CaptionQA: Is Your Caption as Useful as the Image Itself?

Supplementary Material

1. Motivation and Overview

Figure 1 provides a conceptual overview of CaptionQA's evaluation approach. Unlike traditional text-similarity metrics that are fact-blind, multimodal benchmarks that test a different task with sparse supervision, or complex non-deterministic caption evaluation pipelines, CaptionQA measures how "useful" a caption is by testing whether it can stand in for the image on dense, taxonomy-driven question answering. This yields fine-grained diagnostics across domains and aspects, directly measuring the task-relevant information preserved in captions.

2. Question Characteristics

Our pipeline generates predominantly 4-choice multiple-choice questions, which are more challenging than binary yes/no questions. As shown in Figure 2, 87–92% of questions across domains are 4-choice, with the remaining split between 2-choice and 3-choice questions. The Natural domain has a higher proportion of binary questions (30.4%) due to yes/no attribute verification queries (e.g., "Is there a cat in the image?", "Is the door open?"). This distribution reflects the taxonomy-driven generation process, where attribute and hallucination categories naturally lend themselves to binary verification, while other categories require distinguishing between multiple plausible options.

2.1. Question Density Across Domains

The benchmark exhibits consistent question density across images within each domain, with low variance indicating systematic annotation quality. As shown in Figure 3, the Natural domain supports the highest density (66.1 questions per image on average) due to rich visual content including objects, attributes, spatial relationships, and potential hallucinations. The Document domain, while having the lowest average (41.7 questions per image), still provides comprehensive coverage of structural elements, content evaluation, and domain-specific aspects. E-commerce and Embodied AI domains show similar densities (48.6 and 46.4 respectively), reflecting their focus on product attributes and task-relevant perception.

Figures 4–7 show the distribution of questions across top-level taxonomy categories for each domain. In the Natural domain (Figure 4), Attribute questions are most prevalent (28.1%), followed by Object Existence (19.7%) and Hallucination (17.0%). Document questions (Figure 5) focus heavily on Content-Level Evaluation (30.5%) and Structural Elements (21.1%). E-commerce questions (Figure 6) are more evenly distributed across Product Information (18.0%), Contextual Information (16.7%), and Brand/Marketing (16.6%). Embodied AI questions (Figure 7) emphasize Perception (45.0%) and Spatial Context (24.8%), reflecting the task-oriented nature of robotics applications.

3. Caption Prompts

Caption quality is highly sensitive to the instruction given to the MLLM. To study how prompting affects the utility of generated captions, we evaluate each model under four captioning prompts, shared across all domains:

- **Long.** "Write a very long and detailed caption describing the given image as comprehensively as possible."
- **Short.** "Write a very short caption for the given image."
- **Simple.** "Describe this image in detail."
- **Taxonomy-Hinted.** We explicitly condition the caption on our domain-specific taxonomy. Concretely, we ask the model to "Describe this image from the following perspectives. Skip any aspect that does not apply." and then list taxonomy nodes in the form `Top-category -> Subcategory`, e.g., `Object Existence -> Object presence`, `Attribute -> Color`, etc.

These prompts yield captions of substantially different lengths, as shown in Figure 8. On average across all models and domains, the **Short** prompt produces captions of 22 words, **Simple** produces 356 words, **Long** produces 510 words, and **Taxonomy-Hinted** produces 650 words. The Taxonomy-Hinted prompt produces the longest captions, as explicitly providing the taxonomy encourages models to address all categories systematically. For the standard CaptionQA evaluation, we suggest adopting **Simple**, as it demonstrates good performance across models while maintaining reasonable caption length.

4. Taxonomy Structure

Figure 9 presents the complete hierarchical taxonomy structure across all four CaptionQA domains. The taxonomy guides question generation and ensures comprehensive coverage of domain-specific aspects that captions should capture.

Each domain is organized into top-level categories (6–7 per domain) and their corresponding subcategories (15–22 per domain, 69 total across all domains). The hierarchical structure balances comprehensive coverage with manageable annotation complexity, where top-level categories capture broad semantic areas while subcategories provide specific evaluation dimensions.

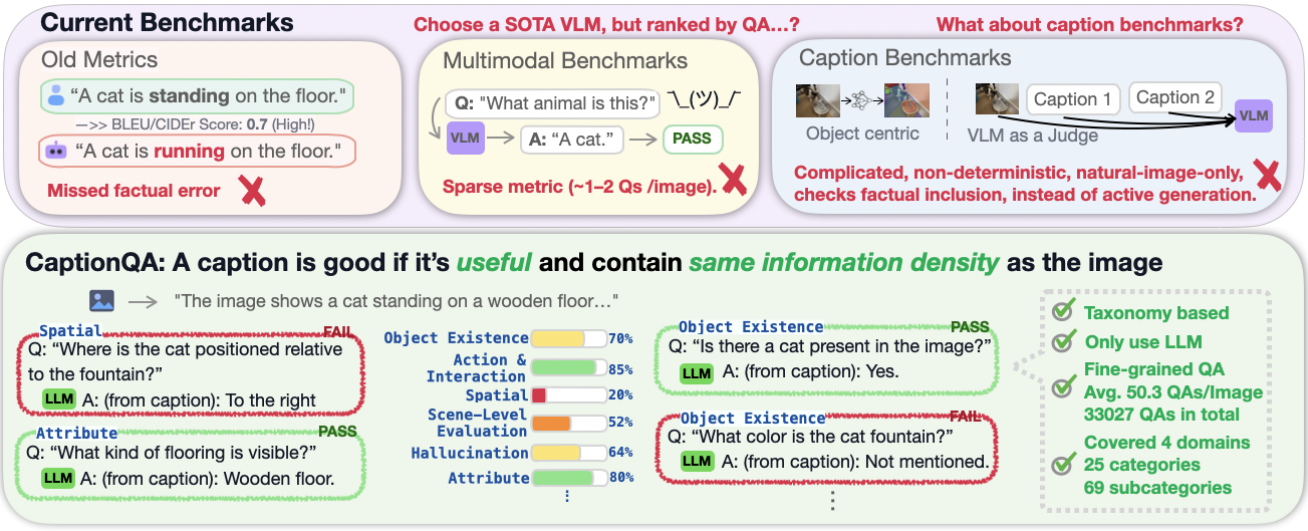

Figure 1. **Defining and evaluating "useful" captions.** Existing practices are either fact-blind (text-similarity metrics) or test a different task with sparse supervision (multimodal benchmarks), or rely on complex, non-deterministic pipelines (caption benchmarks). CaptionQA instead measures how "useful" a caption is by testing whether it can stand in for the image on dense, taxonomy-driven QA, and yields fine-grained diagnostics across domains and aspects.

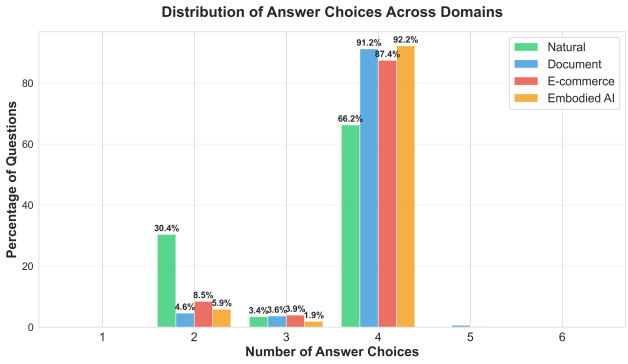

Figure 2. Distribution of answer choices across domains. Most questions are 4-choice (87–92%), making the benchmark more challenging than binary VQA. Natural domain has more binary questions due to attribute verification.

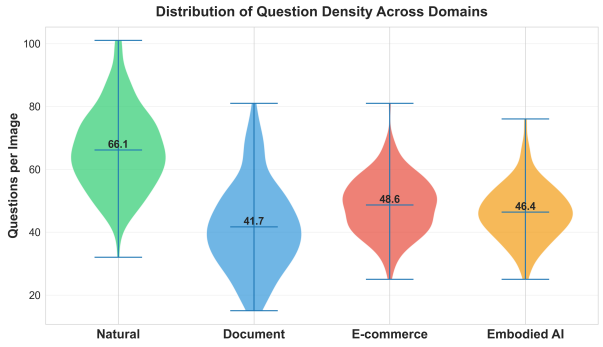

Figure 3. Distribution of question density across domains. The violin plots show the distribution of questions per image, with Natural images supporting the highest density of diverse questions and Document images focusing on specific structural and content elements. The consistent distributions within each domain (low variance) demonstrate systematic annotation quality.

4.1. Design Rationale

Although CaptionQA uses QA-style proxy, the QAs are derived from carefully designed taxonomy of each domain, reflecting the actual information needed in real downstream applications of each domain, such as retrieval, ranking, matching, classification, form filling, agentic pipelines. Each domain's taxonomy was designed through an iterative human-in-the-loop process. Domain experts from industry partners drafted initial categories based on downstream task requirements, which were then refined through several rounds of discussion and generative model refinement. The resulting taxonomies are domain-specific (emphasizing aspects relevant to each domain's applications), comprehensive (covering all salient visual information without redundancy), balanced (no single category dominates, with a maximum of 46.3% for Perception in Embodied AI), and practical (guiding question generation while remaining manageable for annotation). This taxonomy-driven approach ensures CaptionQA evaluates the *right* information–what downstream tasks actually need—rather than arbitrary caption properties.

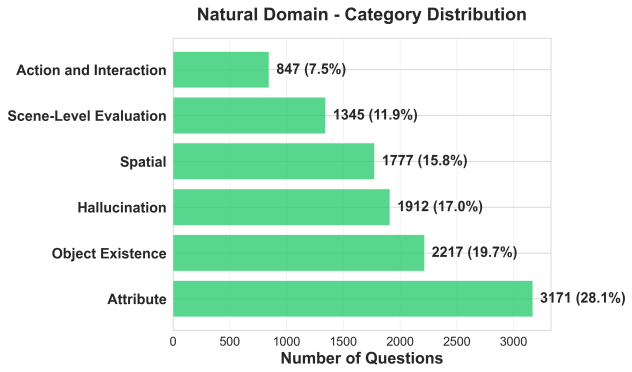

Figure 4. **Natural Domain:** Question distribution across top-level taxonomy categories. Attribute questions dominate, followed by Object Existence and Hallucination detection.

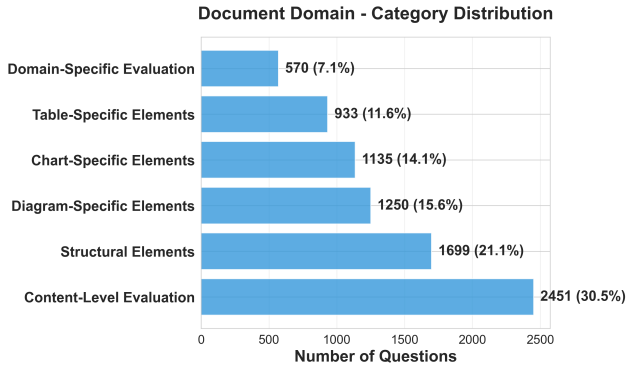

Figure 5. **Document Domain:** Question distribution across top-level taxonomy categories. Content-Level Evaluation and Structural Elements are the primary focus.

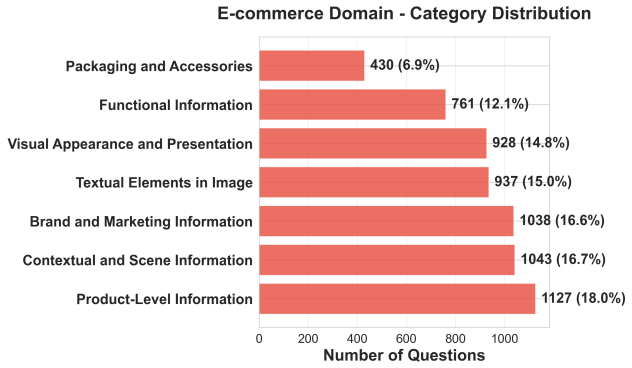

Figure 6. **E-commerce Domain:** Question distribution across top-level taxonomy categories. Questions are evenly distributed across product information, context, and marketing aspects.

5. Image Amount Justification

Instead of collecting tens of thousands of loosely annotated images as in most multimodal benchmarks, CaptionQA

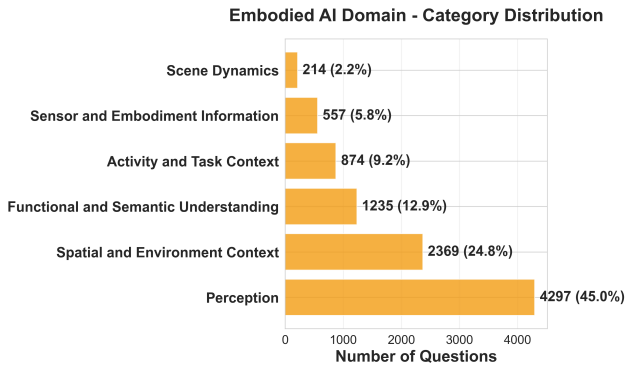

Figure 7. **Embodied AI Domain:** Question distribution across top-level taxonomy categories. Perception and Spatial Context dominate, reflecting robotics task requirements.

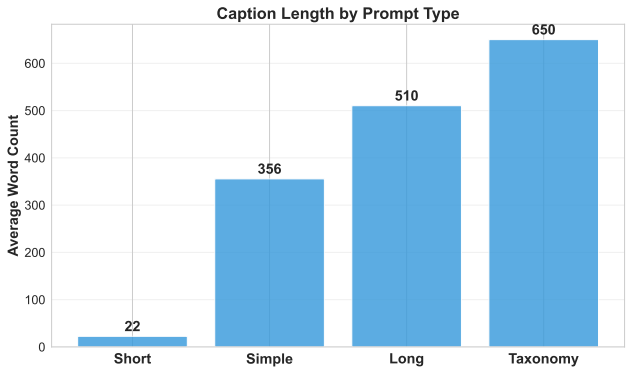

Figure 8. Average caption length (word count) by prompt type, averaged across all models and domains. The Taxonomy-Hinted prompt produces the longest captions (650 words on average), followed by Long (510 words), Simple (356 words), and Short (22 words).

adopts a *high-density* design: each image is paired with an average of 50 carefully curated, taxonomy-grounded questions. This dense annotation strategy makes each image substantially more informative than those in traditional VQA benchmarks, where a single image typically supports only 1–3 questions. Moreover, unlike benchmarks that evaluate short multiple-choice answers, caption evaluation requires generating full-sentence outputs for each image. *Increasing the number of images across domains would therefore linearly inflate the total evaluation time*, as caption generation latency grows with both model size and output length. Our design thus strikes a deliberate balance between semantic coverage and evaluation efficiency, enabling comprehensive yet tractable assessment of multimodal understanding.

