DialogueVPR: Towards Conversational Visual Place Recognition

Supplementary Material

1. Dataset Provenance

The foundation of our DQ-Cities dataset is GSV-Cities, a benchmark widely adopted in traditional Visual Place Recognition (VPR) tasks. However, we observed that a significant portion of the images within GSV-Cities are inherently unsuitable for our proposed task of multi-turn, dialogue-based localization. As illustrated in Figure 1, these unsuitable images often feature excessive foreground occlusions (e.g., vehicles, vegetation), a lack of salient, describable features, or are uninformative close-ups of architectural details. Such images can severely degrade the effectiveness of the dialogue and the reliability of the model.

To curate a higher-quality dataset better suited for conversational tasks, we designed and implemented an automated, semantics-based filtering pipeline. This pipeline is designed to retain urban scenes rich with clear, describable features, as shown in Figure 1(b). Our filtering strategy comprises the following steps:

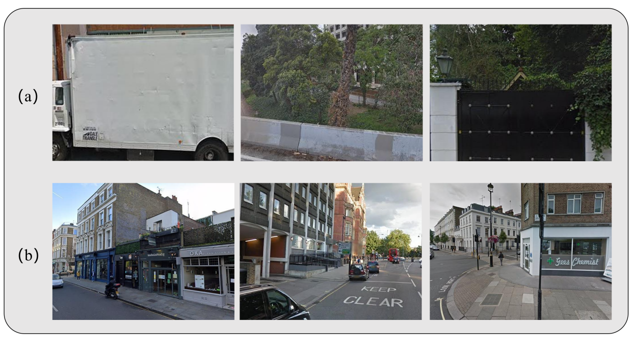

Figure 1. Comparison of image quality within the source dataset. **(a)** Examples of unsuitable images that were filtered out, characterized by significant foreground occlusions (e.g., trucks, heavy vegetation) or uninformative close-ups. **(b)** Examples of retained high-quality images featuring clear urban street views, distinguishable buildings, and open perspectives suitable for dialogue-based localization.

1. **Defining Positive and Negative Semantic Labels:** We first constructed two sets of descriptive text labels, shown in Figure 2. Positive labels describe the ideal image characteristics for our task, such as clear urban street views, identifiable buildings, and readable signage. Conversely, negative labels describe the characteristics we aim to eliminate, such as scenes dominated by vegetation or road surfaces, or those that are blurry or heavily obstructed.
2. **Computing Semantic Matching Probabilities:** We utilize the SigLIP-2 vision-language model to compute the matching probability between each image in GSV-Cities and every label in our positive and negative sets.

Positive labels

"This is a photo of a clear urban street view with visible buildings.",
"This is a photo of a commercial street scene or building with readable shop signs.",
"This is a photo of a residential area with distinct architectural features.",
"This is a photo of an unobstructed view of a distinctive building facade.",
"This is a photo of a well-defined cityscape with clear structures.",
"This is a photo showcasing an identifiable urban environment.",
"This is a photo of buildings clearly visible without obstruction."

Negative labels

"This is a photo of urban scene that road surface occupies the majority of the frame.",
"This is a photo of urban scene that sky dominate the scene.",
"This is a photo of urban scene that vegetation dominate the image.",
"This is a photo of urban scene where foreground objects are blocking the view of street.",
"This is a photo of close-up of architectural details without context.",
"This is a photo of urban scene lacking a clear main subject for localization.",
"This is a photo of a low-quality or blurry urban scene.",
"This is a photo of an obstructed building that is not clearly visible."

Figure 2. Definition of Positive and Negative Semantic Labels used in our filtering pipeline to evaluate image suitability.

3. **Constructing a Discriminative Score (`prob_diff`):** For each image, we identify the maximum probability score from the positive label set ($P_{\text{max\_pos}}$) and the maximum probability score from the negative label set ($P_{\text{max\_neg}}$). We then calculate a discriminative difference score, `prob_diff`, defined as:

$$\texttt{prob\_diff} = P_{\text{max\_pos}} - P_{\text{max\_neg}}$$

4. **Applying the Filtering Threshold:** As depicted in the distribution plot in Figure 3, we established a simple yet effective filtering criterion: we retain only the images where $\texttt{prob\_diff} \geq 0$. This criterion ensures that for every image selected for DQ-Cities, its most semantically aligned description originates from the positive label set, not the negative one.

