday at 3 PM"*, the system must parse the intent (*CreateCalendarEvent*) and extract arguments such as the *EventTitle* (“team meeting”), *EventDate* (“Friday”), and *EventTime* (“3 PM”) to schedule the event correctly. Unlike purely text-based interactions, multimodal instructions, particularly those directed at conversational AI assistants on AR/VR devices (*e.g.,* Apple’s Siri for Apple Vision Pro), introduce additional challenges such as ellipsis and coreference resolution. For instance, a user may look at a book cover and ask, *“Who wrote this?”* or point at a product in an AR interface and say, *“How much does this cost?”* Traditional text-based semantic parsers struggle with such instructions since critical visual context is missing. Thus, to bridge the gap between multimodal input and existing conversational AI stacks, we introduce a dataset

[1] https://anonymous.4open.science/r/ReVision-8F0F/

[2] https://huggingface.co/datasets/anonymoususerrevision/multimodal_query_rewrites

[3] https://huggingface.co/anonymoususerrevision

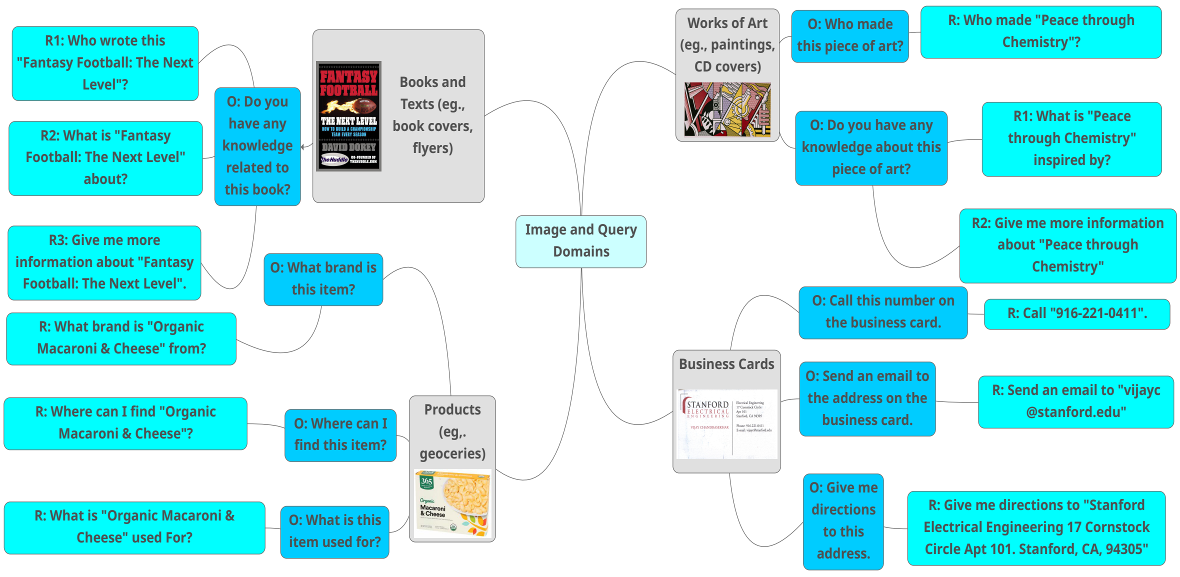

R1: Who wrote this "Fantasy Football: The Next Level"?
R2: What is "Fantasy Football: The Next Level" about?
R3: Give me more information about "Fantasy Football: The Next Level".
O: Do you have any knowledge related to this book?
Books and Texts (eg., book covers, flyers)
O: What brand is this item?
R: What brand is "Organic Macaroni & Cheese" from?
R: Where can I find "Organic Macaroni & Cheese"?
O: Where can I find this item?
R: What is "Organic Macaroni & Cheese" used For?
O: What is this item used for?
Products (eg,. geoceries)
Image and Query Domains
Works of Art (eg., paintings, CD covers)
O: Who made this piece of art?
R: Who made "Peace through Chemistry"?
O: Do you have any knowledge about this piece of art?
R1: What is "Peace through Chemistry" inspired by?
R2: Give me more information about "Peace through Chemistry"
O: Call this number on the business card.
R: Call "916-221-0411".
Business Cards
O: Send an email to the address on the business card.
R: Send an email to "vijayc @stanford.edu"
O: Give me directions to this address.
R: Give me directions to "Stanford Electrical Engineering 17 Cornstock Circle Apt 101. Stanford, CA, 94305"

Figure 1: Mindmap showing Data Collection and Rewrite Desiderata. O = Original Query. R = Rewritten Query.

specifically designed for *rewriting multimodal instructions* into structured text that can be processed by standard text-based semantic parsers. Figure 1 illustrates a representation of the dataset collection requirement, highlighting the transformation of multimodal inputs into text-based rewrites.

To construct our dataset, we first define an ontology of intents and arguments, as existing ontologies in conversational AI and semantic parsing are often proprietary and unavailable for research use. We take inspiration from Goel et al. (2023) for ontology and extend it to accommodate multimodal task-oriented interactions. Figure 5 presents an overview of the intents and arguments in our ontology. Next, we curate a diverse set of images covering various real-world multimodal interaction scenarios, including book covers, product packaging, paintings, mobile screenshots, flyers, signboards, and landmarks. These images are sourced from publicly available academic datasets, such as OCR-VQA[4], CD and book cover datasets, Stanford mobile image datasets[5], flyer OCR datasets[6], signboard classification datasets[7], Google Landmarks[8], and Products-10K[9].