To validate that our image amount design provides sufficient data for reliable model evaluation, we analyze ranking stability as a function of dataset size. For each domain, we

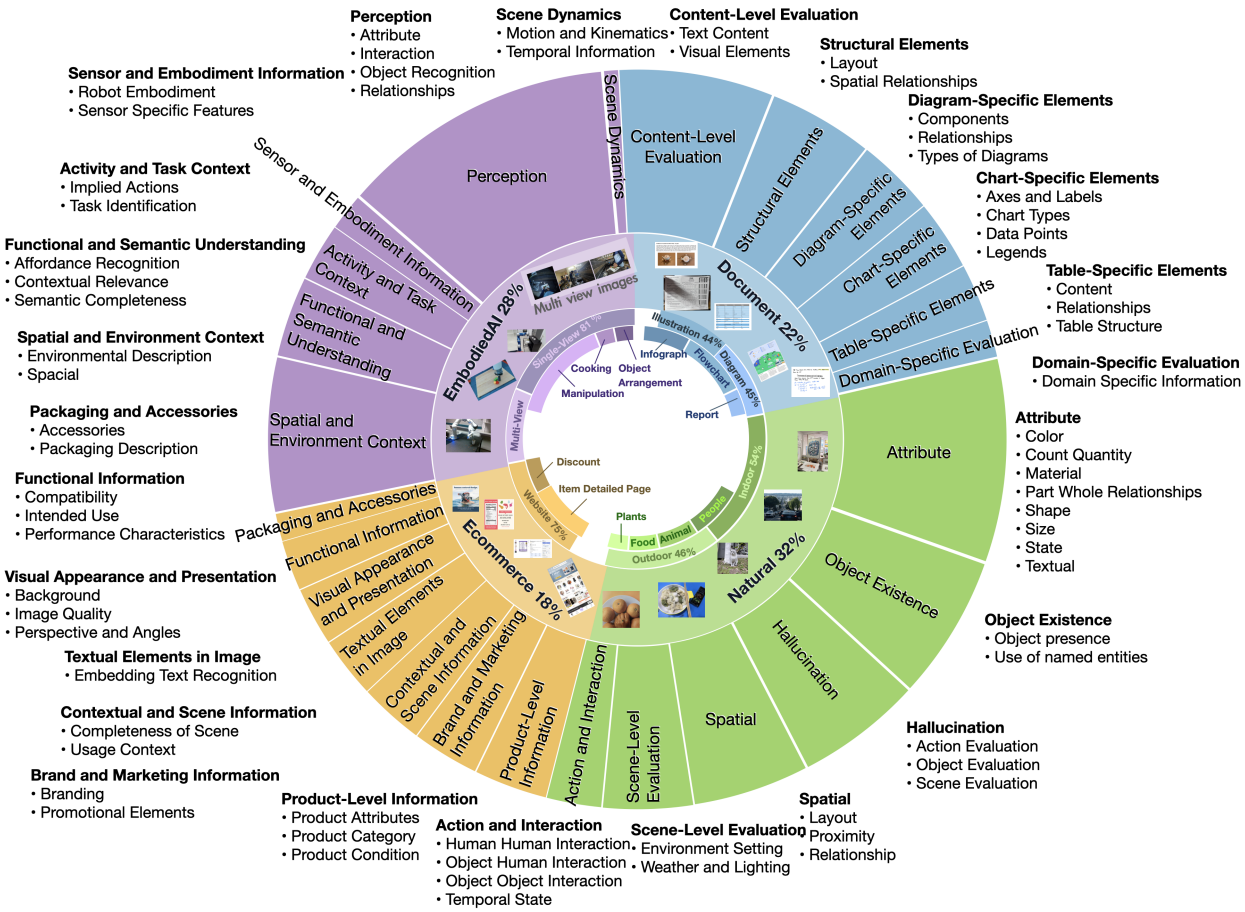

Figure 9. **Taxonomy structure across all four CaptionQA domains**. **(1) Natural domain** contains 6 top-level and 22 subcategories, emphasizing object properties, spatial relationships, and hallucination detection. **(2) Document domain** contains 6 top-level and 15 subcategories, focusing on structural elements, content evaluation, and document-specific features. **(3) E-commerce domain**contains 7 top-level and 16 subcategories, covering product attributes, visual presentation, and marketing information. **(4) Embodied AI domain** contains 6 top-level and 16 subcategories, prioritizing perception, spatial understanding, and task-relevant features for robotics applications.

randomly sample subsets of size $k \in [1, N]$ (where N is the total number of images), compute model performance on each subset, and repeat this 10 times per sample size to account for selection variance.

Figures 10 and 11 show performance trajectories for the top 10 models across all four domains. Three key findings emerge: (1) **Rapid stabilization**: Rankings stabilize within 20-40 images ($\sim$10-20% of full dataset) across all domains. (2) **Stable ordering**: After initial stabilization, model rankings remain consistent—performance curves maintain their relative positions without crossing. (3) **10$\times$ overcapacity**: Quantitative analysis shows rankings achieve Spearman correlation $\rho > 0.95$ with full rankings using only 10% of images, indicating CaptionQA contains approximately 10$\times$ more data than necessary for reliable evaluation.

This analysis empirically validates our high-density design: CaptionQA's dense annotation strategy provides reliable model rankings while maintaining evaluation efficiency. Users can obtain $\rho > 0.95$ correlated rankings using only 30-50 images per domain for preliminary evaluation, confirming that semantic coverage through dense questions is more effective than scale through numerous loosely-annotated images.

6. Cost of Extending CaptionQA to New Domains

One of our design goals is that CaptionQA should be easy to extend beyond the four domains used in the main paper (Natural, Document, E-commerce, Embodied AI). In this section we clarify what needs to be done to add a new domain and how the computational cost scales in our reference implementation.

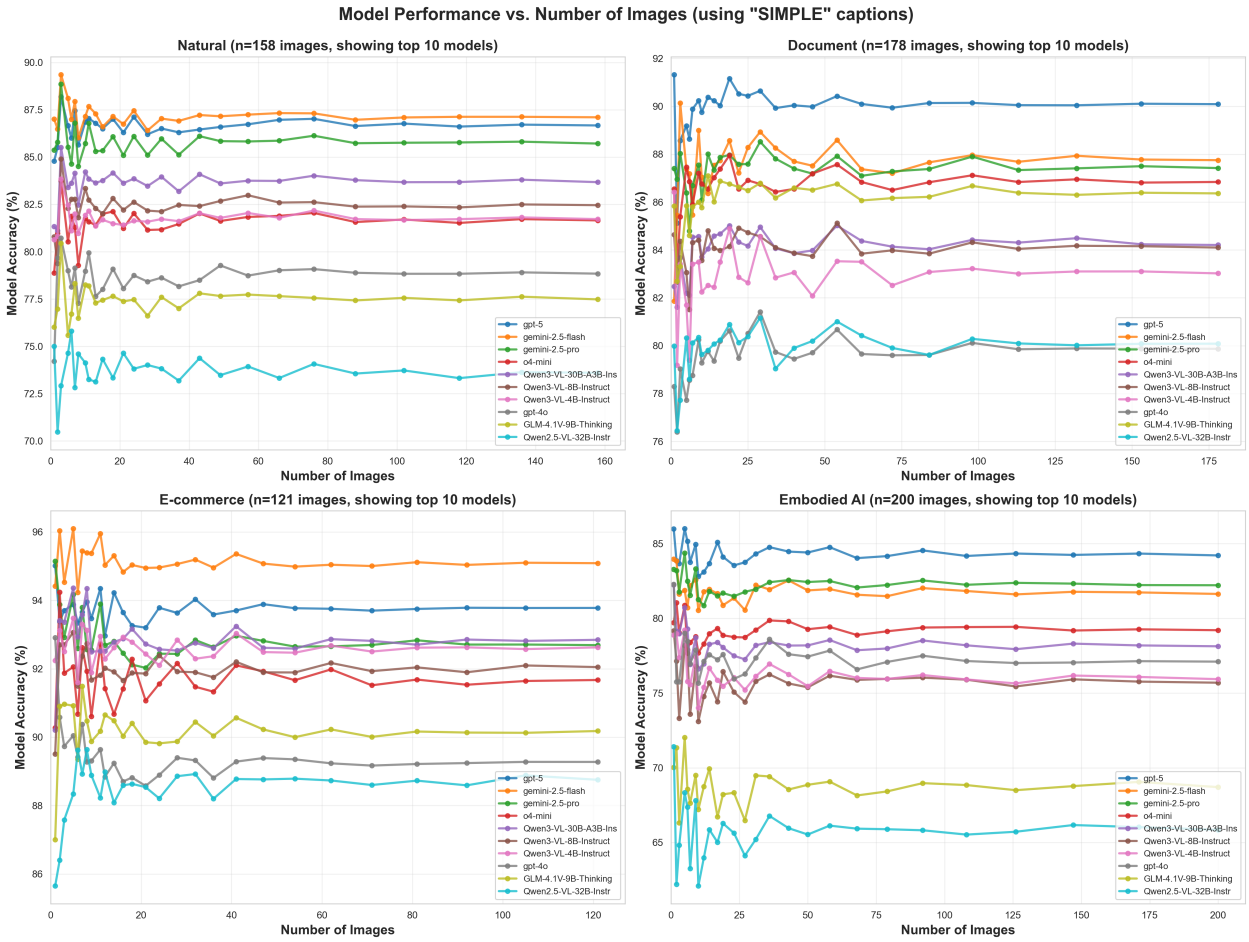

Figure 10. **Model ranking stability vs. number of images (accuracy-based).** Each line represents one model's accuracy trajectory as more images are randomly sampled (10 trials per sample size). Same color indicates the same model across domains. Performance curves plateau rapidly and maintain relative positions, validating data sufficiency. Top 10 models shown (ranked by average performance across domains).

Practical steps for adding a new domain. Given a new domain $\mathcal{D}_{\text{new}}$, extending CaptionQA is a purely mechanical procedure driven by our released code:

1. **Write a taxonomy.** Define a domain-specific taxonomy of information needs (analogous to Table 1-4 in the Supplementary), specifying which aspects (objects, layout, OCR, affordances, etc.) are important for downstream applications in $\mathcal{D}_{\text{new}}$. This is a one-time configuration file.
2. **Collect images.** Curate a set of (around 100-150 images to balance evaluation time and thoroughness) $N_{\text{img}}^{\text{new}}$ representative images for $\mathcal{D}_{\text{new}}$.
3. **Run the pipeline.** Invoke our end-to-end scripts, which automatically (i) generate taxonomy-grounded multiple-choice questions with three VLM agents (GPT 4o, o4-mini, GPT-5), (ii) apply the Qwen-based text-only filter and Qwen3 embedding deduplication, and (iii) perform dual-VLM visual verification before optional human spot-checking.

No manual question authoring is required: once the taxonomy and image list are prepared, all remaining steps are controlled by code scripts.

API cost. For a new domain at the same scale as our Natural split (around $N_{\text{img}}^{\text{new}} \approx 150$ images and $\sim$ 10,000 final QA pairs), the main API usage comes from (i) taxonomy-guided question generation with three VLM agents (GPT-4o, o4-mini, GPT-5), and (ii) dual-VLM quality control with GPT-5 and Gemini 2.5 Pro. In our current implementation this corresponds to roughly 17.7M total tokens across these APIs. Using current list prices for these models, this amounts to about **$40–$60** of one-time API cost to construct a new domain with $\sim$ 10k finalized questions. Since this cost scales linearly with the number of images and the de-

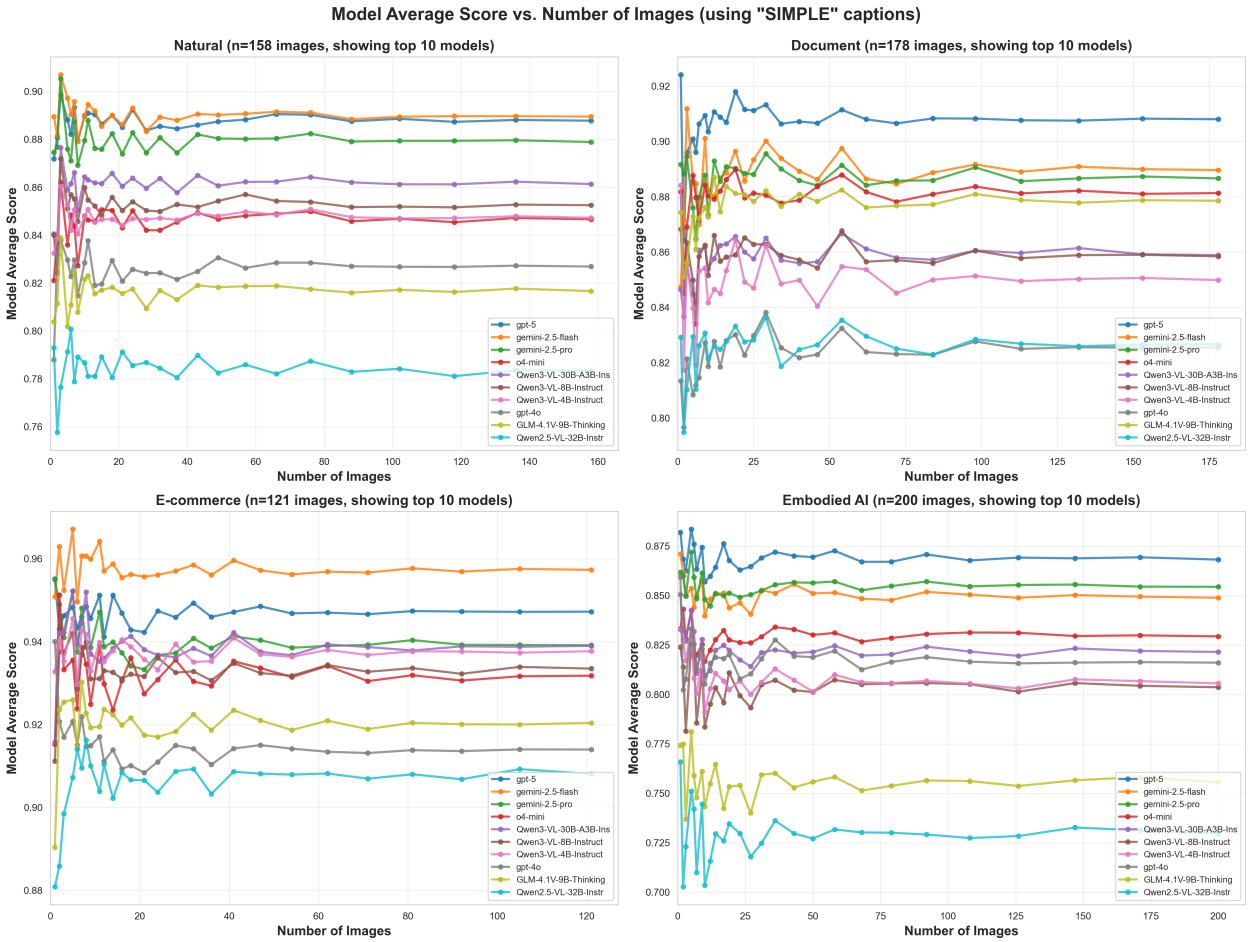

Figure 11. **Model ranking stability vs. number of images (average score-based).** Same analysis using average score (with partial credit for "Cannot answer": $1/n_{\text{choices}} + 0.05$). Patterns mirror Figure 10, confirming data sufficiency holds across evaluation metrics.

sired question density, smaller/larger domains incur proportionally smaller/larger API budgets.

