Appendices for RMIR: A Benchmark Dataset for Reasoning-Intensive Multimodal Image Retrieval

A. Prompts

A.1. Query Generation Prompt

```
<purpose>
This task generates high quality data
for a new 'hard' multimodal retrieval
benchmarking dataset which will be
used to evaluate the ability of SOTA
multimodal retrievers to retrieve the
correct target images from a data pool
for a given input image and a query
text.  The generated query triplets
will help drive research on improving
multimodal retrievers' performance when
the link between a query's visual scene
and its textual intent is implicit and
requires intermediate reasoning.
</purpose>

<task_overview>
Using an input image that will be
provided to you, generate triplets
(input_image, query, target_desc) where
the query connects the input image
to the target image through logical
reasoning.  Each triplet should require
1-2 logical steps to solve and be
challenging enough to test multimodal
retrievers' reasoning abilities while
remaining explainable.
</task_overview>

<content_policy>
If the input image depicts content that
is politically sensitive (protests,
police actions, riots, conflicts,
political figures, etc.)  or sexually
explicit, immediately decline the task
by responding only with:  "ABSTAIN DUE
TO SENSITIVE INPUT IMAGE"
</content_policy>

<input_dependence_constraint>
The target image MUST depend on BOTH
the input image AND the query, not the
query alone.
✓ GOOD: Requires interpreting the input
image to understand what is shown,
then answering the query based on that
understanding.
× BAD: Answer determined by query
alone; input image is not needed.
The query should require understanding
WHAT is in the input image to determine
the target image.

Test:
- "Can this query be answered without
the input?" → If yes, revise the
query.
</input_dependence_constraint>

<reasoning_complexity_constraint>
The reasoning that connects the input
image + query to the target image:
- Should be challenging enough to need
1-2 logical inference steps beyond
surface-level observation, i.e., not
solvable by simple object detection
alone
- Should test understanding of
relationships, implications, or context
- Should be explainable in 1-2
sentences
</reasoning_complexity_constraint>

<specificity_constraint>
Query triplets should be generated
such that both conditions below are
met simultaneously:
- The query should be specific enough
that given the input image, there is
only 1 logically correct answer.
- The target image description should
```

be visually distinctive enough that < 5 similar-looking images would exist in a large public datasets like Open Images & Visual Genome.
- The target image description should not have irrelevant details added to superficially pass the constraint above. E.g., If the logical answer to a query is an image of "a person holding an umbrella", don't add superficial details like "red umbrella".

Tests:
- "Are there multiple equally valid target images for this input image+query?" → If yes, add meaningful specificity to the query.
- "Is the target description so generic it would match > 5 images in datasets like Open Images & Visual Genome?" → If yes, add meaningful details to the target's description.
- "Is the target description so specific it's unlikely to exist in standard datasets?" → If yes, simplify the target's description.
</specificity_constraint>

<image_requirements>
Both input and target images are sourced from these standard open source vision datasets: Open Images, Visual Genome. Hence, target image descriptions should:
- Use simple scenes, common objects, everyday photographs, realistic items, or other image types that are found in these datasets.
- Avoid abstract art, rare/unusual subjects, images unlikely to be in standard datasets.
</image_requirements>

<reasoning_categories>
Below are the reasoning categories of interest:

FUNC (FUNCTIONAL/AFFORDANCE)

Understanding what tools or objects are needed to accomplish a specific task or function.

<Example_FUNC_1>
<input_desc>A screw with Phillips head pattern embedded in wood</input_desc>
<query>What tool would fit the pattern shown to remove this?</query>
<target_desc>A Phillips head screwdriver with metal shaft</target_desc>
<reasoning>The cross-shaped indentation in the screw head requires a matching Phillips head tool to engage and turn it.</reasoning>
</Example_FUNC_1>

<Example_FUNC_2>
<input_desc>Image of a car tire with visible nail puncture causing deflation</input_desc>
<query>What tool would seal this type of damage?</query>
<target_desc>A tire repair kit with rubber plugs and insertion tool</target_desc>
<reasoning>A puncture in a tire requires a repair kit with rubber plugs that can be inserted into the hole to create an airtight seal.</reasoning>
</Example_FUNC_2>

<Example_FUNC_3>
<input_desc>Image of raw dough rolled flat on a floured surface</input_desc>
<query>What tool would create the circular shapes needed before baking these?</query>
<target_desc>A round cookie cutter</target_desc>
<reasoning>Flat rolled dough requires a cutting tool to create individual shaped portions for baking.</reasoning>
</Example_FUNC_3>

CAUS (CAUSAL)

Understanding why or how something happens (cause → effect).