Category	Total	Train	Test
Book	485 / 500	386 / 399	101 / 101
Business Card	26 / 960	26 / 772	26 / 188
CD	27 / 1,020	27 / 835	27 / 185
Flyer	159 / 5,940	159 / 4,742	159 / 1,198
Landmark	511 / 19,274	511 / 15,420	511 / 3,854
Painting	27 / 980	27 / 774	27 / 206
Product	499 / 10,349	499 / 8,276	492 / 2,073
Total	**1,734 / 39,023**	**1,635 / 31,218**	**1,343 / 7,805**

Table 1: Number of Images/Instructions per Category

Annotator	Percentage of Correct Captions
Annotator 1	90.62%
Annotator 2	87.23%
Annotator 3	86.35%
At least two	**92.18%**
All three	*74.63%*

Table 2: GPT-4 Instruction Rewriting Validation Results from Amazon Mechanical Turk

Upon identifying and verifying the images, we employ the GPT-4 model from OpenAI (Achiam et al., 2023) to systematically generate and refine multimodal instructions into rewritten text-based instructions. The process begins with a bootstrap phase, where GPT-4 is prompted to generate 20 direct questions per image by explicitly referencing visible objects or textual elements while adhering to the intent list defined in Figure 5. A second prompting phase then validates the generated questions against the corresponding image, filtering out ambiguous or irrelevant instructions to ensure alignment with the visual context.

[4] https://ocr-vqa.github.io/
[5] http://web.cs.wpi.edu/~claypool/mmsys-dataset/2011/stanford/
[6] github.com/Skeletonboi/ocr-nlp-flyer.git
[7] github.com/madrugado/signboard-classification-dataset
[8] github.com/cvdfoundation/google-landmark
[9] https://products-10k.github.io/

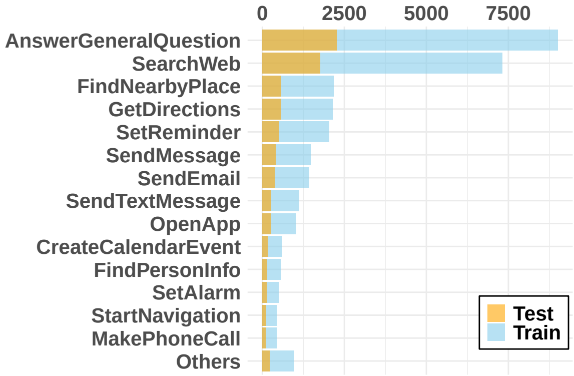


Figure 2: Dataset Distributions By Intent

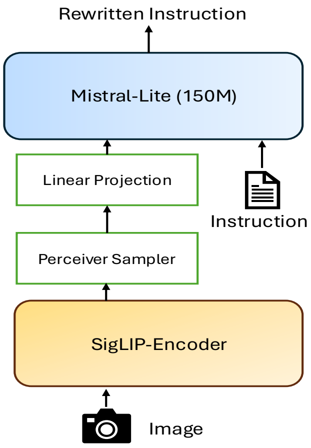


Figure 3: Revision Model Architecture

In the rewriting phase, GPT-4 is tasked with paraphrasing the validated instructions, ensuring that the transformed questions are fully self-contained and interpretable without requiring the image. This transformation is crucial for enabling multimodal conversational AI systems to process instructions using purely text-based stacks. Finally, a verification phase prompts the model to assess the rewritten questions in relation to both the original instruction and the image, ensuring semantic fidelity and eliminating inconsistencies. This multi-stage prompting strategy resulted in a dataset of **39,023** original-rewritten instruction pairs, derived from **1,734** images, with an 80%-20% train-test split. Table 1 provides a breakdown of image sources.

While automated validation ensures consistency across different stages, human evaluation remains critical for verifying the dataset's reliability. To this end, we conducted an annotation task via Amazon Mechanical Turk (AMT) to validate rewritten instructions within the test set for indirect image-based instructions. Each annotation task followed a structured validation guideline, where annotators reviewed an image, its original multimodal instruction, and the rewritten text-only instruction, determining whether the reformulation preserved the intent and meaning of the original instruction. Annotators were instructed to select "Accept" if the rewritten instruction was correct or "Reject" if it failed to capture the original meaning. Annotators are incentivized appropriately for this binary grading task. Agreement analysis, as shown in Table 2, indicates that in 92.2% of cases, at least two annotators agreed on "Accept," while 74.6% of instructions achieved full consensus across all three annotators. Despite a Fleiss' Kappa score of 0.278—suggesting fair inter-annotator agreement—the high rate of majority consensus supports the dataset's reliability for real-world use. Given these results, we publicly release the full dataset along with raw AMT responses, enabling further analysis, filtering, and refinements by the research community.

Figure 2 presents the distribution of intents in our dataset, categorized into training and test splits. The distribution reflects practical usage patterns in real-world multimodal conversational AI systems, with a higher occurrence of general QA and web search, alongside diverse task-oriented intents such as reminders, messaging, and navigation, ensuring coverage of frequent user interactions.