GPU cost. All Qwen models in the pipeline (Qwen2.5-72B for text-only filtering and QA evaluation, and Qwen3-Embedding for deduplication) are run locally on a single AMD Instinct MI325. For a new domain with $\sim$ 10k final QA pairs, and assuming a modest candidate-to-final ratio (e.g., $\sim$ 3 candidates per retained question and 10 blind QA passes per candidate during text-only filtering), the total number of Qwen calls corresponds to roughly **3–4 MI325 GPU-hours** for construction. Once the domain is built, evaluating one captioning model on that domain (QA-on-caption scoring) takes less than **8 minutes** of MI325 time, with no additional API cost since Qwen runs locally. Thus, both construction-time and evaluation-time GPU costs are low and scale linearly with the number of questions.

Human cost. Human effort is deliberately kept minimal. Adding a new domain typically requires: (i) **2–4 hours** of expert time to design and refine the domain taxonomy (often by adapting and editing existing taxonomies), (ii) a short image curation pass to collect around 150 representative images, and (iii) **3–5 hours** of manual checking of questions that are flagged by the dual-VLM checker as potentially ambiguous, ungrounded, or too reasoning-heavy. In practice, this corresponds to roughly **one expert-day ($\sim$6–8 hours)** of human work per new domain, far less than what would be needed to author tens of thousands of multiple-choice questions from scratch.

Takeaway. The main conclusion is that extending CaptionQA to a new domain is entirely feasible: once a taxonomy and image set are specified, the rest of the process reduces to running our public code on a single GPU. The computational cost scales linearly with $N_{\text{img}}^{\text{new}}$ and N_{q} and is dominated by Qwen inference on one AMD MI325, plac-

Table 1. Natural-domain caption-utility taxonomy in **CaptionQA**..

Level-1 top-level category	**Level-2 subcategories (examples in parentheses)**
Object Existence	Object presence Use of named entities
Attribute	Color Shape Size Textual (written content on objects) Material State (e.g., open, closed, whole, broken) Count / Quantity (singular, plural, exact count, range estimate) Part–whole relationships (e.g., wheels of a car, branches of a tree)
Spatial	Positional relationships (above, below, beside) Orientation (facing forward, tilted) Containment (inside, outside) Attachment (connected, detached) Distance between objects (near, far) Clustering (grouped, scattered) Scene composition (foreground, background) Symmetry / asymmetry Overlapping / occlusion
Action and Interaction	Object–object interaction: contact (e.g., collision, gears rotation) Object–object interaction: functional (e.g., key opening the door) Object–human interaction: human activity (running, sitting) Object–human interaction: interaction context (playing sports, using tools) Human–human interaction: nonverbal communication (gestures, body language) Human–human interaction: collaborative / social dynamics (teamwork, group interaction) Human–human interaction: physical / affective interaction (hugs, handshakes) Temporal state: indication of motion (blurred objects, motion trails)
Scene-Level Evaluation	Environment / setting: environment type (natural, urban) Environment / setting: location type (indoor, outdoor, semi-indoor, semi-outdoor) Weather conditions (sunny, rainy, foggy) Time of day (daytime, night, dusk) Shadows and reflections Light source directionality
Hallucination	Object evaluation: object absence (likely but absent objects appropriately omitted) Object evaluation: object ambiguity (e.g., wolf vs. husky) Object evaluation: occluded objects misinterpretation Scene evaluation: scene misinterpretation (snow-covered vs. white sand beach) Scene evaluation: confounding elements (handles, reflections, shadows, clutter) Action evaluation: implied actions (e.g., "running" when the person is stationary)

ing extensions with a few hundred images per domain well within reach for typical academic and industrial users.

7. Rational and Reliability of LLM as QA Reader

Modern industrial systems increasingly rely on large language models not only as standalone chatbots, but as *components* inside downstream pipelines: LLM-based embedding models for retrieval and recommendation, LLM-driven re-ranking in search and feeds, and LLM agentic pipelines that orchestrate tools and multi-step plans. In many of these settings, the LLM consumes text surrogates of visual content (captions, alt-text, OCR transcripts) rather than raw pixels. In other words, a caption is often fed directly into an LLM that must then make a decision, retrieve items, or answer questions. From this perspective, using an LLM to *read* captions in CaptionQA is not a toy setup, but a practical abstraction of how captions are actually used in modern systems.

Table 2. Document-domain caption-utility taxonomy in **CaptionQA**.

Level-1 top-level category	**Level-2 subcategories (examples in parentheses)**
Structural Elements	Layout: key structural elements (title, headers, footnotes, page number) Layout: hierarchical structure (section, subsection) Layout: columns (single-column, multi-column layout) Spatial relationships: alignment (centered, left-aligned, right-aligned) Spatial relationships: overlapping elements (text over images, legends over charts) Spatial relationships: relative positioning (proximity of labels to corresponding elements)
Content-Level Evaluation	Text-content: textual information (extracted text matches the image) Text-content: completeness (missing text, partial text, missing values) Text-content: formatting (bold, italic, underline, font size) Text-content: style differentiation (captions vs. main body text) Text-content: accurate recognition of numbers (percentages, decimal points) Text-content: units and scales (e.g., 10 km vs. 10k) Text-content: symbols and special characters (currency symbols, math symbols, emoji) Visual elements: presence and identification of figures (charts, diagrams, icons) Visual elements: metadata of figures (page number of figures, continuation of figures)
Chart-Specific Elements	Chart types (bar chart, line chart, pie chart, scatter plot) Axes and labels: presence of x-axis and y-axis labels Axes and labels: axis scale (linear, logarithmic) Axes and labels: units of measurement (time, percentage) Legends Data points: correctness of data points Data points: completeness of data point descriptions Data points: trend identification (upward trend, downward trend)
Table-Specific Elements	Table structure: presence of headers (column headers, row headers) Table structure: merged cells (multi-row cells, multi-column cells) Table structure: gridlines (presence of borders, absence of borders) Content: completeness of table content (missing cells, missing rows/columns) Content: correctness of textual and numeric content in cells Content: formatting (bold headers, colored cells for emphasis) Content: units in cells (USD, kg) Relationships: cross-references (footnotes or notes referring to specific cells)
Diagram-Specific Elements	Types of diagrams (flowcharts, network diagrams, UML diagrams, Venn diagrams) Components: nodes (shapes, labels) Components: connections (nodes, type of connections, labels, symbols) Relationships: directionality of connections (one-way, bidirectional) Relationships: hierarchical structure (parent–child relationships in tree diagrams)
Domain-Specific Evaluation	Domain-specific information: financial reports (key metrics such as revenue, profit) Domain-specific information: scientific papers (recognition of equations, symbols) Domain-specific information: legal documents (extraction of clauses, dates)

Different from generic "LLM-as-a-judge" settings. Our use of an LLM QA reader is conceptually different from generic "LLM-as-a-judge" approaches that ask a model to rate a caption on heuristic criteria such as fluency, correctness, or level of detail. In CaptionQA, the LLM is not asked to produce a direct quality score; instead, it is placed in a concrete downstream task: answer multiple-choice questions about an image *using only the caption as input*. The LLM has no access to the image and must treat the caption as its sole evidence. This makes the LLM a stand-in for a real downstream consumer that must act based on textual surrogates, rather than a meta-critic with access to special instructions or reference answers. The quantity we measure is therefore not "how good the caption looks to the LLM", but *whether the caption actually supports successful task completion when an LLM tries to use it.*

Why QA, and why an LLM QA reader? We deliberately choose QA as the interaction between captions and the downstream LLM for two reasons. First, QA is a natural

Table 3. E-commerce-domain caption-utility taxonomy in **CaptionQA**.

Level-1 top-level category	**Level-2 subcategories (examples in parentheses)**
Product-Level Information	Product category (e.g., laptop, running shoes) Product attributes: color (red, matte black) Product attributes: dimensions (12 inches tall) Product attributes: size (set of 3 mugs, a dozen) Product attributes: material (leather, stainless steel) Product attributes: shape (rectangular, oval) Product attributes: texture (smooth finish, rough surface) Product attributes: weight (lightweight backpack) Product condition: new vs. used appearance (brand new, slightly worn); defects or wear (minor scratches)
Contextual and Scene Information	Usage context: indoor vs. outdoor setting (outdoor garden furniture) Usage context: lifestyle depiction (ideal for office use, perfect for outdoor camping) Usage context: human interaction (worn by a model, held by a hand) Usage context: environmental cues (in a kitchen, on a desk) Completeness of scene: completeness of product display (full product shown, partially shown) Completeness of scene: supporting objects (props, background elements) Completeness of scene: scene cleanliness (cluttered, minimalistic background)
Visual Appearance and Presentation	Image quality: blurriness Image quality: lighting (bright, natural light, shadows) Image quality: reflections (glare on metallic surfaces) Background: plain background (white, black) Background: styled background (lifestyle images with props) Background: transparency (images with transparent backgrounds) Perspective and angles: front / side / top / angled views Perspective and angles: close-up shots of key features (zoomed-in details) Perspective and angles: 360-degree or multi-angle views (implied by multiple images)
Functional Information	Intended use: functionality (multi-purpose tool, designed for running) Intended use: usability (easy to use, one-click operation) Intended use: special features (waterproof, wireless connectivity) Performance characteristics: capacity (16 GB RAM, 1 TB storage) Performance characteristics: durability (scratch-resistant, long-lasting) Performance characteristics: safety (child-safe material, non-toxic) Compatibility: compatibility with other products (compatible with iOS and Android) Compatibility: accessories included or sold separately (includes charging cable)
Brand and Marketing Information	Branding: visible logo or brand name Branding: trademark symbols (®, ™) Branding: brand-specific design elements (signature patterns) Branding: model / version identification (iPhone 14 Pro) Promotional elements: sale indicators (50% off tag) Promotional elements: certifications or labels (FDA-approved, eco-friendly) Promotional elements: awards and recognitions (Best product of the year)
Textual Elements in Image	Embedded text recognition: product name or description Embedded text recognition: price tags or discount labels Embedded text recognition: usage instructions or warnings (handle with care) Embedded text recognition: slogans or promotional text
Packaging and Accessories	Packaging description: box, bag, or wrapper (premium gift box) Packaging description: product labels on packaging (organic, recyclable) Packaging description: packaging design (minimalistic, vintage style) Accessories: listing of included accessories (comes with charger and earphones) Accessories: identification of main product vs. accessories (distinguishing primary product from props)

interface for many real applications: agents answering user questions about documents or products, assistants reasoning over screenshots or forms, and recommender systems extracting attributes from item descriptions. In these settings, the LLM rarely "rates" captions in the abstract; instead, it must use captions to answer concrete questions such as *"What color is the dress?", "Is the document signed?", "Which button should the robot press?"*. Our QA items are constructed exactly in this style. For example, in the E-commerce domain, we build a taxonomy around the information that real product recommendation and ranking systems need (category, style, material, pattern, fit, defects, packaging, etc.), and then generate multiple-choice questions that directly query these attributes. Similarly, the Document taxonomy targets fields and layout that drive automation pipelines, and the Embodied AI taxonomy focuses on object states and affordances relevant for agents. As a result, CaptionQA questions are not generic trivia, but explicit probes of domain-specific information needs.

Second, QA is a task where strong LLMs already exhibit high absolute performance and remarkable *stability*. In our experiments (Section 3.5), the QA reader attains high accuracy on image-grounded QA and shows low variance under option shuffling and repeated sampling. This stability is crucial: if the QA model itself were noisy, it would be unclear whether errors come from the caption or from the reader. Our analysis shows that, once the reader is strong enough, the dominant source of failures is the information available in the caption, not randomness in the LLM. In this sense, QA-on-caption with an LLM reader is a practical proxy for downstream usage: it asks whether a caption contains the specific, taxonomy-grounded facts that an LLM-based system (embedding model, recommender, or agent) will actually need in order to make correct decisions.

Our assumption: measure utility from the downstream LLM's perspective. CaptionQA is built around a simple assumption: what ultimately matters is what a *downstream LLM* can recover from the caption. If a caption causes the QA reader to answer incorrectly or become confused, this is evidence that the caption is missing or misleading with respect to that aspect of the image. If the reader still guesses the answer correctly by relying on world knowledge or priors, this also reflects something about caption utility: the caption may be underspecified (the LLM is filling in gaps) or the question may be solvable without the image. In both cases, we treat the combination of caption and QA reader as a black-box downstream system and measure whether it succeeds. This aligns more closely with real deployments, where practitioners care about *end-to-end decisions* made from captions, not about any particular intrinsic caption metric.

Independence from a specific QA model. Our framework is agnostic to the choice of QA reader: any sufficiently strong LLM can in principle be plugged into the CaptionQA evaluation protocol. In the main experiments we adopt Qwen2.5-72B as the default reader, but the benchmark does not depend on this particular model. Swapping the QA reader corresponds to changing the downstream consumer—e.g., using a different embedding model, recommender, or agent backend—while keeping the questions and captions fixed. This makes CaptionQA a flexible tool: users who rely on different LLM stacks in practice can rerun QA-on-caption with their own reader, and directly assess how well a caption model serves *their* downstream LLM-based system.

8. Prompt Transition Analysis: Where Does Length Help?

We analyze accuracy changes across four prompt transitions to identify which categories benefit from longer or more structured prompts.

8.1. Short to Simple: Identifying High-ROI vs. Low-ROI Categories

Figure 12 shows all 25 categories sorted by improvement. Document domain-specific evaluation (+47-51%) and E-commerce textual elements (+44-56%) gain the most. Embodied AI categories (+6-33%) and Natural hallucination (+15%) gain the least. Top categories gain up to 9× more than bottom categories, showing clear domain-specific patterns. The Short to Simple transition captures the majority of gains (mean +33.8%), while Simple to Long adds minimal value (mean +0.4%, shown next).

8.2. Simple to Long: Diminishing Marginal Returns

Figure 13 shows that Simple to Long transitions are nearly flat. All categories show $<2\%$ change, mean +0.35%. Most gains occur in the Short to Simple transition (+33.8% mean), while Simple to Long adds little value. Simple prompts achieve 99% of Long's performance at 70% of the length.