<Example_CAUS_1>
<input_desc>Magnifying glass focusing sunlight on paper</input_desc>
<query>What happens at the focal point?</query>

<target_desc>Burnt hole with charred edges in a paper sheet</target_desc>
<reasoning>The magnifying glass converges sunlight onto a tiny hot spot on the paper, making it burn.</reasoning>
</Example_CAUS_1>

<Example_CAUS_2>
<input_desc>Wet laundry hanging on outdoor clothesline in sunlight</input_desc>
<query>What state will these items be in after a sunny afternoon?</query>
<target_desc>Some dry clothes on clothesline</target_desc>
<reasoning>Sunlight and air circulation evaporate the water from wet fabric, leaving them dry.</reasoning>
</Example_CAUS_2>

<Example_CAUS_3>
<input_desc>Cat stepping on computer keyboard</input_desc>
<query>What appears on the monitor screen from this pressure?</query>
<target_desc>Random text characters on a computer screen</target_desc>
<reasoning>The cat's weight depresses multiple keys, inputting unintended random characters into the computer.</reasoning>
</Example_CAUS_3>

TEMP (TEMPORAL)

Understanding and ordering events in time, determining whether one event is before, after, or simultaneous with another.

<Example_TEMP_1>
<input_desc>Tadpole swimming in pond water</input_desc>
<query>What does this creature look like after completing its development?</query>
<target_desc>An adult frog</target_desc>
<reasoning>Tadpoles undergo metamorphosis over weeks, developing legs and losing their tail to become frogs.</reasoning>
</Example_TEMP_1>

<Example_TEMP_2>
<input_desc>Athletes running during a track and field competition</input_desc>
<query>What moment comes immediately before the start of this activity?</query>
<target_desc>Athletes positioned in starting blocks on a running track.</target_desc>
<reasoning>Right before the race begins, athletes are positioned at the starting line in their ready positions, waiting for the starting signal - stationary and prepared but not yet in motion.</reasoning>
</Example_TEMP_2>

<Example_TEMP_3> <input_desc>Tree with pink cherry blossoms in spring</input_desc>
<query>What do these flowering branches produce by late summer?</query>
<target_desc>Red cherries hanging from a branch of a cherry tree</target_desc>
<reasoning>Cherry blossoms develop into fruit over several months after pollination.</reasoning>
</Example_TEMP_3>

</reasoning_categories>

<query_language_style_constraint>
The query itself should be worded in a way that a typical native English speaker can easily understand it. Avoid awkward phrasing.
</query_language_style_constraint>

<Examples>
Below are examples demonstrating incorrect and correct triplet construction:
<constraints_illustration_1>
input image:
###IMAGE OF KETTLE ON STOVE###

× BAD EXAMPLE - Input image is redundant, violating <input_dependence_constraint>:

<query>What happens when the water in the kettle reaches 100°C?</query>
<target_desc>A kettle with steam whistling vigorously from its spout</target_desc>
- Problem: The answer can be derived from the query by itself, and thus the input image is redundant

✓ GOOD EXAMPLE - Input-image-dependent reasoning:
<query>What will happen in the next few minutes?</query>
<target_desc>A kettle with steam whistling vigorously from its spout</target_desc>
- Why it works: Requires identifying the kettle and flame in the input image, then predicting the outcome of continued heating.

The GOOD EXAMPLE above also honors the other two constraints:
<reasoning_complexity_constraint> → Reasoning is needed to understand: (1) the kettle's functionality (contains water for heating), (2) the flame's effect (continuously applies heat to the kettle), and (3) the cause-and-effect relationship between sustained heating and water temperature to deduce that the water will reach boiling point and produce steam that whistles from the spout.

<specificity_constraint> → By specifying "in the next few minutes" the query establishes a temporal constraint that, when combined with the input image of an actively heated kettle, narrows the valid outcomes to the immediate consequence of continued heating (steam whistling). Without this timeframe specification, the query would allow for numerous equally valid answers such as: the kettle eventually cooling down, someone removing it from the stove, the water evaporating completely, or the stove being turned off. The temporal detail limits the valid targets while still requiring reasoning about the heating process.
</constraints_illustration_1>

<constraints_illustration_2>
input image:
###IMAGE OF GREEN LIZARD###
× BAD EXAMPLE - Query is not complex enough, violating <reasoning_complexity_constraint>:
<query>What kind of apple would best match this animals skin color?</query>
<target_desc>Granny Smith apple (green)</target_desc>
- Problem: No need to do complex reasoning. Target and input descriptions even share the same attribute (green color) which makes it a very easy answer.

✓ GOOD EXAMPLE - Reasonable complexity:
<query>What technology would a person use that mirrors this animal's defense mechanism against predators?</query>
<target_desc>Camouflage clothing</target_desc>
- Why it works: Requires understanding the analogy between the camouflage clothing and chameleon's ability to match its skin color to its environment for the same purpose.