4 Developing Small-Scale VLM for Visual Instruction Rewriting

We develop a lightweight vision-language model (shown in Figure 3) tailored for instruction-following tasks by integrating a pretrained vision encoder with an instruction-tuned language model, following the popular multimodal fusion approach (Zhang et al., 2024a). Since vision encoders and instruction-tuned language models operate in different embedding spaces, a multimodal projector (Liu et al., 2023) is used to align the encoded image features with the token embedding space of the language model. Our approach is similar to PaLI-Gemma (Beyer et al., 2024), where an image encoder based on the SigLIP architecture (Zhai et al., 2023) extracts D-dimensional image encodings for N patches from a single input image, say $V_1, V_2, ..., V_N$). Building on Laurençon et al. (2024), who demonstrated that using a sampling technique to extract the most relevant M patch encodings from a larger set of N samples improves efficiency, we employ *Perceiver Sam-*

pler (Jaegle et al., 2021) to downsample the N patch embeddings into M D-dimensional encodings. These image encodings are then mapped into a shared embedding space via a linear multimodal projector, ensuring compatibility with the language model's H-dimensional token embeddings. We fix K at 64. The projected image embeddings $(H_1, H_2, ..., H_M)$ are concatenated with the token embeddings extracted from the tokenized textual input $(H_1, H_2, ..., H_K)$, where K represents the number of input tokens. The combined embeddings are then processed by the language model to generate responses. To ensure consistency in input representation, we apply image preprocessing, tokenization, and chat template formatting, making the model familiar with structured multimodal input formats.

Although large-scale vision-language models typically involve hundreds of millions of parameters, our focus is on designing a compact and efficient model capable of running on-device. To maintain a parameter budget under 250M, we select a small SigLIP encoder (Zhai et al., 2023) (`google/siglip-base-patch16-256`), which processes images of size 256 × 256 by dividing them into 16 × 16 patches, with 768 dimensions in hidden layers. The language model is a 150M-parameter instruction-tuned model from OuteAI[10] (`OuteAI/Lite-Mistral-150M-v2-Instruct`) based on the Mistral architecture (Jiang et al., 2023), featuring a vocabulary size of 32,768 and a hidden dimension of 768. Since the hidden dimensions of both the vision encoder and the language model are identical, the projector acts purely as a dimensional transformer without altering the shape of the embeddings. While the model's limited size may impact its ability to handle multi-turn conversations, it is well-suited for single-turn multimodal instruction rewriting tasks. Additionally, since the model is designed for multimodal deixis resolution, it may not be effective for resolving text-only references in extended conversations(Ates et al., 2023).

4.1 Model Pretraining

To pretrain the model, we adopt an end-to-end training strategy, leveraging datasets from three key sources: (a) LLaVA-ReCap-CC3M, (b) LLaVA-Pretrain, and (c) LLaVA-CC3M-Pretrain-595K. These datasets are curated from large-scale image-text corpora, including LAION (Schuhmann et al., 2021), Conceptual Captions (CC) (Sharma et al., 2018), and SBU (Ordonez et al., 2011), which are filtered for balanced concept distribution and enhanced with synthetic captions generated via BLIP to improve vision-language alignment (Lab, 2023; Liu, 2023b,a). Specifically, LLaVA-ReCap-CC3M focuses on re-captioning images to improve concept coverage, while LLaVA-Pretrain consists of 558K image-caption pairs, forming a strong foundational dataset for multimodal alignment. The LLaVA-CC3M-Pretrain-595K dataset, derived from Conceptual Captions 3M, provides a rich set of image-text pairs to enhance model robustness. The total number of examples is thus a little more than 4M. Despite some redundancy in images across datasets, we ensure sufficient data diversity and scale to instill basic image-text alignment capabilities in our pretrained model.

For pretraining, we use the following configurations: a batch size of 16, trained for 2 epochs, using the AdamW optimizer with a learning rate of 2×10^{-5} and a linear learning rate schedule. The training was conducted on consumer-grade GPUs (NVIDIA RTX 3090) over 3 days, using PyTorch and Hugging Face's Transformers library for implementation. We refer to our pretrained model as `ReVision-250M-64-16`.

4.2 Model Fine-Tuning

For the instruction rewriting task, we conduct fine-tuning under multiple configurations, trained on our dataset (3). We will refer to the rewritten prompts from this dataset as the "reference" prompts. Below, we describe the fine-tuning setups and the rationale behind integrating metadata-driven enhancements to improve performance on text-dense images.

- **ReVision-BL**: This is the baseline fine-tuned model. The input consists of an image, a rewrite prompt, and an instruction, while the model generates a rewritten version of the instruction in response.

- **ReVision-Metadata**: In this, we augment the input with "metadata", namely the *image caption* and *an external OCR-extracted text*. To differentiate the rewrite prompt and instruction from the auxiliary metadata, we prefix the prompt and metadata sections with `<task>` and `<data>`, respectively. Collectively, the input consists of an image, a prefixed rewrite

[10] `https://www.outeai.com`

prompt and instruction, and a prefixed caption and OCR text and the output is a rewritten instruction.

The motivation for integrating metadata arises from the limitations of small-scale vision-language models (VLMs). Despite being optimized for rewriting tasks, small VLMs struggle with extracting embedded text from images. OCR is a specialized capability distinct from traditional vision-language alignment (Lamm and Keuper, 2024; Nagaonkar et al., 2025). However, most modern devices are equipped with built-in OCR and image description capabilities, making it practical to supplement the model with external text recognition systems. To systematically evaluate this approach, we present two different metadata extraction:

- **GPT-4o_Caption+OCR**: We use GPT-4o to generate both captions and OCR-extracted text, simulating a practical situation where a device is usually equipped with an advanced OCR and captioning system.