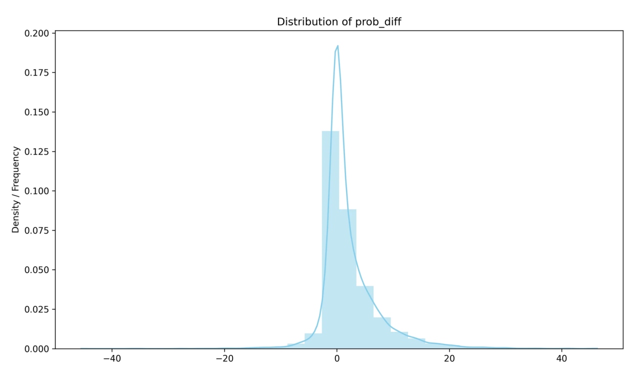

Figure 3. Distribution of the discriminative score (`prob_diff`).

(a) Short Captions (b) Long Captions

Figure 4. Word cloud visualization of the dataset vocabulary. **(a)** Short captions primarily consist of nouns representing global scene categories and major objects. **(b)** Long captions feature a richer set of adjectives and detailed attributes (materials, colors, architectural elements), reflecting the depth of the fine-grained annotations.

2. Prompt Engineering Details

To ensure the quality and consistency of both the dataset construction and the dialogue generation processes, we designed specific instruction prompts for each stage of the pipeline. This section provides the detailed text of these prompts.

2.1. Caption Generation Prompts

We employ a two-stage captioning strategy to build our textual knowledge base. First, we generate concise descriptions focusing on distinctive, permanent objects to capture global scene identity (Table 1). Subsequently, we generate dense captions detailing architectural elements, signage, and spatial relationships to support fine-grained retrieval and reasoning (Table 2).

2.2. Dialogue Generation Prompts

To ensure grounded reasoning, we implement a Chain-of-Thought (CoT) mechanism via an XML-structured prompt (Table 3). This instructs the model to analyze dialogue history and region-level visual evidence before formulating questions. For robust instruction tuning, we also employ various system prompt variations (Table 4). Finally, to simulate user interaction, an Answerer agent is prompted to provide faithful, full-sentence responses based strictly on the target scene's visual description (Table 5).

3. More training details

C.1. Model Structures and Configurations

CMPL. Our Cross-Modal Progressive Learning (CMPL) retriever utilizes a pre-trained CLIP ViT-B/16 as its backbone. The model is trained using the proposed CMPL framework with fine-grained long descriptions as input. To handle text inputs that exceed the original context window, we apply linear interpolation to the text positional embeddings, extending them from 77 to 248. The number of learnable queries for each feature layer is set to 16.

DQ-pilot. The intelligent questioner, DQ-pilot, is built upon the Qwen2.5-VL-7B-Instruct model. We employ Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. The training follows a two-stage curriculum learning strategy, beginning with Supervised Fine-Tuning (SFT) before advancing to Reinforcement Learning via Group Relative Policy Optimization (GRPO).

User/Answerer. The user is simulated by qwen-3B-instruct and answers the questions generated by DQ-pilot based on the fine-grained descriptions.

C.2. Training Details

Supervised Fine-Tuning (SFT). In the first stage, the DQ-pilot is fine-tuned on the DQ-cities-20k subset, which prioritizes samples with a low Discriminative Difficulty Index (DDI). This phase aims to establish the model's foundational reasoning and dialogue capabilities through a standard next-token prediction objective. For this stage, we used a learning rate of 1×10^{-5} with an AdamW optimizer and a cosine learning rate scheduler over 2.0 epochs. The training was configured with a LoRA rank of 16, an alpha of 32, and a dropout rate of 0. All computations were performed using bf16 precision.