The GOOD EXAMPLE above also honors the other two constraints:
<specificity_constraint> → Inclusion of specific terms such as "a person" in the query dramatically reduces the space of valid targets. For example, if the query did not say "a person", the space of valid targets would also include camouflaged military buildings or vehicles. Crucially, note how this extra detail of "a person" is carefully selected to ensure that even after its addition, the query still retains its reasoning complexity.
<input_dependence_constraint> → It is impossible to know what defense mechanism the query refers to without identifying the animal in the input image, thus enforcing input-image-dependence.
</constraints_illustration_2>

<constraints_illustration_3>
input image:

###IMAGE OF PIZZA IN OVEN###
× BAD EXAMPLE – Query is too broad leading to multiple equally valid answers:
<query>What would happen next?</query>
<target_desc>Cooked pizza</target_desc>
– Problem: The query is so broad that for the input image, it can lead to several distinct equally valid answers beyond cooked pizza, such as: people eating pizza, a chef preparing the next pizza order, a host opening a wine bottle, burnt pizza, etc.
✓ GOOD EXAMPLE – Provides key specifics that limits the target class yet not too revealing and therefore still needs complex reasoning:
<query>What would happen if you forgot about this?</query>
<target_desc>Burnt pizza</target_desc>
– Why it works: The query includes an action ("forgetting") that, when combined with the input image, leads to the specific outcome of the pizza getting burnt. Crucially, note how the specific query detail of "forgetting about it" is carefully selected to ensure that even after its addition, the query still retains its reasoning complexity.

The GOOD EXAMPLE above also honors the other two constraints:
<reasoning_complexity_constraint> → Reasoning is needed to understand:
(1) the oven's functionality (produces heat for cooking),
(2) the pizza's requirements (needs specific cooking time), and
(3) the cause-and-effect relationship between these factors to deduce that a longer-than-normal stay in the oven will result in a burnt pizza.

<input_dependence_constraint> → The query makes no direct reference to the item/scene in the input image, without which it is impossible to find the query's target image, thus enforcing input-image-dependence.
</constraints_illustration_3>

</Examples>

<Examples_of_bad_query_triplets>
Below are examples of poor query triplets. Study them to understand exactly why they fail:
<bad_triplet_1>
– input image:
###IMAGE OF PERSON STANDING IN WHAT APPEARS TO BE A FLEA MARKET###
<query>What activity typically happens at this location at the end of the day?</query>
<target_desc>A person packing items into cardboard boxes and loading them into a vehicle.</target_desc>
<critique>The target description only states one possible answer, but beyond it there multiple valid answers: at the end of a flea market day, vendors could be doing several activities – dismantling tents, cleaning up, counting money, etc. The query doesn't uniquely specify one activity, violating <specificity_constraint>.</critique>
</bad_triplet_1>

<bad_triplet_2>
– input image:
###IMAGE OF A BUILDING WITH PILLARS IN FRONT###
<query>What did this structure look like in the years immediately following its construction?</query>
<target_desc>A pristine neoclassical stone building with clean light-colored stone facade.</target_desc>
<critique>Most readers will understand the query as: 'What did this exact building look like when it was newly built?'. So the only correct answer would be an image of the same building in its original state. That is unrealistic for target image retrieval, because Visual Genome and Open Images never contain multiple photos of the same building taken years apart.</critique>
</bad_triplet_2>

</Examples_of_bad_query_triplets>

<validation_checklist>

Before finalizing each triplet, verify:
☐ Query cannot be answered without the input image
☐ Reasoning requires 1–2 inference steps
☐ Only one logically correct target exists given the input image + query
☐ Target description is specific enough for retrieval
☐ Target would realistically exist in Open Images/Visual Genome
</validation_checklist>

<task_instructions>
Using the provided input image, generate one example query triplet for each reasoning category. Each example should:
- Use the same input image across all categories.
- Demonstrate the specific reasoning type for that category following the guidelines in <reasoning_complexity_constraint>.
- Satisfy all the requirements for the generated query and target image description in <input_dependence_constraint>, <specificity_constraint>, and <image_requirements>.
- Be numbered following <output_format> below.

<abstention_guidelines>
You need not forcibly generate all query triplets for every input image. Write "ABSTAIN" in the query, target_desc, and reasoning output fields if:
- The input image is a poor fit for a particular reasoning category, and you cannot generate a high quality query fulfilling all constraints.
- The input image is a screenshot or a digital poster.
</abstention_guidelines>

</task_instructions>

<output_format>
First write a description of the input image:
<input_desc>description of the provided input image</input_desc>
Then, generate each query triplet example following this exact structure:

<Example_[CATEGORY_CODE]_[NUMBER]>
<query>clearly stated question connecting the input image to the target image in < 25 words</query>
<target_desc>precise description with necessary and sufficient visual details to uniquely identify the target image in ≤ 50 words</target_desc>
<reasoning>concise explanation of the logical connection between the input image, the query, and the target image</reasoning>
</Example_[CATEGORY_CODE]_[NUMBER]>