8.3. Long to Taxonomy-Hinted: When Structure Backfires

Figure 14 shows that Taxonomy-Hinted prompts (which list all 69 subcategories) hurt 23/25 categories vs Long, with mean loss of -10.8%. Largest losses occur in categories that already struggle: Document Domain-Specific Evaluation (-33.1%), Embodied AI Perception (-7.8%), Document Structural Elements (-9.3%). When models cannot extract information, explicit category lists may pressure them to fabricate details.

The 2 categories that benefit from Taxonomy-Hinted– Scene-Level Evaluation (Natural) and Visual Appearance

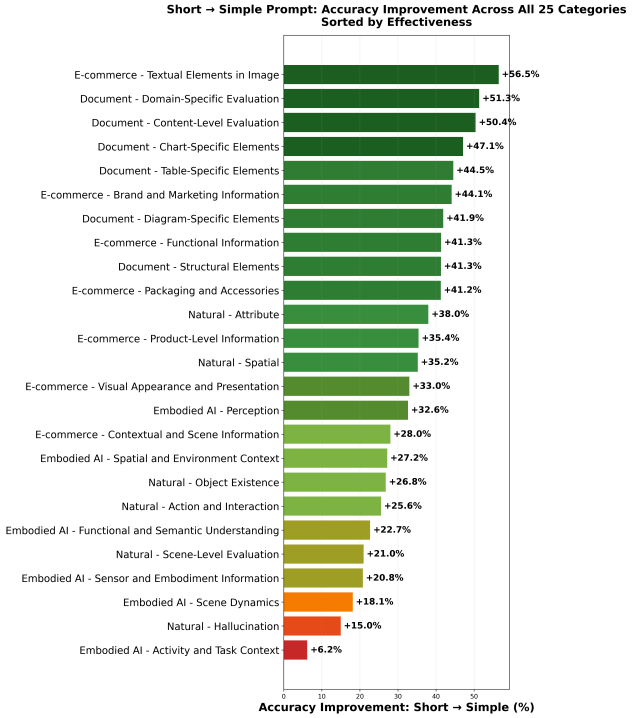

Figure 12. **Short to Simple effectiveness across all 25 categories.** Categories sorted by improvement, color-coded from dark green (+47-56%) to red (+6-21%). Top categories: Document domain-specific evaluation and E-commerce textual elements. Bottom categories: Embodied AI activity context and Natural hallucination. Top categories gain up to 9× more than bottom categories.

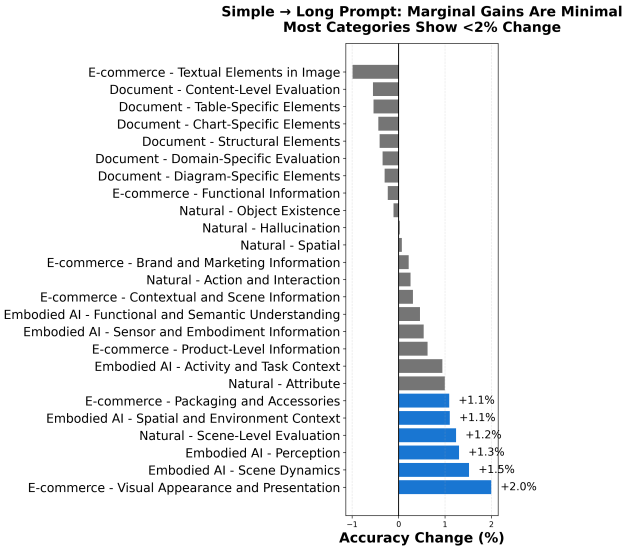

Figure 13. **Marginal gains: Simple to Long.** All 25 categories show <2% change (mean +0.35%). Simple and Long achieve nearly identical scores (75.5% vs 75.7%) despite 1.4× length difference (355 vs 510 words). Additional length provides minimal gain.

(E-commerce)–are both high-level judgments where structure may help organize outputs. Taxonomy-Hinted prompts may work for conceptual categories but not for fine-grained perceptual ones.

8.4. Coverage-Accuracy Relationship Across Prompt Transitions

We examine the relationship between *coverage* (Cannot-Answer rate reduction) and *accuracy* improvement across all four prompt transitions. Figures 15–18 reveal distinct patterns for each transition.

8.4.1. Short to Long: Strong Positive Correlation

Figure 15 shows strong correlation (r=0.905) between coverage and accuracy—longer captions generally improve both. **Below diagonal** (coverage > accuracy): Long captions answer many more questions, but newly answerable questions have lower accuracy (79-84%). Example: Natural Spatial gains 44% coverage but only 35% accuracy–adding substantial content but some is wrong. **Above diagonal** (accuracy > coverage): More efficient captioning that improves accuracy on already-answerable questions without adding much new content. Example: Natural Object Existence gains 27% accuracy but only 5% coverage–fixing errors rather than expanding scope.

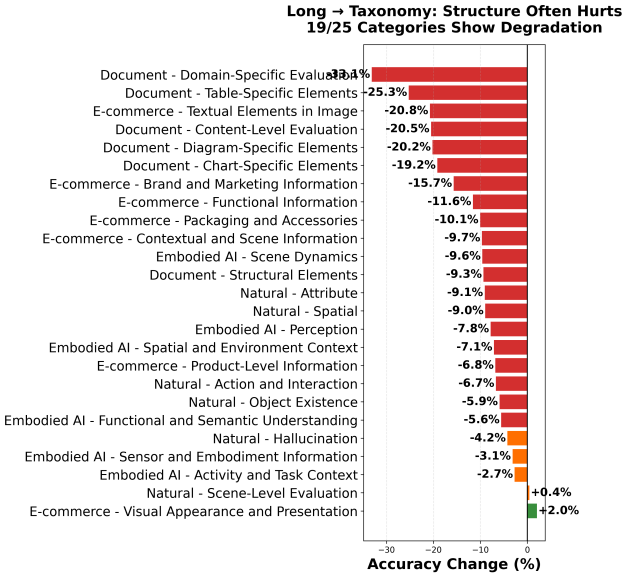

Figure 14. **Taxonomy-Hinted prompts often degrade performance.** 23 of 25 categories show losses (mean -10.8%), with 20 losing >5%. Only 2 categories gain (Visual Appearance +2.0%, Scene-Level Evaluation +0.4%). Largest losses: Document Domain-Specific Evaluation (-33.1%), Embodied AI Perception (-7.8%), Document Structural Elements (-9.3%).

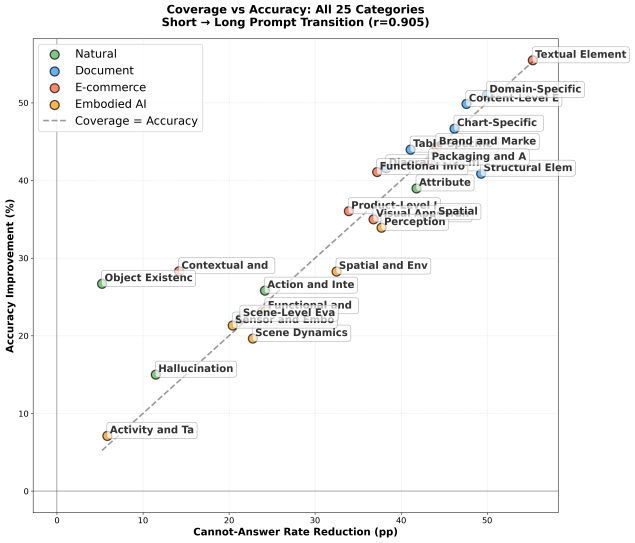

Figure 15. **Coverage vs. Accuracy: Short to Long (r=0.905).** Most categories cluster near diagonal. Below diagonal: Natural Spatial (43.9% coverage vs 35.3% accuracy), Document Structural (49.3% vs 40.9%)—more coverage than accuracy. Above diagonal: Natural Object Existence (5.3% coverage vs 26.7% accuracy), E-commerce Contextual (14.4% vs 28.3%)—more accuracy than coverage.

8.4.2. Short to Simple: Captures Most of Long's Benefits

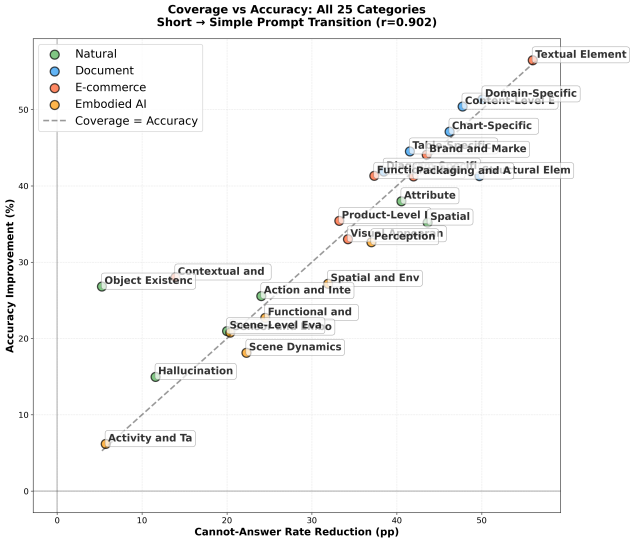

Figure 16. **Coverage vs. Accuracy: Short to Simple (r=0.902).** Mean coverage gain 32.8%, mean accuracy gain 33.8%. Comparing to Short to Long (33.1% coverage, 34.2% accuracy), this transition achieves 99% of Long's gains at 70% of the length.

Figure 16 shows r=0.902 with mean gains of 32.8% coverage and 33.8% accuracy. Comparing to Short to Long (r=0.905, 33.1% coverage, 34.2% accuracy), Simple achieves 99% of Long's benefits. This means the additional Simple to Long step adds minimal value (see next).

8.4.3. Simple to Long: Minimal Changes

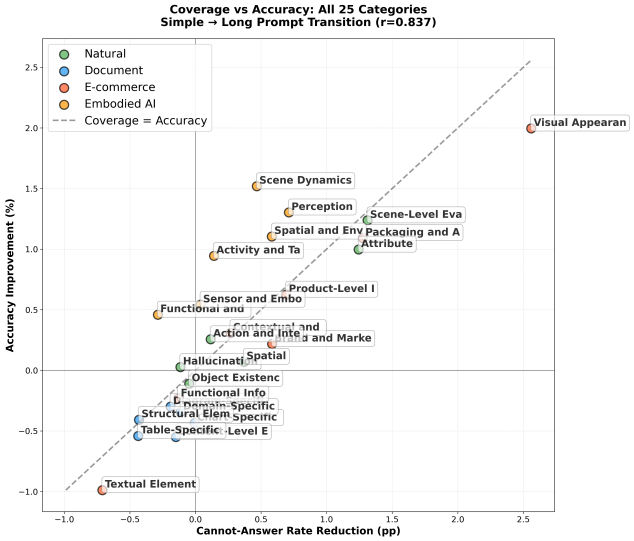

Figure 17. **Coverage vs. Accuracy: Simple to Long (r=0.837).** Clustered near origin. Mean coverage change +0.3%, mean accuracy change +0.4%. E-commerce Visual Appearance is outlier with +2.6% coverage and +2.0% accuracy. Near-zero changes confirm diminishing returns.

Figure 17 shows minimal changes. Most categories cluster near (0,0), confirming Simple prompts achieve nearly all benefits of Long prompts. Both coverage and accuracy changes are $<2\%$ for all categories, showing the Simple to Long step adds negligible value. E-commerce Visual Appearance is the only notable outlier.

8.4.4. Long to Taxonomy-Hinted: Negative Correlation

Figure 18 shows both coverage and accuracy decrease. Taxonomy-Hinted prompts harm both metrics for 23/25 categories. Strong correlation (r=0.966) shows coverage and accuracy decline together. Structured prompts that force ungroundable content cause captioning models to add fabricated details, which are then detected as unreliable by QA models (increasing Cannot-Answer) while also reducing accuracy on answerable questions.

8.5. Extreme Cases: Information Omission in Short Captions

Categories where Short captions yield Cannot-Answer rates exceeding 70%:

- **Perception (Embodied AI)**: 73.7% cannot-answer. Short captions omit fine-grained object attributes for manipulation tasks.

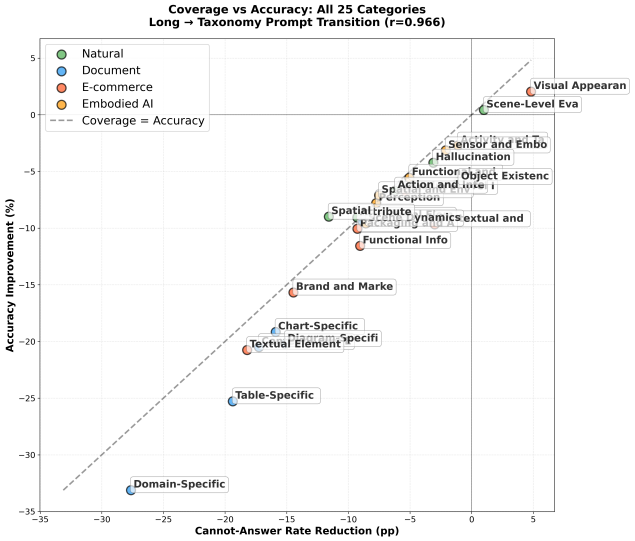

Figure 18. **Coverage vs. Accuracy: Long to Taxonomy-Hinted (r=0.966).** Strong negative correlation. Mean coverage change -8.8%, mean accuracy change -10.8%. Document Domain-Specific is worst outlier (-27.6% coverage, -33.1% accuracy). Only 2 categories show gains. Bottom-left quadrant: Taxonomy-Hinted prompts add wrong content that reduces both coverage and accuracy.

- **Spatial (Natural)**: 74.2% cannot-answer. Short captions exclude precise spatial relationships (distances, orientations).
- **Textual Elements (E-commerce)**: 72.1% cannot-answer. Short captions rarely transcribe product text (labels, specifications).

High Cannot-Answer rates indicate information is missing from the caption, not from the image. This is correct QA behavior—answering would require guessing.

9. Category-Level Statistical Summary

Table 5 shows statistics for all 25 top-level categories under the Simple prompt, aggregated across all models. For each category: mean score, standard deviation (across models), minimum and maximum scores (model variance), Cannot-Answer rate, and question count.

Several insights emerge from Table 5: (1) **Hallucination detection is robust** (85.3% mean, 10.0% Cannot), suggesting models effectively avoid generating false claims; (2) **Spatial reasoning shows high variance** (std=11.6 for Natural Spatial, 10.9 for Embodied Spatial), indicating inconsistent grounding across models; (3) **Cannot-Answer rates correlate with difficulty** (r=0.68), but not perfectly—Perception (Embodied AI) has both low score (65.3%) AND high Cannot rate (36.7%), indicating pervasive information absence.