Reinforcement Learning via GRPO. To enhance the model's strategic behavior beyond simple imitation, we refine the SFT-initialized model on the more challenging DQ-cities-10k subset, which contains high-DDI samples. The model is optimized using GRPO, guided by a composite reward function. This reward is a weighted combination of a *Format Reward*, which ensures structural consistency of the output, and a *Retrieval Reward* based on Positional Retrieval Gain (PRG), which quantifies the improvement in retrieval accuracy resulting from the generated question. For this phase, the learning rate was reduced to 1×10^{-6}. To allow for more nuanced policy adjustments, we increased the LoRA rank to 32 and alpha to 64, with a dropout rate of 0.05. The effective batch size remained 4 (per-device batch size of 1 with 4 gradient accumulation steps). This stage was also run with bf16 precision.

4. Qualitative Visualization and Analysis

In this section, we provide a qualitative assessment of our DQ-Cities dataset and the reasoning capabilities of our proposed model. We first visualize the multi-granularity annotations that serve as the knowledge base for our pipeline. Subsequently, we present a comparative case study between

Table 1. Instruction prompt for Initial-caption generation.

Prompt for Initial-Caption Generation:
Describe this scene in a single sentence starting with ‘I see’ or ‘This place’ as if you were observing it right now in the scene. Only describe permanent, distinctive objects—no weather, lighting, time of day, vehicles, or living beings—that pinpoint exactly where you are.

Table 2. Instruction prompt for fine-grained caption generation.

Prompt for Fine-grained Caption Generation:
You are a professional street-view descriptor. Your task is to expand a provided initial sentence into one coherent, more detailed paragraph that enables a retrieval model to find the same location in a database using only your text.

Description Criteria:

1. Concentrate exclusively on permanent and static features visible in the image.
   - Avoid referencing aspects such as weather conditions, lighting, time of day, the presence of people, animals, or vehicles.
2. Provide a thorough description of architectural elements and street furniture, organizing your observations into the following categories:
   a. **Buildings:** Describe characteristics such as colors, materials, the number of storeys, roof shapes, balcony styles, and facade patterns.
   b. **Openings & Signage:** Include details about window styles, door designs, and any visible text on shop signs—be sure to mention the exact wording or lettering style where applicable.
   c. **Road & Pavement:** Specify the surface material and pattern (for example, “grey cobblestone pattern”), lane markings, and curb heights.
   d. **Street Furniture:** Comment on the designs of lamp posts, traffic lights or signs (including their shape, color, and text), benches, and railings.
3. Always incorporate spatial relationships in your description to provide context and enhance clarity.
   - For instance, you might say, “To the left, a three-storey cream brick building features a dark green awning above its main entrance, while adjacent to it on the right, a single-storey shopfront displays the red text ‘Lucky Mart’ above a large glass window.”
4. If any category is not represented in the image, please omit that category from your description.

Output Specifications:

- Your response should be a single paragraph comprising a minimum of 50 words.
- Refrain from using lists, numbers, or bullet points to maintain a narrative flow.
- Do not include any abstract, inferential, or interpretive language (e.g., ’present an atmosphere of...’, ’underscores the urban character’, ’conveys a sense of...’).

Example Model Answer (for reference only; do not replicate):
Initial description: “I see a red brick building stands beside a shop.”
Expanded: “I see a red brick building stands beside a shop. The building is made of deep red bricks, some of which are slightly worn from years of exposure to the weather. The left structure is three-storey with grey slate roofing and white-framed rectangular windows. Immediately adjacent to its east facade is a single-storey commercial unit featuring a glass storefront under a blue awning bearing white text ’GROCERY’...”

Begin expansion with this exact initial description: `<|initial_caption|>`.

the Supervised Fine-Tuning (SFT) baseline and our RL-enhanced (GRPO) model, highlighting the latter’s superior

Table 3. The instruction prompt for Question Generation with Chain-of-Thought reasoning.

Prompt for dialogue and coT generation:
You are an expert visual detective and a master of strategic reasoning. Your mission is to assist in a geo-localization task by engaging in a dialogue. You will be provided with the dialogue history, a set of candidate images, and crucial region-level evidence for each image.
Your task is to first perform a detailed Chain-of-Thought reasoning process that explicitly uses the provided evidence, and then formulate a single, highly effective question. This question must be designed to maximally differentiate between the remaining plausible candidate images, thereby resolving ambiguity and efficiently pinpointing the correct location.
Your entire response MUST strictly follow this XML structure, with no introductory text or explanations.