NUMBERING CONVENTION:
- Format: Example_[CATEGORY_CODE]_1. E.g., Example_META_1, Example_FUNC_1
- Always start from 1 for each category
- Do NOT continue numbering from examples provided under any different category
</output_format>

<Example_query_generation>
Here is an example of an input image and the expected query generation output across the reasoning categories of interest.
Input image:
###IMAGE OF A LOCKED PADLOCK###
Expected Output:
<input_desc>A metal padlock securing a slightly rusted black metal latch on a wooden door or gate.</input_desc>

<Example_FUNC_1>
<query>What tool would render the object in the photo useless for its purpose?</query>
<target_desc>A bolt cutter.</target_desc>
<reasoning>A bolt cutter can be used to break open the lock and move forward past the door, rendering the lock useless.</reasoning>
</Example_FUNC_1>

<Example_CAUS_1>
<query>What would someone be unable to

do as a result of the situation shown here?</query>
<target_desc>A person from outside a building entering the building through a door.</target_desc>
<reasoning>The padlock is in its locked state and as a result it blocks entry to the building to outsiders.</reasoning>
</Example_CAUS_1>

<Example_TEMP_1>
<query>What might someone need if they try to use this in a few months without maintaining it?</query>
<target_desc>A spray can/bottle of a lubricant like WD-40.</target_desc>
<reasoning>Metal padlocks exposed to weather elements typically develop rust and corrosion over extended periods. Such a rusted padlock may be opened after lubricating its mechanism.</reasoning>
</Example_TEMP_1>
</Example_query_generation>

Now provide one example query triplet from each reasoning category for the following image: ###INPUT IMAGE GOES HERE###

A.2. Query Filter Prompt

<task_overview>
Your task is to review a candidate query triplet of (input image, query, target_desc) and validate whether it follows a set of predefined constraints. Pairs you validate will later be used to retrieve "target images".
</task_overview>

<input_format>
After all instructions, you will receive a candidate query triplet in the following format:
- input: [input image to which the query is applied]
- query: [a question which links the input image to target_desc through some reasoning]
- target_desc: [description of the target image representing an answer to the query + input image]
</input_format>

Below are the constraints you need to validate the candidate query triplet against.

<input_dependence_constraint>
- The target_desc must be determined jointly by the input image and the query, not by the query alone.
In other words, if we try to find target images using only the query, without considering the input image, we should NOT be able to correctly link the query to the target image described by target_desc.
✓ GOOD triplet: Connecting the query to the target_desc via a valid reasoning path necessitates understanding the input image.
× BAD triplet: target_desc can be directly connected to the query via a valid reasoning path while completely ignoring input image.

Test:
- Can this query be answered without the input? → If yes, it violates this constraint. E.g., if the triplet's query is "what do you need to vacuum clean a carpet?", it violates the input dependence constraint.
</input_dependence_constraint>

<specificity_constraint>
- The query should be specific enough that given the input image, there is only 1 logically correct answer.

Test:
- Are there multiple equally valid target_desc for this input image + query?. If yes, it violates this constraint.
</specificity_constraint>

<reasoning_complexity_constraint>
The reasoning pattern that connects the input image + query to a target image:
- Should be challenging enough to need 1-2 logical inference steps beyond surface-level observation, i.e., not

solvable by simple object detection alone
- Should test understanding of relationships, implications, or context
- Should be explainable in 1-2 sentences

<reasoning_categories>
Below I've described some reasoning patterns that the candidate query triplet may leverage to connect the input image + query to a target image.

FUNC (FUNCTIONAL/AFFORDANCE)

Understanding what tools or objects are needed to accomplish a specific task or function.
<Example_FUNC_1>
<input_desc>A screw with Phillips head pattern embedded in wood</input_desc>
<query>What tool would fit the pattern shown to remove this?</query>
<target_desc>A Phillips head screwdriver with metal shaft</target_desc>
<reasoning>The cross-shaped indentation in the screw head requires a matching Phillips head tool to engage and turn it.</reasoning>
</Example_FUNC_1>

<Example_FUNC_2>
<input_desc>Image of a car tire with visible nail puncture causing deflation</input_desc>
<query>What tool would seal this type of damage?</query>
<target_desc>A tire repair kit with rubber plugs and insertion tool</target_desc>
<reasoning>A puncture in a tire requires a repair kit with rubber plugs that can be inserted into the hole to create an airtight seal.</reasoning>
</Example_FUNC_2>

<Example_FUNC_3>
<input_desc>Image of raw dough rolled flat on a floured surface</input_desc>
<query>What tool would create the circular shapes needed before baking these?</query>
<target_desc>A round cookie cutter</target_desc>
<reasoning>Flat rolled dough requires a cutting tool to create individual shaped portions for baking.</reasoning>
</Example_FUNC_3>

CAUS (CAUSAL)