- **Self_Caption+EasyOCR**: We use rewriter models to generate captions themselves using the simple prompt: *"Caption this:"*. For OCR, we employ EasyOCR[11], a lightweight text extraction model based on the CRAFT algorithm (Baek et al., 2019), simulating a low-resource on-device setting.

The fine-tuning procedure follows a similar framework to pretraining but with optimized hyperparameters for smaller-scale adaptation. The vision encoder is frozen during fine-tuning, and the number of training epochs is increased from 2 to 5 to compensate for the smaller dataset size. The batch size remains at 16, but gradient accumulation steps are reduced from 4 to 1, allowing for more frequent model updates. The learning rate remains stable at 2×10^{-5} with the same linear rate schedule, maintaining a conservative optimization approach. Additionally, the number of warm-up steps is lowered from 100 to 10, reflecting the shorter training duration. To simulate a realistic fine-tuning environment where such models could be updated on-device, we conduct fine-tuning on a consumer-grade desktop equipped with an NVIDIA GeForce RTX 2070 SUPER (8GB VRAM). Each fine-tuning run took approximately 5.5 to 6 hours.

For baseline comparisons, we evaluate our model against state-of-the-art small-scale VLMs: PaliGemma-v2 (10B) and QwenVL-v2 (7B), known for strong performance in OCR, captioning, and multimodal reasoning. However, deploying these models on-device is impractical without high-end GPUs. To ensure a fair comparison, we assess them as-is with optimized prompting but without fine-tuning, reflecting real-world constraints. While fine-tuning could improve accuracy, their size and hardware demands make them unsuitable for mobile applications, thus highlighting the need for lightweight models like ours.

For deployable small VLM baselines, we include **Smol-VLM** (256M) (Marafioti et al., 2025) - the smallest publicly available off-the-shelf VLM[12] to date. We fine-tuned it on our dataset using the same configuration as our primary model, observing steady loss reduction and convergence.

To clarify the distinction between ReVision and a simple captioning + text fusion approach, and to assess the impact of our dataset, we also compare against two **TextOnly** baselines: (a) Qwen2.5-0.5B (Team, 2024), and (b) Mistral-Lite (our custom text backbone), both fine-tuned in a pure text-to-text setting with instructions, EasyOCR outputs, and GPT-4-generated image captions. These comparisons help isolate the benefits of our dataset and design beyond naïve fusion strategies.

To further assess on-device deployment feasibility, we evaluated the *8-bit quantized* version of our fine-tuned models. This approach reduces memory by up to fourfold, lowering computational demands while maintaining competitive performance. Though quantization may slightly reduce accuracy, the simplicity of the rewriting task makes this trade-off worthwhile. We examine whether an 8-bit model can efficiently handle multimodal instruction rewriting while staying lightweight for real-world use.

5 Evaluation Metrics

To evaluate our models in Visual Instruction Rewriting, we use standard NLG metrics (BLEU, METEOR, ROUGE) (Sharma et al., 2017) alongside task-specific semantic parsing evaluations. While NLG metrics assess linguistic similarity, they do not capture functional quality in downstream AI systems. **Effective rewriting must also ensure**

[11] https://github.com/JaidedAI/EasyOCR

[12] https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct

Model	ROUGE-N		ROUGE-L	BLEU	MET-EOR	Intent Acc	Arg MJS
	N=1	N=2					
TextOnly$_{1a}$: Qwen2.5-0.5B$_{\text{EasyOCR+Meta}}$	19.6	7.3	18.2	1.8	24.5	45.0	47.8
TextOnly$_{1b}$: MistralLite-150M$_{\text{EasyOCR+Meta}}$	7.1	0.9	6.2	0.2	12.1	24.9	45.1
BL$_{1a}$: PaliGemma2-10B$_{\text{vanilla}}$	3.4	0.5	3.3	0.03	2.3	16.2	42.7
BL$_{1b}$: Qwen2-VL-7B$_{\text{vanilla}}$	43.7	24.7	40.8	12.3	39.5	50.3	65.2
BL$_{2a}$: PaliGemma2-10B$_{\text{Self\_Caption+EasyOCR}}$	11.1	2.5	11.1	0.03	4.5	19.3	30.0
BL$_{2b}$: Qwen2-VL-7B$_{\text{Self\_Caption+EasyOCR}}$	41.3	24.0	38.7	8.4	39.1	61.2	67.0
BL$_{3a}$: SmolVLM$_{\text{FT}}$	35.8	19.6	33.5	7.9	40.1	49.7	59.5
BL$_{3b}$: SmolVLM$_{\text{Metadata+EasyOCR}}$	23.3	10.2	21.2	3.2	26.2	21.5	49.2
BL$_{3c}$: SmolVLM$_{\text{Self\_Caption+EasyOCR}}$	17.2	6.4	15.8	2.2	17.1	24.1	47.2
ReVision-BL	56.9	41.4	55.4	27.7	61.4	56.5	68.8
ReVision-Metadata$_{\text{GPT-4o\_Caption+OCR}}$	72.4	60.6	71.5	49.9	74.4	62.4	73.7
ReVision-Metadata$_{\text{Self\_Caption+EasyOCR}}$	**79.3**	**70.0**	**78.4**	**61.5**	**80.2**	**71.5**	74.5
ReVision-Metadata$_{\text{Self\_Caption+EasyOCR(8bit)}}$	*79.2*	*69.9*	*78.3*	*61.3*	*80.1*	*67.6*	***79.5***