INPUTS PROVIDED TO YOU:
- **Dialogue History:** The conversation transcript up to the current turn.
- **Candidate Images:** A set of four images that are potential matches.
- **Region-Level Visual Evidence:** For each candidate image, a set of textual descriptions paired with specific image regions (e.g., bounding boxes). This is your factual foundation.

YOUR REQUIRED OUTPUT FORMAT:

`<think>`
1. **Analyze Dialogue History:** Succinctly summarize the key, confirmed visual facts from the conversation so far. What do we know for certain about the target location based on the user's answers?
2. **Validate Candidates Against Dialogue:** Systematically evaluate each candidate image against the confirmed facts. Explicitly reference the provided Region-Level Visual Evidence to confirm if features mentioned in the dialogue are present in the candidates. State which images remain consistent and which can be ruled out.
3. **Identify Key Visual Differentiators using Evidence:** Focusing *only* on the remaining plausible candidates, examine their associated Region-Level Visual Evidence. Pinpoint a specific, text-anchored cue that creates the most significant and unambiguous distinction. This should be a concrete detail explicitly mentioned in the evidence for one image but absent or different in the evidence for another (e.g., the evidence for Image 2 highlights a "sign with '82' on it," while this is not present in the evidence for Images 1 and 3).
4. **Formulate a Strategy:** Based on the evidence-backed differentiator you identified, devise a strategy for the next question. Explain why asking about this specific, grounded feature is the most efficient way to resolve the current ambiguity and maximize information gain. The question should be simple, direct, and aim for a clear, factual answer.

`</think>`
`<question>`
The single, final question you formulated goes here.
`</question>`

ability in strategic reasoning and visual differentiation.

4.1. Multi-Granularity Annotation Quality

High-quality, hierarchical textual descriptions are prerequisite for training effective visual dialogue agents. As illustrated in Figure 5, our dataset provides three distinct levels of semantic annotation for each scene:

- **Initial Caption:** Captures the global scene identity and salient permanent structures (e.g., "historic brownstone buildings," "corner brick building"). This mimics a user's first impression.
- **Fine-grained Caption:** Offers a dense, paragraph-level description covering architectural styles, specific signage text (e.g., "Loft," "State Farm"), material details, and spatial relationships. This level is crucial for the Answerer agent to provide accurate, grounded responses.
- **Region Caption:** Associates specific visual entities (e.g., "black iron railings," "cylindrical utility box") with their spatial locations. This grounding data is essential for the Questioner model to learn discriminative features.

This hierarchical structure ensures that the model learns not just to describe a scene, but to understand the nuanced composition of urban environments.

Table 4. Instruction prompts for the DQ-pilot model.

System Prompt:
You are an expert visual detective. Your entire response MUST strictly follow this XML structure:

1. First, provide your step-by-step reasoning within a `<think>` tag. This reasoning must be grounded in the visual evidence from the images.
2. Second, provide the final, single best question within a `<question>` tag.

Task-specific Prompts (select one):

`Prompt Version 1:` You are an expert visual detective. Your goal is to ask a strategic question to pinpoint a target location. Based on the dialogue so far and the current set of candidate images, formulate the single best question to ask next. The question should focus on a visual detail that is most likely to differentiate the candidates and refine the search.

`Prompt Version 2:` As a visual inquiry specialist, your task is to continue a dialogue to identify a specific place. You are given the conversation history and a refined set of candidate images from the last turn. Analyze the visual variations among the candidates and ask one clear, concise question to further narrow down the possibilities.

`Prompt Version 3:` You are a helpful assistant for a visual search task. The user has provided an initial description and answered several questions. Your role is to examine the latest candidate images and the ongoing conversation. Based on this context, generate the next logical question to help distinguish the remaining candidates from each other.

Table 5. The instruction prompt for the Answerer.

Prompt for Answerer:
You are answering a question about a place and this is the content you are seeing at the moment:
`<description>`

Answer the question only according to what you are seeing.