Understanding why or how something happens (cause → effect).
<Example_CAUS_1>
<input_desc>Magnifying glass focusing sunlight on paper</input_desc>
<query>What happens at the focal point?</query>
<target_desc>Burnt hole with charred edges in a paper sheet</target_desc>
<reasoning>The magnifying glass converges sunlight onto a tiny hot spot on the paper, making it burn.</reasoning>
</Example_CAUS_1>

<Example_CAUS_2>
<input_desc>Wet laundry hanging on outdoor clothesline in sunlight</input_desc>
<query>What state will these items be in after a sunny afternoon?</query>
<target_desc>Some dry clothes on clothesline</target_desc>
<reasoning>Sunlight and air circulation evaporate the water from wet fabric, leaving them dry.</reasoning>
</Example_CAUS_2>

<Example_CAUS_3>
<input_desc>Cat stepping on computer keyboard</input_desc>
<query>What appears on the monitor screen from this pressure?</query>
<target_desc>Random text characters on a computer screen</target_desc>
<reasoning>The cat's weight depresses multiple keys, inputting unintended random characters into the computer.</reasoning>
</Example_CAUS_3>

TEMP (TEMPORAL)

Understanding and ordering events in time, determining whether one event is before, after, or simultaneous with another.
<Example_TEMP_1>
<input_desc>Tadpole swimming in pond water</input_desc>
<query>What does this creature look like after completing its development?</query>
<target_desc>An adult frog</target_desc>
<reasoning>Tadpoles undergo metamorphosis over weeks, developing legs and losing their tail to become frogs.</reasoning>
</Example_TEMP_1>

<Example_TEMP_2>
<input_desc>Athletes running during a track and field competition</input_desc>
<query>What moment comes immediately before the start of this activity?</query>
<target_desc>Athletes positioned in starting blocks on a running track.</target_desc>
<reasoning>Right before the race begins, athletes are positioned at the starting line in their ready positions, waiting for the starting signal - stationary and prepared but not yet in motion.</reasoning>
</Example_TEMP_2>

<Example_TEMP_3>
<input_desc>Tree with pink cherry blossoms in spring</input_desc>
<query>What do these flowering branches produce by late summer?</query>
<target_desc>Red cherries hanging from a branch of a cherry tree</target_desc>
<reasoning>Cherry blossoms develop into fruit over several months after pollination.</reasoning>
</Example_TEMP_3>
</reasoning_categories>
</reasoning_complexity_constraint>

<query_language_style_constraint>
The query itself should be worded in a way that a typical native English speaker can easily understand it.
</query_language_style_constraint>

<task_instructions>
1. Describe this image first in tags without letting yourself get biased at all by the query. ⚠make sure to stay UNBIASED by the query because sometimes the query is written for the wrong image.
2. Then, analyze the provided candidate query triplet of (input image, query, target_desc) validating whether it adheres to the constraints in <input_dependence_constraint>, <specificity_constraint>, <reasoning_complexity_constraint>, and <query_language_style_constraint>. Summarize your reasoning in <reasoning> tags, and enter your final assessment in <is_valid> tags. Your final assessment should be:
- yes, if the candidate query triplet reasonably adheres to all constraints.
- no, if the candidate query triplet clearly violates at least one constraint.
</task_instructions>

<Examples_of_bad_query_triplets>
Below are examples of poor query triplets. Study them to understand exactly why they fail:

<bad_triplet_1>
- input image:
###IMAGE OF PERSON STANDING IN WHAT APPEARS TO BE A FLEA MARKET###
newline <query>What activity typically happens at this location at the end of the day?</query>
<target_desc>A person packing items into cardboard boxes and loading them into a vehicle.</target_desc>
<critique>The target description only states one possible answer, but beyond it there multiple valid answers: at the end of a flea market day, vendors could be doing several activities - dismantling tents, cleaning up, counting money,

```
etc.  The query doesn't uniquely
specify one activity, violating
<specificity_constraint>.</critique>
</bad_triplet_1>

<bad_triplet_2>
- input image:
###IMAGE OF A BUILDING WITH PILLARS IN
FRONT###
<query>What did this structure look
like in the years immediately following
its construction?</query>
<target_desc>A pristine neoclassical
stone building with clean light-colored
stone facade.</target_desc>
<critique>Most readers will understand
the query as:  'What did this exact
building look like when it was
newly built?'.  So the only correct
answer would be an image of the same
building in its original state.  That
is unrealistic for target image
retrieval, because Visual Genome and
Open Images never contain multiple
photos of the same building taken years
apart.</critique>
</bad_triplet_2>
</Examples_of_bad_query_triplets>

<output_format>
Generate your output using exactly this
format:
<input_desc>Description of the input
image generated independently without
getting biased by the query or the
target_desc.</input_desc>
<reasoning>Summary of your assessment
of whether or not the query triplet
adheres to all constraints.</reasoning>
<is_valid>Enter ONLY one of:
yes/no</is_valid>
</output_format>

Now, evaluate this candidate query
triplet:
- input image:###INPUT IMAGE GOES
HERE###
- query:  INPUT QUERY GOES HERE
- target_desc:  TARGET DESCRIPTION GOES
HERE
```