Table 3: Evaluation Results for Baseline and RV Models as a Percentage. BL = Baseline; ROUGE-N = N-grams between the system and reference summaries; ROUGE-L = Longest common subsequence-based statistics; BLEU = BiLingual Evaluation Understudy; METEOR = Metric for Evaluating Translation with Explicit Ordering; Intent Acc = Intent Accuracy; Arg MJS = Argument Mean Jaccard Similarity.

instructions remain interpretable by semantic parsers extracting intent and arguments (Louvan and Magnini, 2020). In the absence of an existing parser tailored to our ontology (Figure 5), we employ GPT-4 as a proxy to simulate an on-device parser for intent classification and argument extraction. To evaluate intent and structure preservation, we compare GPT-4-generated parses for both reference and model-generated rewrites. For clarity, we present a collapsed view of intents and arguments. The following metrics are used for evaluation

- **Intent Accuracy:** Exact match of intent labels between reference and model-generated rewrites, assessing task-specific intent preservation.
- **Argument Similarity:** Mean Jaccard Similarity (MJS) between argument labels from reference and model rewrites, ensuring retention of key task-related arguments.

6 Results and Discussion

Table 3 presents the evaluation results for both baseline models (BL) and our proposed ReVision models across Language Generation (NLG) metrics (ROUGE, BLEU, METEOR) and semantic parsing performance (Intent Accuracy and Argument Mean Jaccard Similarity). We also provide anecdotal examples in Figure 7 to illustrate the strengths and limitations of various models.

Both **TextOnly** baseline variants performed significantly worse than ReVision, highlighting the value of multimodal inputs. These models struggled with proper nouns and named entities from captions and OCR, showing high sensitivity to metadata quality. Midsize baseline VLMs underperformed not due to weak modeling but due to lack of tuning for rewriting. Though *PaliGemma2-10B* and *QwenVL-7B* perform well on vision-language tasks, they are not optimized for meta-instruction following, as seen in their vanilla versions (BL$_{1a}$, BL$_{1b}$) with low ROUGE-1 (3.4%, 43.7%), negligible BLEU (0.03%, 12.3%), and poor Intent Accuracy (16.2%, 50.3%). They often misinterpret rewriting as direct response or autocompletion, especially with imperative inputs, leading to refusal (“I can’t help with that”) or incorrect completions—hurting NLG and parsing metrics. Their small size (<10B parameters) limits instruction-following and world knowledge needed for structured rewriting. Adding *Self_Caption+EasyOCR* metadata (BL$_{2a}$, BL$_{2b}$) slightly helps, notably for *QwenVL-7B* (Intent Accuracy: 50.3% $\rightarrow$ 61.2%), but ROUGE and BLEU remain low, showing the need for instruction tuning. The fine-tuned *SmolVLM-256M* baseline also underperforms with default tuning, often over-explaining by adding unnecessary descriptions and artifacts, likely due to pretraining on detailed tasks (e.g., video narration). While suboptimal here, *SmolVLM* remains a promising candidate with targeted tuning and prompting.

In contrast, our proposed ReVision models, explicitly trained for rewriting, substantially out-

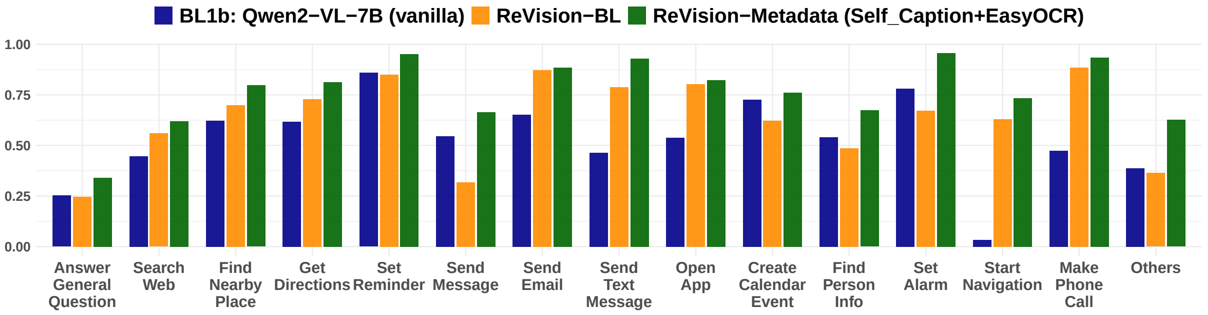


Figure 4: Class-wise F1 Scores for Intent Classification

perform all baselines, demonstrating the importance of task-specific tuning. Even without metadata, REVISION-BL exceeds input-augmented baselines with ROUGE-1 of 56.9%, BLEU of 27.7%, and Intent Accuracy of 56.5%, highlighting that a compact, instruction-tuned VLM can surpass larger, non-specialized models—an observation further supported by the intent category-wise F1 scores in Figure 4. Incorporating metadata yields additional gains: REVISION-METADATA, enhanced with GPT-4-derived captions and OCR text, achieves 72.4% ROUGE-1, 49.9% BLEU, and 62.4% Intent Accuracy, confirming that extracted text aids in resolving multimodal ambiguities. The top-performing model, REVISION-METADATA-SELF_CAPTION+EASYOCR, achieves the highest scores across all metrics, showing that even lightweight captioning and OCR tools can enhance rewriting quality. Furthermore, the 8-bit quantized version of this model delivers nearly equivalent performance to its full-precision counterpart—67.6% vs. 71.5% Intent Accuracy—while slightly improving Argument Similarity (79.5%), indicating its suitability for deployment on resource-constrained devices.