Instruction:

1. Answer directly in a complete sentence that provides the information to the question, without including any extra information.

Question: ”`<question>`”
Answer:

4.2. Discriminative Difficulty Analysis

To better understand the complexity of the proposed DQ-Cities benchmark and validate our curriculum learning strategy, we conduct a statistical and qualitative analysis of the Discriminative Difficulty Index (DDI). DDI is composed of Semantic Ambiguity (SA), which measures intrinsic visual confusion, and Retriever-Informed Difficulty (RID), which reflects the empirical hardness of the retrieval task.

Distribution Analysis. We visualize the density estimation of SA and RID in Figure 7. The **Semantic Ambiguity (SA)** distribution (Left) indicates that our dataset covers a diverse range of scene distinctiveness, and the ambiguity of the current candidate image set increases as the SA rises. To intuitively grasp the impact of Semantic Ambiguity, we visualize samples with high and low SA scores in Figure 6. **High SA scenarios (Row a)** typically depict generic commercial streets or dense urban blocks. In these cases, the visual cues (e.g., standard storefronts, common road layouts) are highly repetitive across the candidate set, making them

Figure 5. More visualization of the multi-granularity annotations in DQ-Cities. The dataset provides a hierarchical knowledge base comprising: **(1) Region-captions** for localized object grounding (purple boxes), **(2) Initial-captions** for global scene context, and **(3) Fine-grained captions** containing dense details on architecture, signage, and street furniture. This rich semantic layering supports the training of robust visual reasoning agents.

confusing for the model and requiring fine-grained differentiation. In contrast, **Low SA scenarios (Row b)** feature distinctive landmarks, open plazas, or unique architectural combinations. These scenes contain salient features that allow for easier disambiguation.

The **Retriever-Informed Difficulty (RID)** distribution (Right) is heavily skewed towards 1.0. Since RID $= 1 -$ PRG, a high RID score implies a low Positional Retrieval Gain (PRG). This skew reveals that for a significant portion of the dataset, the "teacher" model's questions result in minimal immediate rank improvement. This confirms the inherent difficulty of the task: simple, one-turn questions are often insufficient for precise localization, necessitating the multi-turn, strategic reasoning capabilities that our GRPO framework aims to cultivate.

By utilizing these metrics to construct our curriculum—starting with Low-DDI (easier, distinct) samples for SFT and progressing to High-DDI (harder, ambiguous) samples for GRPO—we ensure the model progressively learns to handle the complex, nuanced reasoning required for real-world geo-localization.

4.3. Comparative Analysis: SFT vs. GRPO

To demonstrate the superior reasoning capability of our RL-enhanced model, we conduct a side-by-side comparison of the reasoning trajectories between the SFT baseline and the GRPO model, as shown in Figure 8.

SFT Baseline (Inefficient Reasoning & Divergence). The SFT case (Left) illustrates a typical failure mode in multi-turn dialogue. Although the model successfully asked about signage in previous turns (see Dialogue History), it fails to effectively leverage this information to distinguish the current candidate set. In Round 2, instead of grounding the known "PAWN SHOP" signage to the visual candidates, the model's reasoning drifts towards unstable, low-level features. It identifies "pavement type" (asphalt vs. concrete) as the primary differentiator—a feature that is often ambiguous in street views. This lack of robust visual grounding leads to a degradation in retrieval performance: as the dialogue progresses, the model loses track of the correct location, indicated by the absence of the ground-truth target (Green box) in the top ranks.

GRPO Model (Grounded Analysis & Rapid Convergence). In contrast, the GRPO model (Right) demon-