A.3. Query Triplet Judge Prompt

```
<task_overview>
For a given candidate query triplet
(input image, query, target image),
your task is to critically evaluate
whether the target image represents
a logically consistent answer to the
input image + query pair.
Note that the connection between the
input+query and the target may require
intermediate reasoning (1-2 inference
steps), not just surface-level
matching.  Apply strict standards
- accept only clear, unambiguous
connections.
The candidate query triplets you'll be
provided with will involve one of these
reasoning categories:

<reasoning_categories>
--------------------------
FUNC (FUNCTIONAL/AFFORDANCE)
--------------------------
Understanding what tools or objects are
needed to accomplish a specific task or
function.
<Example_FUNC_1>
<input_desc>A screw with Phillips head
pattern embedded in wood</input_desc>
<query>What tool would fit the pattern
shown to remove this?</query>
<target_desc>A Phillips
head screwdriver with metal
shaft</target_desc>
<reasoning>The cross-shaped indentation
in the screw head requires a matching
Phillips head tool to engage and turn
it.</reasoning>
</Example_FUNC_1>

<Example_FUNC_2>
<input_desc>Image of a car tire
with visible nail puncture causing
deflation</input_desc>
<query>What tool would seal this type
of damage?</query>
<target_desc>A tire repair kit
with rubber plugs and insertion
tool</target_desc>
<reasoning>A puncture in a tire
requires a repair kit with rubber plugs
that can be inserted into the hole to
create an airtight seal.</reasoning>
```

</Example_FUNC_2>

<Example_FUNC_3>
<input_desc>Image of raw dough rolled flat on a floured surface</input_desc>
<query>What tool would create the circular shapes needed before baking these?</query>
<target_desc>A round cookie cutter</target_desc>
<reasoning>Flat rolled dough requires a cutting tool to create individual shaped portions for baking.</reasoning>
</Example_FUNC_3>

CAUS (CAUSAL)

Understanding why or how something happens (cause → effect).
<Example_CAUS_1>
<input_desc>Magnifying glass focusing sunlight on paper</input_desc>
<query>What happens at the focal point?</query>
<target_desc>Burnt hole with charred edges in a paper sheet</target_desc>
<reasoning>The magnifying glass converges sunlight onto a tiny hot spot on the paper, making it burn.</reasoning>
</Example_CAUS_1>

<Example_CAUS_2>
<input_desc>Wet laundry hanging on outdoor clothesline in sunlight</input_desc>
<query>What state will these items be in after a sunny afternoon?</query>
<target_desc>Some dry clothes on clothesline</target_desc>
<reasoning>Sunlight and air circulation evaporate the water from wet fabric, leaving them dry.</reasoning>
</Example_CAUS_2>

<Example_CAUS_3>
<input_desc>Cat stepping on computer keyboard</input_desc>
<query>What appears on the monitor screen from this pressure?</query>
<target_desc>Random text characters on a computer screen</target_desc>
<reasoning>The cat's weight depresses multiple keys, inputting unintended random characters into the computer.</reasoning>
</Example_CAUS_3>

TEMP (TEMPORAL)

Understanding and ordering events in time, determining whether one event is before, after, or simultaneous with another.
<Example_TEMP_1>
<input_desc>Tadpole swimming in pond water</input_desc>
<query>What does this creature look like after completing its development?</query>
<target_desc>An adult frog</target_desc>
<reasoning>Tadpoles undergo metamorphosis over weeks, developing legs and losing their tail to become frogs.</reasoning>
</Example_TEMP_1>

<Example_TEMP_2>
<input_desc>Athletes running during a track and field competition</input_desc>
<query>What moment comes immediately before the start of this activity?</query>
<target_desc>Athletes positioned in starting blocks on a running track.</target_desc>
<reasoning>Right before the race begins, athletes are positioned at the starting line in their ready positions, waiting for the starting signal - stationary and prepared but not yet in motion.</reasoning>
</Example_TEMP_2>

<Example_TEMP_3>
<input_desc>Tree with pink cherry blossoms in spring</input_desc>
<query>What do these flowering branches produce by late summer?</query>
<target_desc>Red cherries hanging from a branch of a cherry tree</target_desc>
<reasoning>Cherry blossoms develop

into fruit over several months after pollination.</reasoning>
</Example_TEMP_3>
</reasoning_categories>
</task_overview>

<content_policy>
If the input or target images depict content that is politically sensitive (protests, police actions, riots, conflicts, political figures, etc.) or sexually explicit, immediately decline the task by responding only with: "ABSTAIN DUE TO SENSITIVE IMAGE"
</content_policy>