9.1. Domain-Specific Subcategory Analysis

While Table 5 provides top-level statistics, Figures 19–22 break down performance across all 69 fine-grained subcategories, grouped by domain. Each radar chart shows all 24 evaluated models across all subcategories within that domain.

9.1.1. Natural Domain

Natural domain contains 22 subcategories spanning scene understanding, object recognition, attributes, actions, and spatial reasoning. Key findings: (1) **Scene-level categories are easiest**: Scene type classification (88%), environment recognition (85%), and overall scene description (82%) achieve high scores across all models. (2) **Spatial subcategories are hardest**: Distance estimation (45%), orientation (52%), and relative positioning (58%) show 20-30 point drops vs perceptual categories. (3) **Attribute granularity matters**: Coarse attributes (color, size: 75%) outperform fine-grained ones (texture, material: 55%).

9.1.2. Document Domain

Document domain contains 15 subcategories for charts, tables, diagrams, and text documents. Key findings: (1) **Content evaluation outperforms structure parsing**: Content-level evaluation (82%) and domain-specific reasoning (80%) are 15-20 points higher than structural parsing (60-65%). (2) **Chart elements are inconsistent**: Axis labels (62%), legends (58%), and data point extraction (55%) show high variance (std>12). (3) **Gemini advantage on tables**: Gemini 2.5 Pro leads by 5-8 points on table-specific subcategories (cell content, row/column understanding), suggesting better OCR integration.

9.1.3. E-commerce Domain

E-commerce domain contains 16 subcategories for product images. Key findings: (1) **Highest overall performance**: Mean across subcategories (81%) exceeds other domains by 8-12 points. Product images may have cleaner backgrounds and clearer focal objects. (2) **Text extraction type-dependent**: Brand names (85%) and product titles (82%) are captured well, but fine print (68%) and specifications (65%) are often missed. (3) **Visual appearance is relative**: Color description (75%) works but style matching (62%) and material appearance (58%) are subjective and harder to ground.

9.1.4. Embodied AI Domain

Embodied AI domain contains 16 subcategories for robotics and embodied perception. Key findings: (1) **Task context vs perception gap**: Activity recognition (83%) and task-level reasoning (78%) are 15-20 points higher than object properties (60%), affordances (55%), and manipulation planning (52%). High-level semantics are easier than

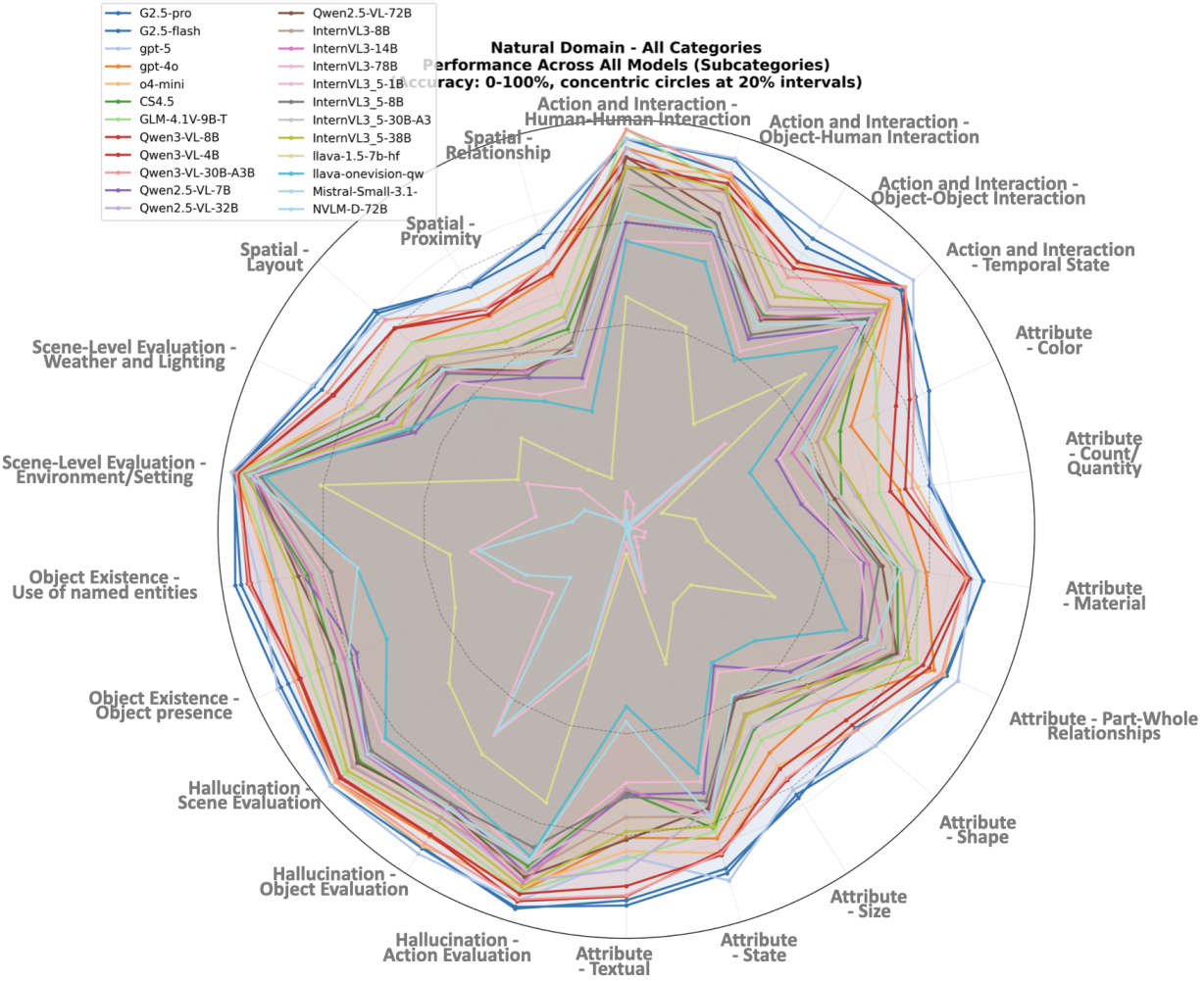

Figure 19. **Natural domain: 22 subcategories.** Models perform best on scene-level evaluation (80-92%) and object existence (75-90%), but struggle with spatial reasoning (40-65%) and fine-grained attributes (50-70%). GPT-5 leads on most categories. Spatial subcategories (distance, orientation, relative position) show the largest performance gaps and highest variance.

action-relevant details. (2) **Sensor information is omitted**: Depth cues (48%), camera viewpoint (52%), and embodiment context (55%) have >30% Cannot-Answer rates, indicating captions rarely describe sensor geometry. (3) **Dynamics are hard**: Temporal dynamics (58%), motion prediction (53%), and scene changes (50%) show captioning models struggle with implicit temporal reasoning from static images.

10. Detailed Model Performance Analysis

Figure 23 presents a unified view of all 24 evaluated models across all 69 subcategories and 4 domains. Each axis represents one subcategory, with axes colored by domain (Natural=green, Document=blue, E-commerce=red, Embodied AI=orange) and separated by black radial lines marking domain boundaries.

Three findings emerge from the comprehensive view. First, **model strengths are domain-specific**: GPT-5 leads on Natural and E-commerce categories, while Gemini models perform better on Document structural elements. Second, **spatial reasoning is hard for all models**: scores drop to 35-60% on Embodied AI spatial categories vs 75-90% on perceptual categories. Third, **category matters more than model**: within-model variance across categories (35-95%) exceeds between-model variance on the same category (5-15 points), showing that *what* to caption is harder than *how well* to caption it.

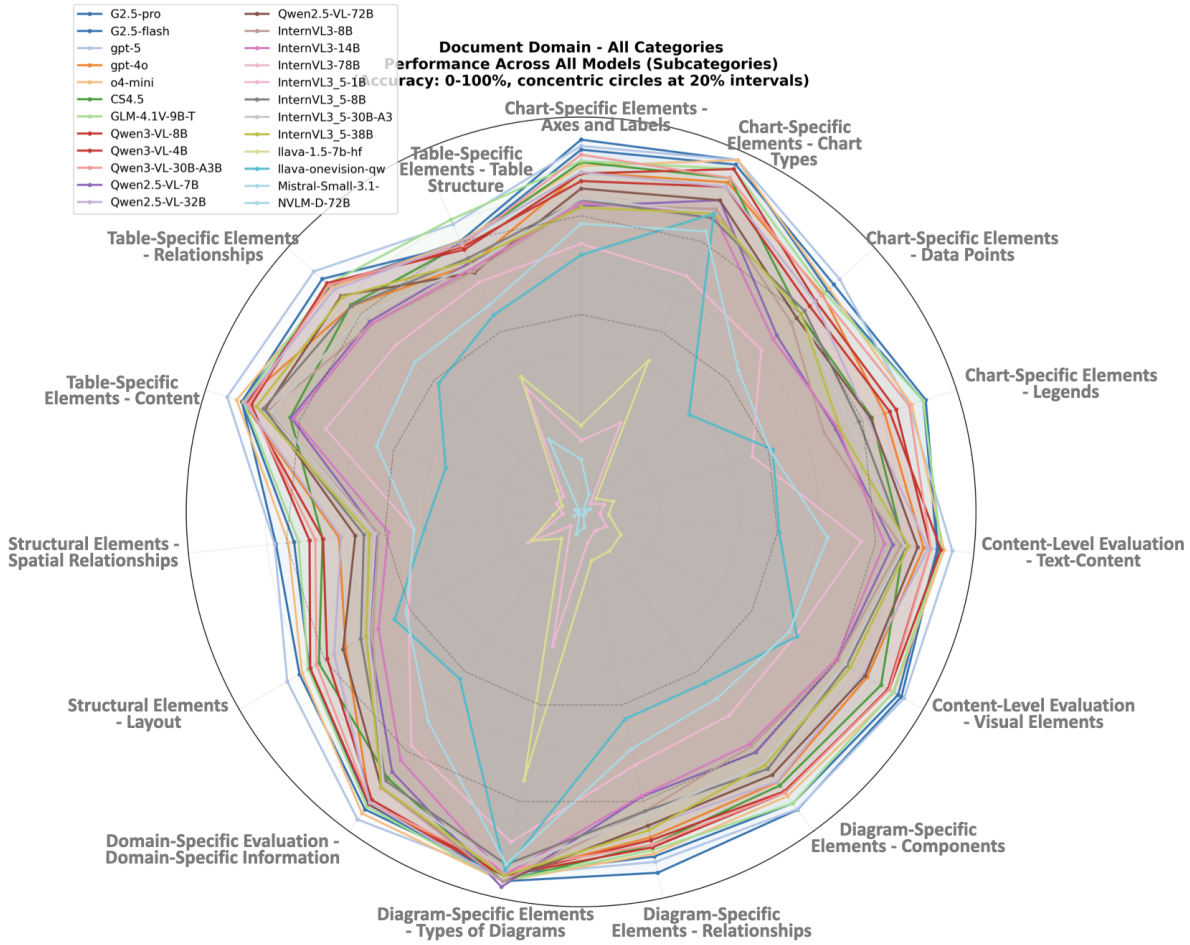

Figure 20. **Document domain: 15 subcategories.** Models excel on high-level evaluation (80-93%) but struggle with structural elements (50-75%). Gemini models show relative strength on table/chart parsing. Variance is high across subcategories, with chart-specific elements (axis labels, legends) being particularly challenging.

11. Question Difficulty Distribution

To assess whether CaptionQA provides adequate discrimination across different capability levels, we analyze question difficulty based on the percentage of models that answer each question correctly.

11.1. Difficulty Categorization

We categorize questions into three difficulty levels based on the proportion of models (out of 24 total) that answer correctly:

- **Easy:** $\geq$80% of models answer correctly
- **Medium:** 50-80% of models answer correctly
- **Hard:** $<$50% of models answer correctly

11.2. Distribution Across Domains

Figure 24 shows the difficulty distribution for each domain.

11.3. Examples of Hardest and Easiest Questions

Figure 25 shows concrete examples of the hardest and easiest questions with their corresponding images, answer choices, and correct answers across all four domains.

12. Full Results

We report the full CaptionQA results as shown in Table 7–Table 22 for all evaluated models, prompts, and domains in this section. These tables complement the main-paper summary (Table 3) by providing per-domain, per-prompt breakdowns, and by including all three metrics: Score, Acc, and Cannot. *Models in the tables are ranked by score*.

Models covered. In total we evaluate **24** multimodal LLMs, spanning both open-source and proprietary systems and covering a wide range of scales (from 1B to 78B param-

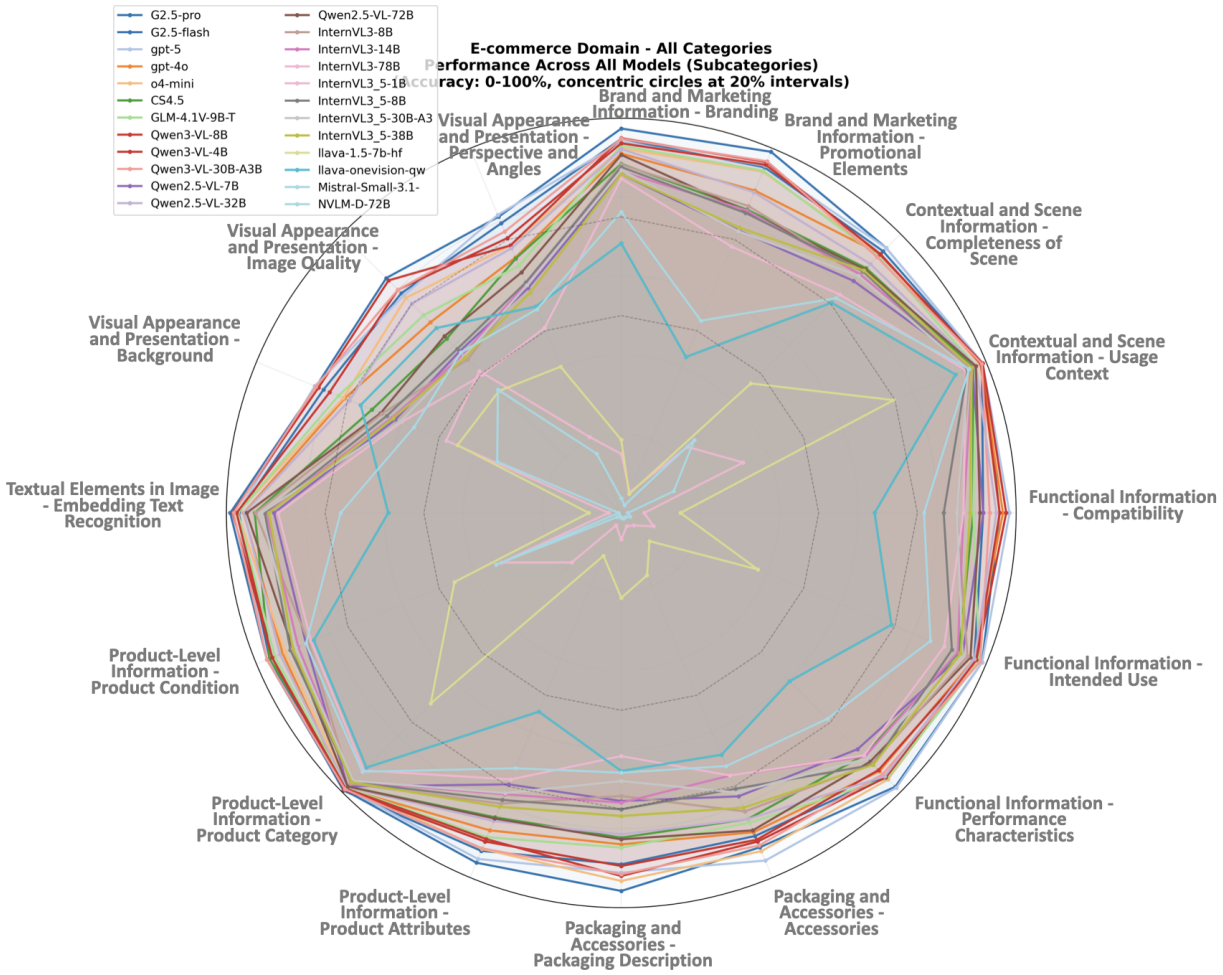

Figure 21. **E-commerce domain: 16 subcategories.** Models achieve highest overall scores (70-96%) across all domains. Contextual understanding (85-96%) and product-level information (82-94%) are strengths. Visual appearance details (color matching, style) are harder (60-75%). Text extraction varies by text type.

eters for open-source models, and large API-only models on the proprietary side).