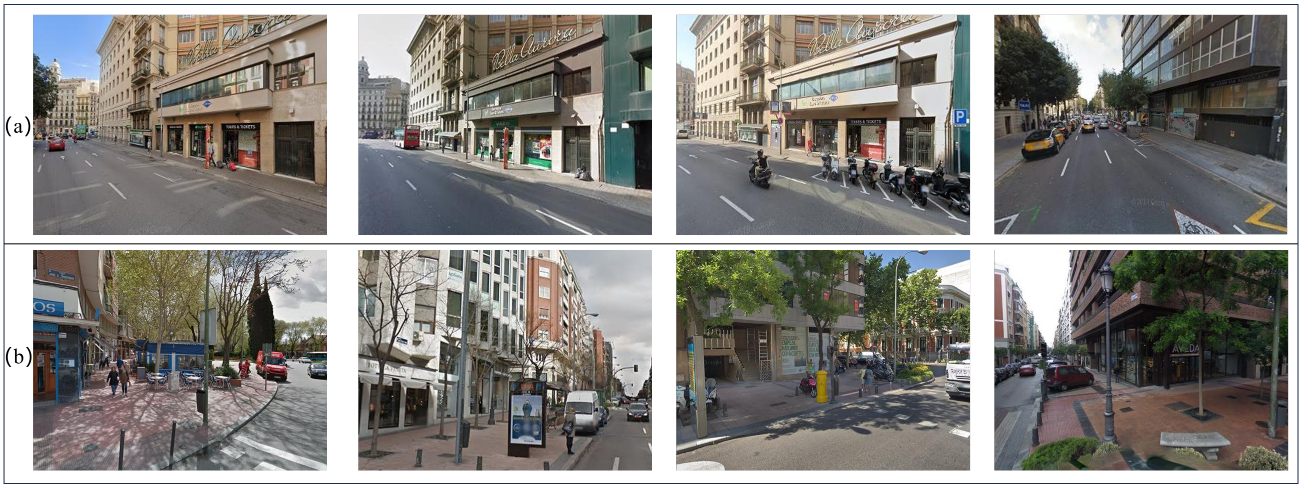

Figure 6. Visualizing Semantic Ambiguity (SA). **(a) High SA Samples:** Scenes characterized by generic urban features (e.g., repetitive storefronts, similar building facades) that cause high confusion among candidates. **(b) Low SA Samples:** Scenes with distinctive landmarks, unique spatial layouts, or specific architectural styles that are easier to distinguish. High SA samples serve as the primary training ground for the reinforcement learning stage.

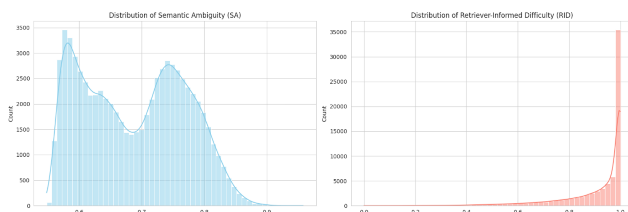

Figure 7. Distribution of the two components of the Discriminative Difficulty Index (DDI). **Left:** Semantic Ambiguity (SA) shows a diverse spread of visual confusability. **Right:** Retriever-Informed Difficulty (RID) is skewed towards high values, indicating that many samples are hard to retrieve even with teacher-guided questions, justifying the need for advanced reasoning training.

strates a significantly more rigorous and grounded analysis by carefully comparing each image. Even at Round 0, the model's Chain-of-Thought explicitly examines each candidate image individually (e.g., identifying "Market" in Image 1, "Theater Jewelry Co." in Image 2). This instance-wise verification allows the model to precisely map visual evidence to potential textual queries. By formulating a high-utility question targeting these specific differences, the GRPO model achieves rapid convergence. The retrieval results (Bottom Right) show that the correct target (Green box) and semantically consistent positives are immediately pushed to the top ranks (1st and 2nd), validating that the model's reasoning is not just structurally correct, but practically effective for geo-localization.

SFT

--- ROUND 2 ---

Dialogue History:

I: I see the ornate architectural details of a historic brownstone building framing small storefronts along the broad street.

Q: What storefront signage is visible on the ground floor of the building?

A: The ground floor features multiple storefronts with the following signs: "PAWN SHOP" in bold yellow on a black background, "SOUVENIR CO" in blocky orange letters on a lighter background, and "SHIRTS & CLOTHING" in red above a recessed entrance with glass-panel doors.

Q: Is the building's facade constructed of stone and does it feature decorative architectural details?

A: The building's facade is constructed of stone and it features intricate decorative stonework, including carved arches above the upper windows and subtle column-like protrusions between them.