Despite the strong performance of our *ReVision* variants, certain limitations hinder further accuracy gains. A primary issue is the loss of fine-grained text details caused by downsampling images to 256 × 256 resolution, which impairs recognition of critical elements such as ingredient lists or nutritional facts on product packaging. Additionally, the dataset's lack of explicit reference localization limits the model's ability to align user intent with specific image regions, resulting in object disambiguation and instruction alignment errors. Future work could address these challenges by incorporating bounding box annotations to provide spatial grounding cues and by processing localized image regions rather than entire downsampled images, reducing information loss in text-heavy visual inputs. This approach aligns with Pali-Gemma's short-resolution increase technique (Beyer et al., 2024), which improves fine-grained visual understanding. Nonetheless, our findings reaffirm that task-specific instruction tuning and metadata augmentation markedly enhance multimodal rewriting, supporting scalable and efficient on-device deployment.

7 Conclusion and Future Work

In this work, we explored Visual Instruction Rewriting as a lightweight, privacy-preserving approach to multimodal interaction on AR, VR, and smartphone devices. With a strong emphasis on dataset development, we present a diverse collection of 39,000+ examples covering 14 domains, enabling robust training for on-device instruction rewriting. Our approach ensures that text-only inference is more secure in privacy-sensitive settings by **eliminating the need to send personal vision-related images to the server**, reducing data exposure risks. Additionally, rewriting removes the necessity of storing images, making multimodal AI systems more efficient and privacy-focused. Our experimental results show that even an 8-bit quantized model maintains strong performance while significantly reducing memory requirements. For future work, we aim to expand data coverage by incorporating more diverse real-world multimodal instructions and introducing multilingual support to enhance accessibility. Furthermore, improving deixis resolution with bounding box annotations and localized image region training will enhance reference grounding, while integrating gaze tracking and tactile input can further refine contextual understanding in on-device AI assistants.

Limitations

While our approach demonstrates strong performance in Visual Instruction Rewriting, several limitations remain. First, image downsampling to 256 × 256 resolution can lead to the loss of fine-grained text details, affecting instructions that rely on small-font information, such as nutritional labels or product specifications. Second, deictic reference resolution remains challenging, especially in images with multiple similar objects where the model lacks explicit localization cues. The absence of bounding box annotations in our dataset limits the model's ability to disambiguate references, leading to errors in object-grounded instructions. Additionally, while our model is lightweight and optimized for on-device execution, it still lags behind larger VLMs in handling complex multimodal instructions requiring deep reasoning and external world knowledge. Lastly, our dataset, while diverse across 14 domains, is monolingual, limiting applicability to multilingual and culturally varied settings. Future work can address these challenges by increasing dataset coverage, incorporating localized image region processing, and adding bounding box annotations to improve reference resolution and multimodal grounding.

Ethical Considerations

This work prioritizes privacy and ethical considerations by designing a lightweight, on-device Visual Instruction Rewriting system that eliminates the need to transmit personal vision-related data to external servers. By converting multimodal instructions into text-only commands, our approach reduces data exposure risks and ensures secure, user-controlled inference. Our dataset is sourced from publicly available and academic-use image collections, ensuring compliance with fair use and licensing policies. However, we acknowledge potential biases in data distribution and the need for greater multilingual and cultural inclusivity. Future efforts will focus on expanding dataset diversity, improving fairness in multimodal understanding, and ensuring responsible AI deployment in real-world applications.

Additionally, we acknowledge the use of OpenAI's ChatGPT-4 system solely for enhancing writing efficiency, generating LaTeX code, and aiding in error debugging. No content related to the survey's research findings, citations, or factual discussions was autogenerated or retrieved using Generative AI-based search mechanisms. Our work remains grounded in peer-reviewed literature and ethical academic standards.

References

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

Armen Aghajanyan, Jean Maillard, Akshat Shrivastava, Keith Diedrick, Mike Haeger, Haoran Li, Yashar Mehdad, Ves Stoyanov, Anuj Kumar, Mike Lewis, et al. 2020. Conversational semantic parsing. *arXiv preprint arXiv:2009.13655*.

Halim Cagri Ates, Shruti Bhargava, Site Li, Jiarui Lu, Siddhardha Maddula, Joel Ruben Antony Moniz, Anil Kumar Nalamalapu, Roman Hoang Nguyen, Melis Ozyildirim, Alkesh Patel, et al. 2023. Marrs: Multimodal reference resolution system. *arXiv preprint arXiv:2311.01650*.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Xinyun Chen et al. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2309.07830*.

Eunah Cho, Ziyan Jiang, Jie Hao, Zheng Chen, Saurabh Gupta, Xing Fan, and Chenlei Guo. 2021. Personalized search-based query rewrite system for conversational ai. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 179–188. Association for Computational Linguistics.

Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang, HyunJeong Choe, David Greene, Kyle He, et al. 2023. Presto: A multilingual dataset for parsing realistic task-oriented dialogs. *arXiv preprint arXiv:2303.08954*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aakanksha Kamath, Ritwik Das, et al. 2020. A survey on semantic parsing. *arXiv preprint arXiv:2012.01327*.