Open-source VLMs. Our open-source pool includes 18 models from five major families:

- **Qwen3-VL**: 4B, 8B, and 30B-A3B.
- **Qwen2.5-VL**: 7B, 32B, and 72B.
- **GLM-4.1V**: 9B.
- **InternVL family**: InternVL3.5 (1B, 8B, 30B-A3B, 38B) and InternVL3 (8B, 14B, 78B).
- **Other baselines**: NVLM-D-72B, LLaVA-OneVision-7B, LLaVA-1.5-7B, and Mistral-Small-24B.

Proprietary VLMs. We further evaluate 6 proprietary models:

- **OpenAI**: GPT-5, GPT-4o, and GPT-o4-mini.
- **Google**: Gemini 2.5 Pro and Gemini 2.5 Flash.
- **Anthropic**: Claude Sonnet 4.5.

These models represent the current generation of large, API-only VLMs commonly deployed in production systems.

Prompts. Every model is evaluated under the same set of **four caption prompts** described in Section 3.3 of the main paper:

- **Long**: "Write a very long and detailed caption describing the given image as comprehensively as possible."
- **Short**: "Write a very short caption for the given image."
- **Simple**: "Describe this image in detail." (our recommended default setting).
- **Taxonomy-Hinted**: We supply the domain taxonomy as a list of aspect prompts (Top-level → Subcategory) and ask the model to describe the image from those perspectives.

All four prompts are applied to all four domains for each captioning model, so every model is evaluated in 16 settings

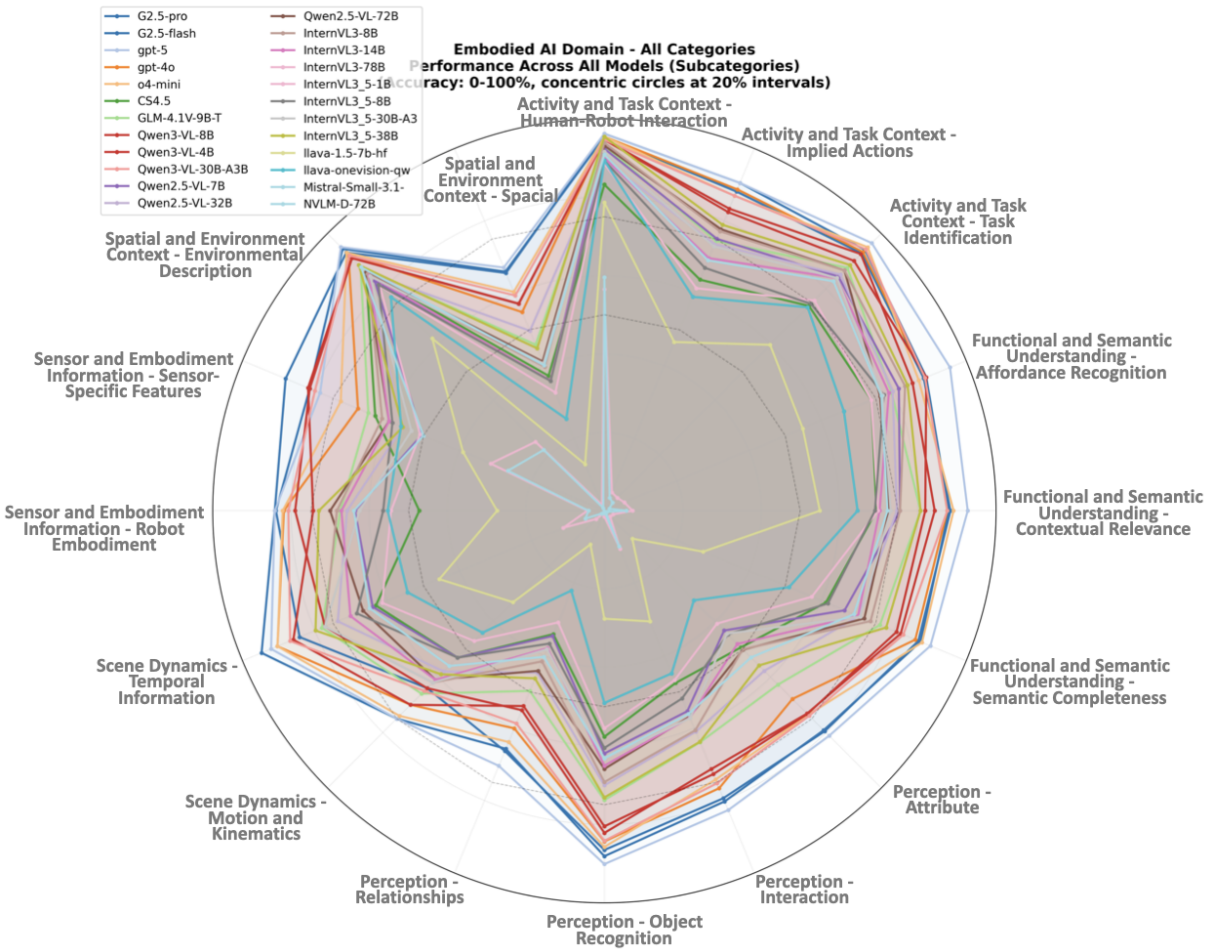

Figure 22. **Embodied AI domain: 16 subcategories.** Most challenging domain overall (50-85% range). Activity/task understanding (80-93%) is relatively strong, but perception subcategories (object properties, affordances, manipulation) drop to 40-70%. Sensor-specific information (depth, embodiment viewpoint) is systematically under-described.

(4 domains $\times$ 4 prompts).

Domains and metrics. The tables are organized by domain (**Natural**, **Document**, **E-commerce**, and **Embodied AI**). Within each domain we report one table per prompt (Long, Short, Simple, Taxonomy-Hinted). Each table contains three metrics:

- **Score** (%): our utility-oriented multiple-choice score averaged over all questions.
- **Acc** (%): CaptionQA accuracy, i.e., fraction of questions where the QA reader selects the correct option.
- **Cannot** (%): fraction of questions where the QA reader selects "Cannot answer from the caption.", indicating that the caption is judged insufficient for that question.

Rows are grouped into *Open-Source VLMs* and *Proprietary VLMs*, and sorted by Score.

Table layout. For clarity, we mirror the same layout across domains. For example, Table 7 reports all models on the **Natural** domain under the **Long** prompt; analogous tables report the other prompts on the Natural domain, and similarly for **Document**, **E-commerce**, and **Embodied AI**. Together, these tables provide a complete view of CaptionQA performance across *all* model–prompt–domain combinations evaluated in this work.

References

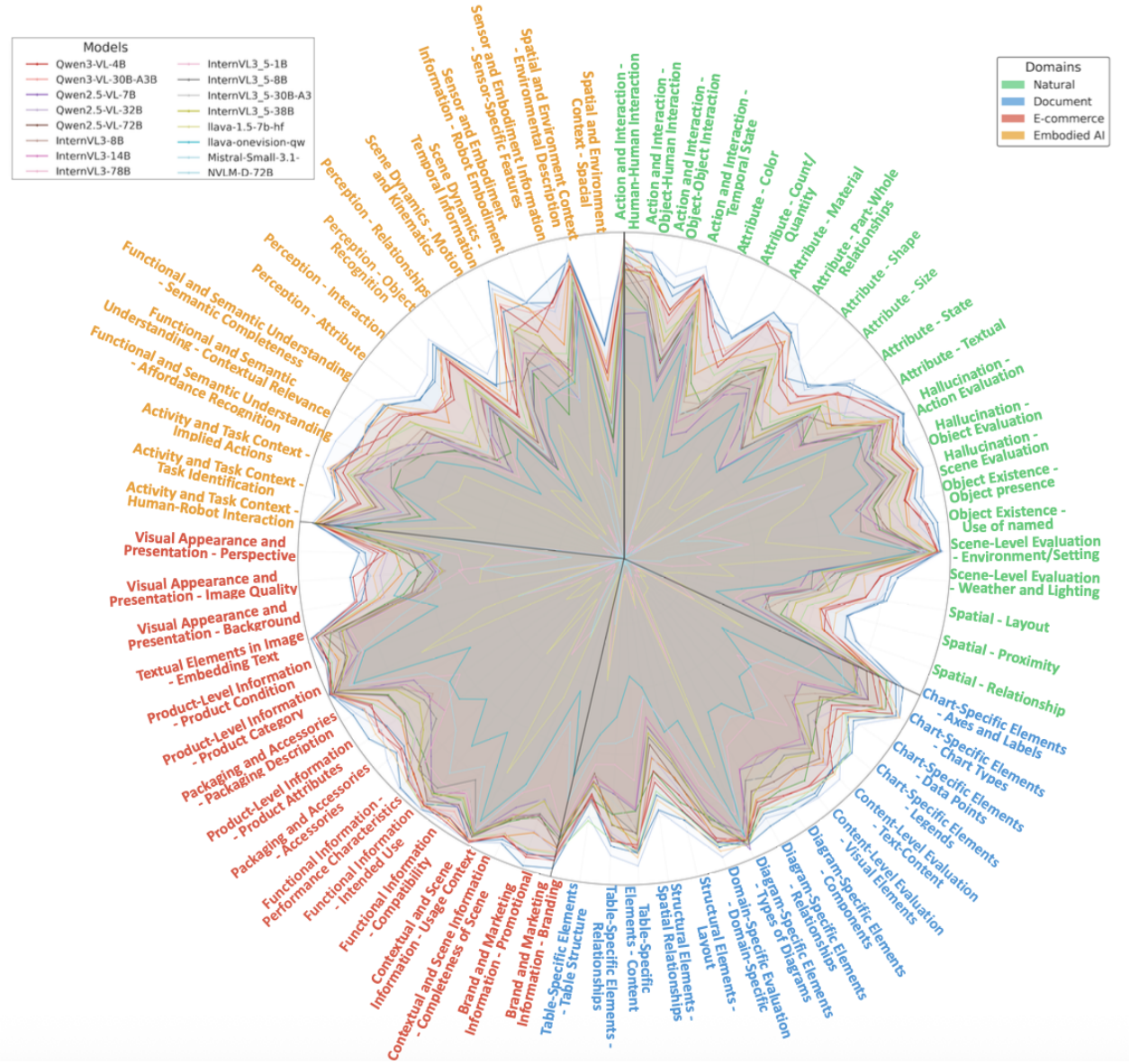

Figure 23. **Comprehensive model performance across all 69 subcategories and 4 domains.** Each colored line represents one of 24 evaluated models across 69 fine-grained subcategories. The chart is divided into four domain sections (Natural, Document, E-commerce, Embodied AI) separated by black radial lines. Concentric circles indicate accuracy levels from 0-100% at 20% intervals. Top-performing models include **GPT-5** and **Gemini 2.5 Flash** among proprietary models, and **Qwen3-VL 30B-A3B** and **GLM-4.1V 9B** among open-source models, though performance varies dramatically across categories within each model (30-95% range), revealing that no single model excels uniformly across all task types. Notably, categories requiring spatial reasoning (Embodied AI section) show consistently depressed performance across all models (40-70%) compared to perceptual categories (70-90%), indicating a systematic capability gap in current vision-language models.

Table 4. Embodied-AI-domain caption-utility taxonomy in **CaptionQA**.

Level-1 top-level category	**Level-2 subcategories (examples in parentheses)**
Perception	Object recognition: object identification (e.g., cup, door handle) Object recognition: object category (e.g., furniture, tools, appliances) Attribute: color Attribute: shape Attribute: size Attribute: material Attribute: state Attribute: orientation Relationships: positional relationships (on top of, next to) Relationships: containment (object inside a box) Relationships: attachment (tool attached to a robotic arm) Relationships: occlusion (partially visible object behind another object) Interaction: contact (robot holding a bottle) Interaction: manipulation (grasping, pushing, pulling) Interaction: proximity (object within reach, object far from reach)
Spatial and Environment Context	Spatial: proximity (near the edge of the table) Spatial: distance estimation (approximately 2 meters away) Spatial: perspective (view from a high angle, low-angle view) Environmental description: indoor vs. outdoor (indoor kitchen, outdoor garden) Environmental description: room type (living room, office, workshop) Environmental description: surroundings (surrounded by shelves, in an open space) Environmental description: surface properties (wooden floor, metal surface)
Activity and Task Context	Task identification: navigation tasks (robot navigating a hallway) Task identification: object manipulation tasks (picking up a tool) Task identification: cleaning tasks (sweeping debris) Task identification: inspection tasks (inspecting a pipe for damage) Implied actions: action in progress (robot approaching a table) Implied actions: action completed (door successfully opened) Implied actions: task outcome (object successfully placed in the bin) Human–robot interaction: human presence (person standing nearby) Human–robot interaction: interaction type (handing an object to the robot) Human–robot interaction: collaborative actions (robot assisting a person with a task)
Scene Dynamics	Motion and kinematics: robot motion (robot moving forward, robot arm rotating) Motion and kinematics: object motion (ball rolling on the floor) Motion and kinematics: velocity estimation (object moving quickly, slow movement) Temporal information: time-specific context (morning light coming through the window) Temporal information: sequential actions (after opening the drawer, picking up the tool)
Sensor and Embodiment Information	Sensor-specific features: camera type (RGB, depth, thermal) Sensor-specific features: depth perception (distance to object measured by depth sensor) Sensor-specific features: field of view (wide-angle view, narrow focus on object) Sensor-specific features: sensor artifacts (glare on metallic surface, low-light noise) Robot embodiment: robot components in frame (robot arm, gripper) Robot embodiment: self-awareness (robot's shadow visible, robot base in view) Robot embodiment: tool attachment (screwdriver attached to gripper)
Functional and Semantic Understanding	Affordance recognition: affordances of objects (graspable handle, pourable bottle) Affordance recognition: tool usability (wrench ready to tighten a bolt) Affordance recognition: interaction potential (button pressable by robot finger) Semantic completeness: completeness of scene description (all key objects and actions described) Semantic completeness: avoidance of hallucination (no mention of non-existent objects or actions) Contextual relevance: task-specific relevance (focusing on objects necessary for the task) Contextual relevance: importance weighting (emphasizing key objects over background elements)

Table 5. Per-category performance summary across all models (Simple prompt). Categories are sorted by mean score within each domain. High std indicates large inter-model variance; high Cannot rate indicates systematic omissions.