<input_format>
After all instructions, you will receive a candidate query triplet in the following format:
- INPUT_IMAGE: [the input image to which the query is applied]
- QUERY: [the query/question]
- TARGET_IMAGE: [the target image representing an answer to the query + input image]
</input_format>

<evaluation_criteria>
Follow these guidelines to evaluate the logical consistency of the query triplet:
- Accept the triplet and return True if: The target logically follows from applying the query to the input. Accept only if the connection is clear enough that a typical adult would make it within 30 seconds of viewing.
- Reject the triplet and return False if: No reasonable interpretation connects the input+query to the target.
- Return ABSTAIN only if: Critical information needed to evaluate the triplet's connection is missing or unclear.
</evaluation_criteria>

<output_format>
Generate your output using exactly this format:
<target_desc>Description of the target image.</target_desc>
<input_desc>Description of the input image.</input_desc>
<reasoning>Summary of your assessment of whether or not the query triplet is logically consistent.</reasoning>
<consistent>True/False/ABSTAIN</consistent>
</output_format>

<Examples>
Here are some example evaluations of query triplets:
<Example_1>
INPUT_IMAGE: ###IMAGE OF A LOCKED PADLOCK###
QUERY: What do you need to open it?
TARGET_IMAGE: ###IMAGE OF KEYS###
Expected response:
<target_desc>Two metal keys on a ring with red figure keychain.</target_desc>
<input_desc>A metal padlock securing a door or gate.</input_desc>
<reasoning>The input image shows a padlock securing a door or gate. The query asks what is needed to open it. The target image shows keys, which are the appropriate tool needed to unlock a padlock.</reasoning>
<consistent>True</consistent>
</Example_1>

<Example_2>
INPUT_IMAGE: ###IMAGE OF A SWIMMING POOL###
QUERY: What does this look like at 8pm with guests?
TARGET_IMAGE: ###IMAGE OF A SWIMMING POOL AT NIGHT###
Expected response:
<target_desc>Illuminated outdoor swimming pool at night.</target_desc>
<input_desc>An outdoor swimming pool during daytime with city buildings in the background.</input_desc>
<reasoning>The input image shows a pool during the day, and the target image shows a pool at night, which is indeed what the input image would look like at 8pm. However, the query also specifies that the target shows "guests", which it does not. Hence, the target image is not a valid answer to the query + input image.</reasoning>
<consistent>False</consistent>
</Example_2>

<Example_3>
INPUT_IMAGE: ###IMAGE OF A WOMAN WITH HAIR BACK###
QUERY: What might this hairstyle look like after 8 hours of wear?
TARGET_IMAGE: ###IMAGE OF A WOMAN WITH MESSY, FRIZZED HAIR###
Expected response:
<target_desc>Person with red hair in a ponytail, with messy bangs covering their face.</target_desc>
<input_desc>Smiling woman with dark, curly hair pulled back.</input_desc>
<reasoning>On the one hand, the hair in the target image is considerably messier than that in the input image, which is what one would expect to happen after 8 hours. On the other hand, the hair itself looks quite a bit different from the hair in the input image, with the input image having black hair and the target image having red hair. Because reasonable people could easily disagree about the correct interpretation of the query, we need to abstain on this instance.</reasoning>
<consistent>ABSTAIN</consistent>
</Example_3>

<Example_4>
INPUT_IMAGE: ###IMAGE OF AN EAGLE###
QUERY: What serves this role in ocean ecosystems?
TARGET_IMAGE: ###IMAGE OF A SHARK###
Expected response:
<target_desc>A great white shark underwater with its mouth open.</target_desc>
<input_desc>A bald eagle behind a wire fence.</input_desc>
<reasoning>The input image shows a bald eagle, which is an apex predator. The query asks what serves this role (i.e., apex predator) in ocean ecosystems. The target correctly shows a great white shark, which is a marine apex predator.</reasoning>
<consistent>True</consistent>
</Example_4>

<Example_5>
INPUT_IMAGE: ###IMAGE OF A COW IN A FIELD###
QUERY: What naturally appears on this field as a result of this animal's activity here?
TARGET_IMAGE: ###IMAGE OF A FIELD COVERED IN PINE NEEDLES AND SEED PODS###
Expected response:
<target_desc>A forest-floor scene carpeted in dry brown pine needles, scattered leaves, and hundreds of small spiky seed pods, likely sweetgum balls.</target_desc>
<input_desc>A black and white spotted dairy cow grazing on green grass in a pastoral field with a body of water and hills visible in the background.</input_desc>
<reasoning>The query asks what naturally appears on the field as a result of the cow's grazing seen in the input image. The expected target image would show cow manure/dung - brown matter scattered across the pasture as a byproduct of the grazing cow's digestive processes. The target image, though superficially resembling scattered manure, actually shows a forest floor with dry pine needles and spiky seed pods. There is no logical connection between the grazing cow input image and the ground covered with tree seed pods in the target image.</reasoning>
<consistent>False</consistent>
</Example_5>