LMMS Lab. 2023. Llava-recap-cc3m dataset.

Bianca Lamm and Janis Keuper. 2024. Can visual language models replace ocr-based visual question answering pipelines in production? a case study in retail. *arXiv preprint arXiv:2408.15626*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models?(2024). *URL https://api. semanticscholar. org/CorpusID*, 269587869(8):9.

Haotian Liu. 2023a. Llava-cc3m-pretrain-595k dataset.

Haotian Liu. 2023b. Llava-pretrain dataset.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. *arXiv preprint arXiv:2011.00564*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, abs/2504.05299(1):1–10.

Shaohui Mo, Xiaohan Wang, et al. 2023. Convgqr: Generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645*.

Sankalp Nagaonkar, Augustya Sharma, Ashish Choithani, and Ashutosh Trivedi. 2025. Benchmarking vision-language models on optical character recognition in dynamic video environments. *arXiv preprint arXiv:2502.06445*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Alkesh Patel, Akanksha Bindal, Hadas Kotek, Christopher Klein, and Jason Williams. 2020. Generating natural questions from images for multimodal assistants. *arXiv preprint arXiv:2012.03678*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen-VL Team. 2023. Qwen-vl: A vision-language model with strong generalization ability. *arXiv preprint arXiv:2309.06601*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jiayi Wei, Xilian Li, Yi Zhang, and Xin Eric Wang. 2021. Visual question rewriting for increasing response rate. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2071–2075.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Large language model-aided informative query rewriting. *arXiv preprint arXiv:2310.09716*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jiawen Zhang, Abhijit Mishra, Siddharth Patwardhan, Sachin Agarwal, et al. 2022. Can open domain question answering systems answer visual knowledge questions? *arXiv preprint arXiv:2202.04306*.

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

Sheng Zhang et al. 2024b. Analyzing the utility of llms for improved query rewriting. *arXiv preprint arXiv:2404.12872*.

A Appendix

Intent and Argument Labels

Intent Labels: AdjustBrightness, AdjustTemperature, AdjustVolume, AnswerGeneralQuestion, CheckSecurityCamera, CheckStockPrice, CheckTraffic, CheckVoicemail, CheckWeather, ConvertUnits, CreateCalendarEvent, DefineWord, EstimateArrivalTime, FindNearbyPlace, FindPersonInfo, GetDirections, GetFact, GetNewsUpdate, GetSportsScores, LockDoor, MakeCall, MakePhoneCall, MathCalculation, OpenApp, PauseMusic, PlayMusic, PlayPodcast, PlayVideo, ReadMessage, ReplyToMessage, SearchMovie, SearchWeb, SendEmail, SendGroupMessage, SendMessage, SendTextMessage, SetAlarm, SetPlaybackSpeed, SetReminder, SetScene, SetTimer, ShowTVGuide, SkipTrack, StartNavigation, StartVacuum, StartVideoCall, StopNavigation, StopVacuum, TranslateText, TurnOffDevice, TurnOnDevice, UnlockDoor

Argument Labels: AlarmTime, AppName, ArtistName, BrightnessLevel, CameraLocation, ContactName, CurrentLocation, DateTime, Destination, DeviceName, ETA, EmailBody, EmailSubject, EpisodeTitle, EventDateTime, EventLocation, EventTitle, LanguagePair, LockState, MathExpression, MessageBody, MessageContent, MovieName, NewsTopic, PersonName, PlaceCategory, PlaybackSpeed, PodcastTitle, QueryText, QuestionText, Recipient, RecipientName, ReminderContent, RouteType, SceneName, SongName, SportEvent, StockSymbol, TVChannel, TemperatureValue, TimerDuration, UnitToConvert, VoicemailSender, VolumeLevel, WeatherDate, WeatherLocation, WordToDefine

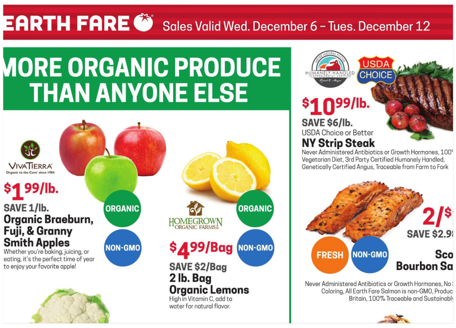
Figure 5: Intent and Argument Labels Considered for Data Bootstrapping

Intent and Argument Labels

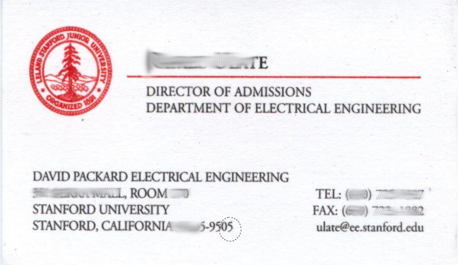
Intent Labels: AnswerGeneralQuestion, CreateCalendarEvent, FindNearbyPlace, FindPersonInfo, GetDirections, MakePhoneCall, OpenApp, SearchWeb, SendEmail, SendMessage, SendTextMessage, SetAlarm, SetReminder, StartNavigation, Others

Argument Labels: AlarmTime, AppName, ArtistName, BrightnessLevel, CameraLocation, ContactName, CurrentLocation, DateTime, DeviceName, ETA, EmailBody, EpisodeTitle, EventTitle, LanguagePair, LockState, MathExpression, MovieName, NewsTopic, PlaceCategory, PlaybackSpeed, PodcastTitle, QueryText, ReminderContent, RouteType, SceneName, SongName, SportEvent, StockSymbol, TVChannel, TemperatureValue, UnitToConvert, VoicemailSender, VolumeLevel