Domain	Category	Mean	Std	Min	Max	Cannot %	#Q
Natural	Scene-Level Evaluation	82.3	8.1	65.2	91.4	14.7	1247
	Object Existence	77.7	9.3	58.1	89.2	4.3	2215
	Action and Interaction	76.9	8.7	61.3	88.5	22.4	1682
	Attribute	69.0	10.2	48.9	84.7	32.5	3178
	Hallucination	85.3	7.4	71.2	95.1	10.0	1913
	Spatial	62.8	11.6	42.1	78.9	30.6	1210
Document	Domain-Specific Evaluation	80.8	12.3	58.9	93.2	18.9	982
	Content-Level Evaluation	79.0	11.8	59.3	91.7	20.6	2448
	Diagram-Specific Elements	77.6	10.9	58.2	90.1	17.8	467
	Chart-Specific Elements	75.5	10.5	57.9	88.3	21.0	891
	Table-Specific Elements	74.6	10.2	56.8	87.4	17.3	1124
	Structural Elements	64.7	9.8	47.3	79.2	23.7	1510
E-commerce	Contextual and Scene Info	86.5	8.9	71.2	95.8	6.1	1041
	Product-Level Information	85.4	9.2	69.8	94.3	17.4	1125
	Textual Elements in Image	83.5	11.7	61.9	95.2	17.7	687
	Brand and Marketing Info	83.4	10.1	66.7	93.5	20.8	1037
	Functional Information	86.3	9.4	70.1	95.9	16.0	589
	Packaging and Accessories	78.8	10.6	61.2	91.3	23.9	682
	Visual Appearance	71.8	9.8	55.4	85.7	22.6	725
Embodied AI	Activity and Task Context	82.5	9.7	66.8	93.2	18.1	1689
	Perception	65.3	11.4	45.9	81.7	36.7	4294
	Sensor and Embodiment Info	70.4	10.8	52.6	85.1	27.8	741
	Functional/Semantic Underst.	78.4	10.2	61.7	90.3	23.9	1247
	Scene Dynamics	70.1	11.6	51.3	86.4	24.1	478
	Spatial and Environment	71.8	10.9	54.2	87.6	25.9	825

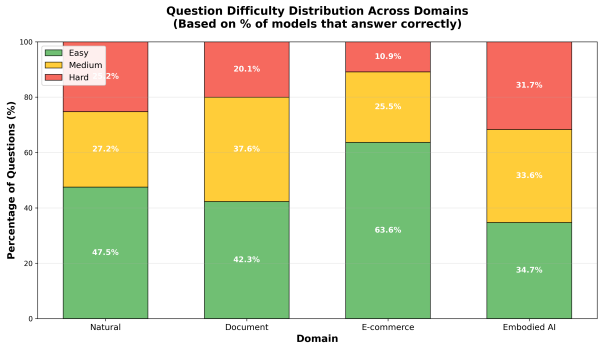

Figure 24. **Question Difficulty Distribution Across Domains.** Each domain exhibits a different difficulty profile: E-commerce has more easy questions (64%), while Embodied AI is the most challenging (32% hard questions). This diversity ensures CaptionQA can discriminate between models at different capability levels.

Table 6. **Question Difficulty Distribution Summary.**

Domain	Easy	Medium	Hard	Total
Natural	47.5%	27.2%	25.2%	10,445
Document	42.3%	37.6%	20.1%	7,417
E-commerce	63.6%	25.5%	10.9%	5,884
Embodied AI	34.7%	33.6%	31.7%	9,273

Table 7. Results on the **Natural** domain with the **Long** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	8B	85.69	83.11	11.8
Qwen3-VL	30B-A3B	85.62	83.18	11.2
Qwen3-VL	4B	83.95	80.87	14.2
GLM-4.1V	9B	82.92	79.07	17.6
Qwen2.5-VL	32B	78.95	74.58	20.0
InternVL3.5	38B	78.70	73.94	21.9
Qwen2.5-VL	72B	77.29	72.10	23.9
InternVL3.5	30B-A3B	74.33	68.06	29.0
Qwen2.5-VL	7B	73.84	67.61	28.7
InternVL3.5	8B	73.73	67.24	29.9
NVLM-D	72B	72.66	66.20	29.8
InternVL3	14B	72.41	65.68	31.3
InternVL3.5	1B	72.22	65.65	30.5
LLaVA-OneVision	7B	70.20	62.94	33.6
InternVL3	8B	66.39	57.10	42.9
LLaVA-1.5	7B	53.06	39.14	64.7
InternVL3	78B	38.66	19.44	89.6
Mistral Small	24B	35.81	15.34	95.6
Proprietary VLMs				
GPT-5	–	90.34	88.89	6.7
Gemini 2.5 Pro	–	89.44	87.71	7.9
Gemini 2.5 Flash	–	89.28	87.57	7.8
GPT-o4-mini	–	85.99	83.62	11.0
GPT-4o	–	84.71	81.60	14.3
Claude Sonnet 4.5	–	77.78	72.96	22.2

Table 8. Results on the **Natural** domain with the **Simple** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	30B-A3B	86.14	83.68	11.3
Qwen3-VL	8B	85.25	82.46	12.7
Qwen3-VL	4B	84.73	81.72	13.7
GLM-4.1V	9B	81.67	77.48	19.2
Qwen2.5-VL	32B	78.35	73.60	21.8
InternVL3.5	38B	78.26	72.95	24.4
Qwen2.5-VL	72B	75.26	69.34	27.4
InternVL3.5	30B-A3B	74.58	68.43	28.5
InternVL3-14B	14B	74.16	67.71	29.8
NVLM-D	72B	73.13	66.74	29.6
InternVL3.5	8B	72.97	66.43	30.2
Qwen2.5-VL	7B	71.64	64.58	32.6
InternVL3.5	1B	70.82	63.55	33.6
LLaVA-OneVision	7B	66.56	57.91	40.1
LLaVA-1.5	7B	52.51	38.53	65.0
InternVL3	78B	38.86	19.72	89.3
Mistral Small	24B	35.91	15.26	96.5
Proprietary VLMs				
Gemini 2.5 Flash	–	88.95	87.10	8.4
GPT-5	–	88.78	86.67	9.7
Gemini 2.5 Pro	–	87.89	85.72	10.0
GPT-4o	–	82.69	78.84	17.7
GPT-o4-mini	–	84.66	81.66	13.8
Claude Sonnet 4.5	–	76.56	71.12	25.0

Table 9. Results on the **Natural** domain with the **Short** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	62.43	50.58	54.8
Qwen3-VL	4B	60.52	48.63	55.0
Qwen3-VL	30B-A3B	59.48	47.23	56.7
Qwen3-VL	8B	57.38	44.00	62.0
InternVL3.5	38B	56.52	43.21	61.7
InternVL3	8B	56.50	43.13	61.9
InternVL3.5	1B	55.28	41.55	63.7
InternVL3.5	30B-A3B	55.16	41.28	64.4
InternVL3-14B	14B	55.13	41.12	64.9
InternVL3.5	8B	54.89	40.94	64.7
Qwen2.5-VL	7B	54.77	40.73	65.1
Qwen2.5-VL	32B	54.74	40.68	65.2
Qwen2.5-VL	72B	54.70	40.42	66.1
NVLM-D	72B	51.29	36.02	70.8
LLaVA-OneVision	7B	51.01	35.58	71.6
LLaVA-1.5	7B	49.05	32.92	74.9
InternVL3	78B	38.24	18.48	92.2
Mistral Small	24B	36.21	15.50	96.7
Proprietary VLMs				
Claude Sonnet 4.5	–	56.77	43.38	62.0
GPT-5	–	55.65	41.62	64.6
Gemini 2.5 Pro	–	55.33	41.17	65.6
Gemini 2.5 Flash	–	55.13	40.67	67.0
GPT-o4-mini	–	54.76	40.61	65.6
GPT-4o	–	54.19	39.67	67.4

Figure 25. **Example Questions with Images and Answer Choices: Hardest vs Easiest.** Left column shows hardest questions (0-10% of models answered correctly), highlighting challenges in fine-grained spatial reasoning, technical detail recognition, and complex relational understanding. Right column shows easiest questions (90-100% of models answered correctly), typically involving basic object presence or simple binary attributes. For each example, the image is shown on the left with the question text and multiple-choice options displayed together in a white box with domain-colored borders (green=Natural, blue=Document, red=E-commerce, orange=Embodied AI). The correct answer is marked with a checkmark (✓). Note that different images are used for hardest vs easiest questions to avoid confounding factors.

Table 10. Results on the **Natural** domain with the **Taxonomy-Hinted** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	8B	78.99	73.78	23.9
Qwen3-VL	4B	76.30	70.27	27.7
GLM-4.1V	9B	75.87	69.79	27.9
Qwen2.5-VL	32B	74.64	68.51	28.2
Qwen3-VL	30B-A3B	74.46	68.29	7.2
Qwen2.5-VL	72B	72.93	66.25	30.6
NVLM-D	72B	72.63	66.25	29.4
InternVL3	8B	71.57	64.10	34.4
InternVL3.5	38B	69.30	61.24	37.2
Qwen2.5-VL	7B	68.46	59.66	40.7
InternVL3-14B	14B	68.27	59.60	40.1
InternVL3.5	8B	65.77	56.29	43.9
InternVL3.5	30B-A3B	65.66	56.12	44.1
InternVL3.5	1B	65.58	56.13	43.7
LLaVA-OneVision	7B	63.24	53.14	46.8
LLaVA-1.5	7B	51.24	36.58	68.1
InternVL3	78B	38.39	19.33	88.9
Mistral Small	24B	35.95	15.71	94.5
Proprietary VLMs				
Gemini 2.5 Flash	–	87.55	85.11	11.0
Gemini 2.5 Pro	–	87.47	85.01	11.2
GPT-5	–	86.87	83.79	14.1
GPT-o4-mini	–	82.50	78.57	18.0
GPT-4o	–	78.30	72.82	25.1
Claude Sonnet 4.5	–	77.36	72.14	24.1

Table 11. Results on the **Document** domain with the **Long** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	88.34	87.00	4.6
Qwen3-VL	8B	86.63	85.19	4.9
Qwen3-VL	30B-A3B	86.05	84.46	5.5
Qwen3-VL	4B	84.63	82.72	6.6
Qwen2.5-VL	32B	82.14	79.54	9.0
Qwen2.5-VL	72B	80.36	77.11	11.2
InternVL3.5	38B	78.78	74.92	13.4
InternVL3.5	8B	77.29	73.21	14.1
InternVL3.5	30B-A3B	76.45	71.86	15.9
Qwen2.5-VL	7B	76.35	72.08	14.8
InternVL3	8B	72.95	67.71	18.1
InternVL3	14B	71.88	66.35	19.1
InternVL3.5	1B	67.50	60.33	24.7
NVLM-D	72B	65.39	57.73	26.5
LLaVA-OneVision	7B	61.93	52.41	32.9
LLaVA-1.5	7B	36.22	12.05	83.7
InternVL3	78B	34.29	9.01	87.5
Mistral Small	24B	30.75	2.56	97.6
Proprietary VLMs				
GPT-5	–	90.01	89.11	3.1
Gemini 2.5 Pro	–	88.67	87.60	3.7
GPT-o4-mini	–	88.38	87.15	4.2
Gemini 2.5 Flash	–	86.90	85.34	5.4
Claude Sonnet 4.5	–	85.08	82.70	8.2
GPT-4o	–	82.18	79.34	9.8

Table 12. Results on the **Document** domain with the **Simple** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	87.86	86.36	5.1
Qwen3-VL	30B-A3B	85.89	84.21	5.8
Qwen3-VL	8B	85.85	84.10	6.0
Qwen3-VL	4B	84.99	83.02	6.8
Qwen2.5-VL	32B	82.67	80.08	8.9
Qwen2.5-VL	72B	80.56	77.42	10.8
InternVL3.5	38B	78.91	74.92	13.8
InternVL3.5	8B	78.56	74.72	13.3
InternVL3.5	30B-A3B	77.72	73.64	14.1
InternVL3	8B	75.83	71.66	14.4
Qwen2.5-VL	7B	75.85	71.07	16.5
InternVL3	14B	74.17	69.55	16.1
InternVL3.5	1B	68.08	61.30	23.4
NVLM-D	72B	65.25	57.27	27.6
LLaVA-OneVision	7B	61.45	52.11	32.3
LLaVA-1.5	7B	36.48	12.22	84.0
InternVL3	78B	34.19	9.03	87.2
Mistral Small	24B	30.81	2.56	97.9
Proprietary VLMs				
GPT-5	–	90.81	90.09	2.5
Gemini 2.5 Flash	–	88.97	87.75	4.1
Gemini 2.5 Pro	–	88.66	87.42	4.3
GPT-o4-mini	–	88.14	86.84	4.5
Claude Sonnet 4.5	–	83.09	80.18	10.0
GPT-4o	–	82.55	79.87	9.3