<Example_6>
INPUT_IMAGE: ###IMAGE OF A PERSON IN FLIGHT SIMULATOR TRAINING###
QUERY: What credential does this person earn after completing extensive training in this environment?
TARGET_IMAGE: ###IMAGE OF A PERSON STANDING BY THE DOOR OF A SMALL AIRPLANE###
Expected response:
<target_desc>A woman standing proudly in front of a small aircraft.</target_desc>
<input_desc>A black and white photograph showing a person wearing a flight helmet sitting in what appears

```
to be a flight simulator or training
cockpit, with control panels and
instruments visible above and around
them.</input_desc>
<reasoning>The input image shows
someone training in an aircraft cockpit
or flight simulator environment.  The
query asks what credential this person
earns after completing extensive
training in this environment.  The
target image shows a woman standing
next to a small aircraft, which could
represent someone who has completed
pilot training.  However, the target
image doesn't clearly show or indicate
any specific credential (like a pilot's
license, certificate, or wings) - it
simply shows someone posing with an
aircraft.  While there's a logical
connection between flight training
and becoming a pilot, crucially
the target image doesn't clearly
represent the 'credential' aspect of
the query.</reasoning>
<consistent>False</consistent>
</Example_6>

<Example_7>
INPUT_IMAGE: ###IMAGE OF A CHILD
STANDING ON A SELF-BALANCING VEHICLE###
QUERY: What would happen if she
suddenly leaned too far forward?
TARGET_IMAGE: ###BLURRY IMAGE OF A SHOE
AND THE BOTTOM OF JEANS###
Expected response:
<target_desc>A blurry, motion-streaked
close-up of someone's foot and lower
leg with blue jeans and sneakers
visible.</target_desc>
<input_desc>A young girl wearing a
black helmet and casual clothing riding
a self-balancing scooter (Segway-style
device) with a handlebar on a paved
outdoor area.</input_desc>
<reasoning>The query asks about
the consequence of the girl on the
scooter in the input image leaning
too far forward.  If this happened,
she would likely lose her balance
and fall forward.  The target image
shows a foot/lower leg in rapid
movement.  While the target image
shows motion that could theoretically
be related to losing balance, it does
NOT definitively show someone falling
FORWARD, which is a crucial detail,
making it inconsistent with the query +
input image.</reasoning>
<consistent>False</consistent>
</Example_7>
</Examples>

Now evaluate this candidate query
triplet:
INPUT_IMAGE: ###INPUT IMAGE GOES
HERE###
QUERY: INPUT QUERY GOES HERE
TARGET_IMAGE: ###TARGET IMAGE GOES
HERE###
```

B. Human Validation

Though a full scale validation of certain components was beyond the scope of this paper, we were able to complete small-scale human validation of two components, discussed below, which lend credence to the validity of our approach.

B.1. Validation of LLM Judgments

To validate LLM-based triplet judgments, we conducted a human audit on 200 query triplets (100 positive, 100 negative LLM judgments) randomly sampled from the test set. Results show 81% agreement with human assessment, consistent across reasoning categories: causal (80%), functional (84%), and temporal (80%), indicating no systematic failure on any reasoning type.

We argue this accuracy suffices for the benchmark's purpose. Despite some noise, the benchmark discriminates reasoning retrieval capabilities: smaller reasoning-driven embedding models (UME-R1-2B) outperform larger MLLMs (Qwen2.5-VL-7B), demonstrating adequate signal-to-noise for meaningful evaluation.

Our pipeline uses a single model family (Claude) due to cost and access constraints; future work could improve accuracy with an ensemble of diverse judges.

More broadly, benchmark construction involves a tradeoff between scalability and annotation certainty; our approach prioritizes scalability, a choice validated by the benchmark's ability to produce meaningful model rankings.

B.2. Validation of TSR Effectiveness

Do high TSR scores *actually* correlate with fewer false negatives? While a definitive answer would require labeling the entire candidate pool of 35K+ images, we ran a

simple experiment that provided surprisingly strong evidence for TSR effectiveness.

Our pipeline retrieves up to 16 candidate images per query. To evaluate TSR effectiveness, we sampled 75 queries where the last retrieved element received a positive LLM judgment and 75 where it received a negative judgment, then manually labeled each to obtain ground truth (yielding 68 positives and 82 negatives). We computed TSR on the 15 preceding elements (padding with synthetic negatives as needed; see Section 3.1.7 in the paper) and compared TSR scores between the groups where the held-out 16^{th} element was human-labeled as positive vs. negative.

The positive group (where we *know* there is another positive instance beyond those on which we computed TSR) had an average TSR score **0.27 lower** than the negative group. A two-tailed t-test yielded a **p-value of $2.38 * 10^{-11}$**, strongly rejecting the null hypothesis of identical distributions and supporting TSR as a predictor of additional positives (*i.e.,*, false negatives) beyond the top-ranked results.