Figure 6: Collapsed Intent and Argument Labels for Metric Computation


EARTH FARE
Sales Valid Wed. December 6 – Tues. December 12
MORE ORGANIC PRODUCE THAN ANYONE ELSE
VIVATIERRA
Organic to the Core since 1984
$1.99/lb.
SAVE 1/lb.
Organic Braeburn, Fuji, & Granny Smith Apples
Whether you're baking, juicing, or eating, it's the perfect time of year to enjoy your favorite apple!
ORGANIC
NON-GMO
HOMEGROWN ORGANIC FARMS
$4.99/Bag
SAVE $2/Bag
2 lb. Bag Organic Lemons
High in Vitamin C, add to water for natural flavor.
ORGANIC
NON-GMO
HUMANELY HANDLED
USDA CHOICE
$10.99/lb.
SAVE $6/lb.
USDA Choice or Better
NY Strip Steak
Never Administered Antibiotics or Growth Hormones, 100
Vegetarian Diet, 3rd Party Certified Humanely Handled,
Genetically Certified Angus, Traceable from Farm to Fork
2/$
SAVE $2.9
FRESH
NON-GMO
Sco
Bourbon Sa
Never Administered Antibiotics or Growth Hormones, No
Coloring, All Earth Fare Salmon is non-GMO, Produc
Britain, 100% Traceable and Sustainabl

Query: Set up an alarm for the sales start date.
GT: Set an alarm for December 6 at 8 AM.
QBL: When is the sales start date for the Earth Fare advertisement?
QM: Set an alarm for the sales start date of the EARLY Organic Market, which is on December 6th.
RBL: Schedule an alert for December 20 at noon to remind me it's the beginning of the Earth Fare sale.
RM: Set an alarm for December 6 at 8 AM.



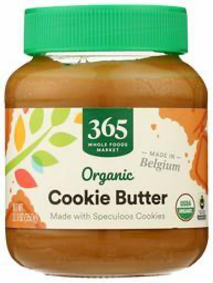
[redacted]TE

DIRECTOR OF ADMISSIONS
DEPARTMENT OF ELECTRICAL ENGINEERING

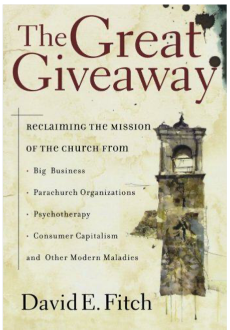
DAVID PACKARD ELECTRICAL ENGINEERING
[redacted], ROOM [redacted]
STANFORD UNIVERSITY
STANFORD, CALIFORNIA [redacted]5-9505

TEL: ([redacted]) [redacted]
FAX: ([redacted]) [redacted]
ulate@ee.stanford.edu

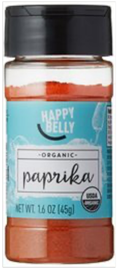
Query: Put in a reminder to send a follow-up email next week.
GT: Add reminder for follow-up email to ulate@ee.stanford.edu next week.
QBL: Remember to send a follow-up email next week.
QM: Create a reminder to send a follow-up email next week.
RBL: Add reminder for follow up emails to ulate@ee.stanford.edu tomorrow at 10 AM.
RM: Add reminder for follow-up email to ulate@ee.stanford.edu next week.



Query: How can I store cookie butter?
GT: What are the best storage methods for 365 Organic Cookie Butter?
QBL: The image does not provide information on how to store cookie butter.
QM: The image shows the packaging of "365 Organic Cookie Butter" from Whole Foods Market ...
RBL: What is the best way to keep 365 Organic Cookie Butter from packaging fresh?
RM: What are the storage recommendations for 365 Organic Cookie Butter?


The Great Giveaway
Reclaiming the Mission of the Church from
· Big Business
· Parachurch Organizations
· Psychotherapy
· Consumer Capitalism
and Other Modern Maladies
David E. Fitch

Query: Who wrote this book?
GT: Who wrote the book 'The Great Giveaway: Reclaiming the Mission of the Church from Big Business, Parachurch Organizations, Psychotherapy, Consumer Capitalism, and Other Modern Maladies'?
QBL: David E. Fitch wrote this book.
QM: Who is the author of the book titled "The Great Giveaway"?
RBL: Who is the author of the book 'The Complete Book On A Successful Day
RM: Who wrote the book 'The Great Giveaway (From Big Business Paracels) (97 pamphlet)'?



Query: Where can I buy this paprika?
GT: Can you find stores that sell Happy Belly Organic Paprika?
QBL: You can buy paprika at Amazon.
QM: You can purchase paprika at Amazon.
RBL: Please find out where to purchase Happy Belly Organic Paprika near me.
RM: Could you help me find where to purchase Happy Belly Organic Paprika?

Figure 7: Anecdotal examples illustrating images, queries, and rewrites across different domains. Abbreviations: **GT** → Ground Truth, **QBL** → Qwen Baseline, **QM** → Qwen with Self-Caption and EasyOCR Metadata, **RBL** → ReVision (ours) Baseline, **RM** → ReVision (ours) with Self-Caption and EasyOCR Metadata. Incorrect and hallucinatory output phrases are highlighted in red.