Table 13. Results on the **Document** domain with the **Short** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	55.69	40.47	52.7
Qwen3-VL	30B-A3B	48.90	32.04	58.5
Qwen3-VL	4B	48.12	31.28	58.4
Qwen2.5-VL	32B	45.32	26.20	66.2
InternVL3.5	8B	45.26	26.01	66.7
InternVL3.5	30B-A3B	44.76	25.60	66.4
InternVL3	8B	44.58	25.13	67.4
InternVL3	14B	44.57	25.54	66.0
InternVL3.5	38B	44.24	24.50	68.4
InternVL3.5	1B	43.69	23.61	69.6
Qwen2.5-VL	7B	43.32	23.07	70.1
Qwen3-VL	8B	43.27	22.98	70.3
Qwen2.5-VL	72B	42.93	22.33	71.4
NVLM-D	72B	40.99	19.26	75.3
LLaVA-OneVision	7B	38.28	14.70	81.7
LLaVA-1.5	7B	35.33	10.12	87.3
InternVL3	78B	33.70	7.48	90.8
Mistral Small	24B	30.78	2.26	98.8
Proprietary VLMs				
GPT-o4-mini	–	46.40	27.99	63.7
Gemini 2.5 Pro	–	46.25	27.73	64.1
Claude Sonnet 4.5	–	46.04	27.64	63.7
GPT-5	–	44.85	25.79	66.0
Gemini 2.5 Flash	–	44.09	24.51	67.8
GPT-4o	–	42.84	22.29	71.2

Table 14. Results on the **Document** domain with the **Taxonomy-Hinted** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	72.40	67.10	18.3
Qwen2.5-VL	32B	71.44	65.77	19.5
Qwen3-VL	8B	70.82	64.83	20.7
Qwen3-VL	4B	66.63	58.82	27.1
Qwen3-VL	30B-A3B	62.49	52.26	35.3
Qwen2.5-VL	72B	62.48	53.66	30.5
NVLM-D	72B	59.61	49.29	35.7
Qwen2.5-VL	7B	57.05	44.95	41.8
LLaVA-OneVision	7B	56.57	44.48	41.8
InternVL3.5	1B	56.37	44.04	42.7
InternVL3	8B	55.05	42.42	43.6
InternVL3	14B	54.52	41.36	45.5
InternVL3.5	8B	52.60	38.08	50.1
InternVL3.5	38B	51.94	37.38	50.4
InternVL3.5	30B-A3B	51.83	37.53	49.5
LLaVA-1.5	7B	35.65	10.94	85.5
InternVL3	78B	33.41	7.29	90.4
Mistral Small	24B	30.60	2.39	97.8
Proprietary VLMs				
Claude Sonnet 4.5	–	83.23	80.90	8.0
Gemini 2.5 Flash	–	81.89	79.36	8.7
Gemini 2.5 Pro	–	81.34	78.71	9.0
GPT-o4-mini	–	75.08	69.62	18.8
GPT-5	–	72.36	65.39	24.1
GPT-4o	–	63.48	54.58	30.8

Table 15. Results on the **E-commerce** domain with the **Long** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	30B-A3B	94.27	93.34	3.6
Qwen3-VL	8B	93.95	92.86	4.2
Qwen3-VL	4B	93.62	92.54	4.2
GLM-4.1V	9B	93.03	91.45	6.1
Qwen2.5-VL	32B	90.76	88.97	6.9
Qwen2.5-VL	72B	89.85	87.58	8.8
InternVL3.5	38B	87.53	84.20	12.8
Qwen2.5-VL	7B	87.21	84.13	11.8
InternVL3.5	30B-A3B	85.78	81.87	15.1
InternVL3.5	8B	85.77	82.03	14.5
InternVL3	8B	84.53	80.50	15.6
InternVL3	14B	84.05	79.77	16.6
InternVL3.5	1B	80.12	74.29	22.5
NVLM-D	72B	79.05	73.07	23.1
LLaVA-OneVision	7B	77.24	70.51	26.1
LLaVA-1.5	7B	48.39	30.43	69.6
InternVL3	78B	38.46	16.51	85.1
Mistral Small	24B	34.59	10.75	92.5
Proprietary VLMs				
GPT-5	–	96.11	95.62	1.9
Gemini 2.5 Pro	–	95.60	94.94	2.6
Gemini 2.5 Flash	–	95.55	94.94	2.4
GPT-o4-mini	–	94.22	93.32	3.5
GPT-4o	–	91.14	89.07	8.0
Claude Sonnet 4.5	–	91.11	88.94	8.4

Table 16. Results on the **E-commerce** domain with the **Simple** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	30B-A3B	93.90	92.85	4.1
Qwen3-VL	4B	93.77	92.63	4.4
Qwen3-VL	8B	93.35	92.05	5.1
GLM-4.1V	9B	92.04	90.18	7.1
Qwen2.5-VL	32B	90.81	88.75	7.9
Qwen2.5-VL	72B	89.07	86.20	11.0
Qwen2.5-VL	7B	85.38	81.31	15.7
InternVL3	8B	87.01	83.59	13.2
InternVL3-14B	14B	86.17	82.35	14.8
InternVL3.5	38B	86.47	82.69	14.6
InternVL3.5	8B	86.60	82.98	13.9
InternVL3.5	30B-A3B	85.79	81.80	15.4
InternVL3.5	1B	82.69	77.76	19.0
NVLM-D	72B	78.46	72.43	23.4
LLaVA-OneVision	7B	75.09	67.38	30.0
LLaVA-1.5	7B	49.00	31.12	69.2
InternVL3	78B	38.47	16.58	84.9
Mistral Small	24B	34.52	10.31	93.8
Proprietary VLMs				
Gemini 2.5 Flash	–	95.73	95.09	2.5
GPT-5	–	94.73	93.78	3.6
Gemini 2.5 Pro	–	93.91	92.69	4.7
GPT-o4-mini	–	93.18	91.67	5.8
GPT-4o	–	91.40	89.28	8.2
Claude Sonnet 4.5	–	88.86	85.91	11.4

Table 17. Results on the **E-commerce** domain with the **Short** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	64.05	50.97	50.6
Qwen3-VL	4B	61.69	48.32	51.6
Qwen3-VL	30B-A3B	61.23	47.60	52.6
InternVL3.5	38B	59.16	44.75	55.7
Qwen2.5-VL	32B	57.89	43.17	56.9
Qwen3-VL	8B	57.66	42.71	57.8
InternVL3.5	30B-A3B	57.58	42.32	59.0
Qwen2.5-VL	72B	57.06	41.68	59.4
InternVL3	8B	56.99	41.62	59.4
Qwen2.5-VL	7B	56.88	41.39	59.8
InternVL3.5	8B	56.67	41.17	59.9
InternVL3-14B	14B	56.27	40.60	60.5
InternVL3.5	1B	55.23	39.26	61.7
NVLM-D	72B	50.88	33.28	68.1
LLaVA-OneVision	7B	49.80	31.28	71.6
LLaVA-1.5	7B	43.91	23.38	79.4
InternVL3	78B	35.32	11.79	91.1
Mistral Small	24B	33.03	8.17	96.3
Proprietary VLMs				
Claude Sonnet 4.5	–	59.31	44.90	55.7
Gemini 2.5 Pro	–	58.98	44.02	57.8
GPT-5	–	57.36	41.98	59.4
GPT-o4-mini	–	56.84	41.40	59.6
Gemini 2.5 Flash	–	56.77	41.03	60.8
GPT-4o	–	55.48	39.29	62.5

Table 18. Results on the **E-commerce** domain with the **Taxonomy-Hinted** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	8B	84.19	80.29	15.1
GLM-4.1V	9B	83.87	79.92	15.4
Qwen2.5-VL	72B	83.80	79.95	15.0
Qwen2.5-VL	32B	83.55	79.68	15.1
Qwen3-VL	4B	82.98	78.83	16.1
Qwen3-VL	30B-A3B	76.27	68.72	29.3
NVLM-D	72B	76.12	69.54	25.6
InternVL3	8B	75.82	69.57	24.3
Qwen2.5-VL	7B	75.21	68.71	25.3
InternVL3	14B	74.83	67.84	27.2
InternVL3.5	38B	73.46	65.65	30.4
InternVL3.5	30B-A3B	73.27	65.34	30.8
InternVL3.5	8B	71.70	63.49	31.9
LLaVA-OneVision	7B	69.78	60.24	36.9
InternVL3.5	1B	68.54	59.16	36.4
LLaVA-1.5	7B	47.99	29.56	71.4
InternVL3	78B	39.11	18.01	81.9
Mistral Small	24B	35.87	12.98	88.8
Proprietary VLMs				
Gemini 2.5 Flash	–	93.41	92.37	4.1
Gemini 2.5 Pro	–	91.50	89.79	6.7
Claude Sonnet 4.5	–	89.74	87.71	7.9
GPT-5	–	88.63	85.96	10.4
GPT-o4-mini	–	86.89	83.74	12.3
GPT-4o	–	81.51	76.73	18.6

Table 19. Results on the **Embodied AI** domain with the **Long** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	30B-A3B	82.55	78.79	13.2
Qwen3-VL	8B	81.06	76.87	14.7
Qwen3-VL	4B	80.35	75.67	16.4
GLM-4.1V	9B	79.01	73.68	18.7
InternVL3.5	38B	75.41	68.84	23.1
Qwen2.5-VL	72B	73.77	66.72	24.7
Qwen2.5-VL	32B	73.73	67.20	22.9
InternVL3.5	30B-A3B	70.02	61.28	30.6
Qwen2.5-VL	7B	68.74	59.41	32.8
InternVL3.5	8B	68.41	58.89	33.5
NVLM-D	72B	67.15	56.78	36.5
InternVL3	14B	66.89	56.71	35.7
LLaVA-OneVision	7B	65.67	55.16	36.9
InternVL3	8B	65.05	53.55	40.4
InternVL3.5	1B	64.68	53.97	37.7
LLaVA-1.5	7B	50.01	32.29	62.4
Mistral Small	24B	39.24	15.05	85.2
InternVL3	78B	33.84	7.34	93.5
Proprietary VLMs				
Gemini 2.5 Flash	–	86.97	84.27	9.4
Gemini 2.5 Pro	–	86.78	84.14	9.3
GPT-o4-mini	–	83.33	80.34	10.5
GPT-4o	–	83.21	79.50	13.0
GPT-5	–	82.83	80.41	1.6
Claude Sonnet 4.5	–	69.90	61.12	30.9

Table 20. Results on the **Embodied AI** domain with the **Simple** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	30B-A3B	82.15	78.13	14.1
Qwen3-VL	4B	80.56	75.93	16.2
Qwen3-VL	8B	80.37	75.70	16.3
GLM-4.1V	9B	75.56	68.71	24.1
InternVL3.5	38B	74.68	67.65	24.7
Qwen2.5-VL	32B	72.98	65.85	25.1
InternVL3	8B	72.07	63.97	28.4
Qwen2.5-VL	72B	71.60	63.36	28.9
NVLM-D	72B	70.31	61.75	30.1
InternVL3	14B	69.75	61.00	30.7
InternVL3.5	30B-A3B	69.75	60.75	31.6
Qwen2.5-VL	7B	68.36	58.83	33.5
InternVL3.5	8B	67.24	57.25	35.1
InternVL3.5	1B	64.46	53.73	37.7
LLaVA-OneVision	7B	61.01	48.39	44.4
LLaVA-1.5	7B	49.84	31.73	63.8
Mistral Small	24B	33.78	6.34	96.5
InternVL3	78B	34.32	7.88	93.2
Proprietary VLMs				
GPT-5	–	86.82	84.21	9.1
Gemini 2.5 Pro	–	85.45	82.22	11.3
Gemini 2.5 Flash	–	84.89	81.63	11.4
GPT-o4-mini	–	82.94	79.21	13.1
GPT-4o	–	81.61	77.11	15.7
Claude Sonnet 4.5	–	67.27	57.08	35.8

Table 21. Results on the **Embodied AI** domain with the **Short** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
GLM-4.1V	9B	61.29	46.72	51.1
Qwen3-VL	4B	59.63	44.33	53.7
Qwen3-VL	30B-A3B	57.56	41.23	57.4
InternVL3.5	38B	57.16	40.78	57.5
Qwen3-VL	8B	56.24	39.24	59.7
InternVL3	8B	55.17	37.88	60.8
Qwen2.5-VL	72B	54.83	37.22	61.8
InternVL3.5	30B-A3B	54.61	37.33	60.7
InternVL3	14B	53.77	35.68	63.5
Qwen2.5-VL	32B	53.58	35.66	63.0
Qwen2.5-VL	7B	53.27	35.12	63.8
InternVL3.5	1B	52.66	34.73	63.0
InternVL3.5	8B	52.12	33.59	65.1
NVLM-D	72B	51.53	32.97	65.2
LLaVA-OneVision	7B	49.41	29.63	69.6
LLaVA-1.5	7B	46.75	26.02	73.0
InternVL3	78B	34.03	7.36	94.1
Mistral Small	24B	33.95	6.59	96.3
Proprietary VLMs				
GPT-o4-mini	–	57.76	41.39	57.5
GPT-5	–	57.73	41.36	57.5
Gemini 2.5 Pro	–	57.40	40.76	57.9
Gemini 2.5 Flash	–	55.78	38.43	61.0
GPT-4o	–	54.47	36.50	63.1
Claude Sonnet 4.5	–	53.16	34.97	64.0

Table 22. Results on the **Embodied AI** domain with the **Taxonomy-Hinted** prompt. Score and Acc are reported in %. Cannot is the percentage of questions where the model caption does not contain an answer.

Model	Size	Score↑	Acc↑	Cannot↓
Open-Source VLMs				
Qwen3-VL	8B	74.85	67.32	26.3
GLM-4.1V	9B	72.59	64.33	28.9
Qwen2.5-VL	32B	70.02	61.21	30.8
Qwen3-VL	30B-A3B	70.70	61.67	31.7
Qwen2.5-VL	72B	69.66	60.21	33.2
Qwen3-VL	4B	69.55	59.76	34.2
InternVL3	8B	67.46	57.27	35.7
InternVL3.5	38B	66.35	55.41	38.4
InternVL3.5	30B-A3B	65.15	53.84	39.5
NVLM-D	72B	64.42	53.18	39.4
Qwen2.5-VL	7B	63.44	51.65	41.3
InternVL3	14B	62.86	50.87	42.0
LLaVA-OneVision	7B	62.03	49.32	44.7
InternVL3.5	8B	61.62	48.96	44.4
InternVL3.5	1B	60.61	47.88	44.7
LLaVA-1.5	7B	46.92	27.44	68.6
InternVL3	78B	34.40	8.13	92.7
Mistral Small	24B	34.17	7.22	95.0
Proprietary VLMs				
Gemini 2.5 Pro	–	85.35	81.86	12.1
GPT-5	–	85.10	81.37	13.1
Gemini 2.5 Flash	–	82.81	78.42	15.4
GPT-o4-mini	–	80.08	74.69	18.9
GPT-4o	–	76.27	69.30	24.4
Claude Sonnet 4.5	–	70.44	62.23	28.8