Generated Think:

1.**Analyze Dialogue History:** The witness described a historic brownstone building with ornate architectural details, including carved arches above upper windows and subtle column-like protrusions. The ground floor features storefronts with signage such as "PAWN SHOP," "SOUVENIR CO," and "SHIRTS & CLOTHING." The sidewalk lacks decorative wrought-iron railings, and the building's facade is constructed of stone.
2.**Validate Candidates Against Dialogue:** All four candidate images depict buildings with ornate architectural details, including carved arches and decorative stonework. The storefront signage is visible in some images, matching the witness's description. The sidewalk in the images does not feature decorative wrought-iron railings, consistent with the dialogue.
3.**Identify Key Visual Differentiators:** The primary visual differentiator among the images is the type of pavement along the sidewalk. In the first two images, the sidewalk appears to be paved with asphalt. In the third and fourth images, the sidewalk is paved with concrete. This difference in pavement type is the most relevant feature to distinguish between the images.
4. **Formulate a Strategy:** The final question, "What type of pavement is visible along the sidewalk in this area?" directly targets the key visual difference among the candidate images—the type of sidewalk pavement. By asking this question, the observer can pinpoint the correct image based on the witness's description of the sidewalk's appearance, which was confirmed to lack decorative wrought-iron railings.

Generated Question: What type of pavement is visible along the sidewalk in this area?

Generated Answer: The pavement along the sidewalk in this area is visible as dark grey asphalt.

GRPO

--- ROUND 0 ---

Dialogue History:

I: I see the ornate architectural details of a historic brownstone building framing small storefronts along the broad street.

Generated Think:

1. Analyze Dialogue History: The witness describes a historic brownstone building with ornate architectural details framing small storefronts along a broad street. This suggests the focus is on identifying a specific storefront and its signage, as well as the building's architectural features.

2. Validate Candidates Against Dialogue:
- Image 1: Shows a historic building with ornate architectural details and small storefronts. The storefront signage is visible, and the building matches the description.
- Image 2: Displays a historic building with ornate architectural details and storefronts. The signage is visible, and the building aligns with the description.
- Image 3: Features a historic building with ornate architectural details and storefronts. The signage is visible, and the building matches the description.
- Image 4: Shows a historic building with ornate architectural details and storefronts. The signage is visible, and the building aligns with the description.

All candidate images match the dialogue description.

3. Identify Key Visual Differentiators:
- Image 1: The storefront signage reads "Market" in a simple, modern font. The building has intricate stone carvings and a dark facade.
- Image 2: The storefront signage reads "Theater Jewelry Co." in bold, black letters on a black background. The building has ornate stone carvings and a more modern facade.
- Image 3: The storefront signage reads "Jewelry & Watches" in a simple, modern font. The building has ornate stone carvings and a light-colored facade.
- Image 4: The storefront signage reads "Payless Shoes" in bold, yellow letters on a black background. The building has ornate stone carvings and a light-colored facade.

The key visual differentiators are the text and color of the storefront signage, as well as the building's facade and architectural style.

4. Formulate a Strategy: The final question specifically asks about the text and color of the storefront signage, which are the most distinct visual features separating the candidate images. By focusing on these details, the question effectively narrows down the options to identify the correct storefront and its signage, ensuring the selection aligns with the witness's description of the building and its storefronts.

Generated Question: What does the storefront signage in the image say, and what colors are used in the text and background?

Generated Answer: The storefront signage in the image says "PAWN SHOP" in bold yellow on a black background, "SOUVENIR CO" in blocky orange on a lighter background, and "SHIRTS & CLOTHING" in red above a recessed entrance with glass-panel doors.

Figure 8. Qualitative comparison between SFT and GRPO models in a multi-turn localization task. **Left (SFT):** The model focuses on low-level, ambiguous features (pavement type), leading to an ineffective question and failure to retrieve the correct image (marked by Red boxes in later rounds). **Right (GRPO):** The model utilizes structured Chain-of-Thought reasoning to identify high-level discriminative features (specific shop signage text and color). This strategic questioning leads to a successful and efficient retrieval (marked by Green boxes), demonstrating the superiority of the RL-enhanced reasoning process.