VISUAL CONTEXT COMPOSITION: RENDERING AND GROUPING EVIDENCE FOR MULTI-IMAGE VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-image tasks require vision-language models (VLMs) to relate images, labels, and instructions. Text can enter the language decoder directly or pass through the visual encoder when rendered in an image. Placing multiple elements in one image also allows them to be visually encoded together. We introduce Visual Context Composition (VCC), a training-free method that controls these processing routes while keeping images with their associated labels. Across five VLMs and four benchmark families, the selected VCC input configurations improve average task scores over baseline inputs. Giving each labeled example its own image performs best on average for fine-grained classification; separating task roles, such as instructions, references, and queries, performs best in the other benchmark families. Fine-grained gains persist for two models when individual image sizes or total input pixel counts are matched in separate comparisons. With rendered content held fixed, changing which task roles share an image also changes accuracy. We hypothesize that joint visual encoding represents task-relevant relationships. Assigned labels can be read from reference-image features. Token-removal experiments further show that information propagated from a region to other visual tokens contributes to predictions even after that region's own tokens are removed.

1 INTRODUCTION

Many tasks require a vision-language model (VLM) to relate evidence across multiple images and text. A model may need to associate an image with its description, compare several images, or use demonstrations to answer a new query. Multi-image and in-context learning benchmarks evaluate these uses of context (Zong et al., 2025; Wang et al., 2025; Chen et al., 2025; Zhao et al., 2024a).

Input format determines the route information takes through a VLM. Text supplied separately from images reaches the language decoder as text tokens, whereas text rendered into an image is processed by the visual encoder. In models that encode images separately, such as Qwen3-VL, elements sharing an image can interact during visual encoding (Qwen Team, 2025). These choices change the visual representations supplied to the decoder. Which elements should pass through visual encoding together, and how does this affect performance?

We introduce Visual Context Composition (VCC) to change these processing routes with model weights and task evidence fixed. Rendering scope selects elements to draw as pixels; image grouping determines which are encoded together. PixelWorld studies a common pixel representation (Lyu et al., 2025), while CollagePrompt and MIRB examine image arrangement and merging (Xu et al., 2025; Zhao et al., 2024a). We control both choices for images and associated text within a task and compare joint and separate encoding of fixed rendered evidence. Figure 3 shows three groupings: *Joint* combines the selected elements; *Role* separates instructions, the full reference set, and the query; *Example* further separates labeled references. Each reference image stays with its label.

Across the evaluated configurations, *Example* has the highest mean for fine-grained classification and *Role* elsewhere. Fine-grained gains remain supported under separate size and pixel controls for Qwen3-VL-32B and GPT-5.4. On Concentric, a synthetic color-matching task, fixed-pixel comparisons with Qwen3-VL-32B show that joining instructions with references improves accuracy, while joining references with the query reduces it for textual queries; image-query effects are not signif-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

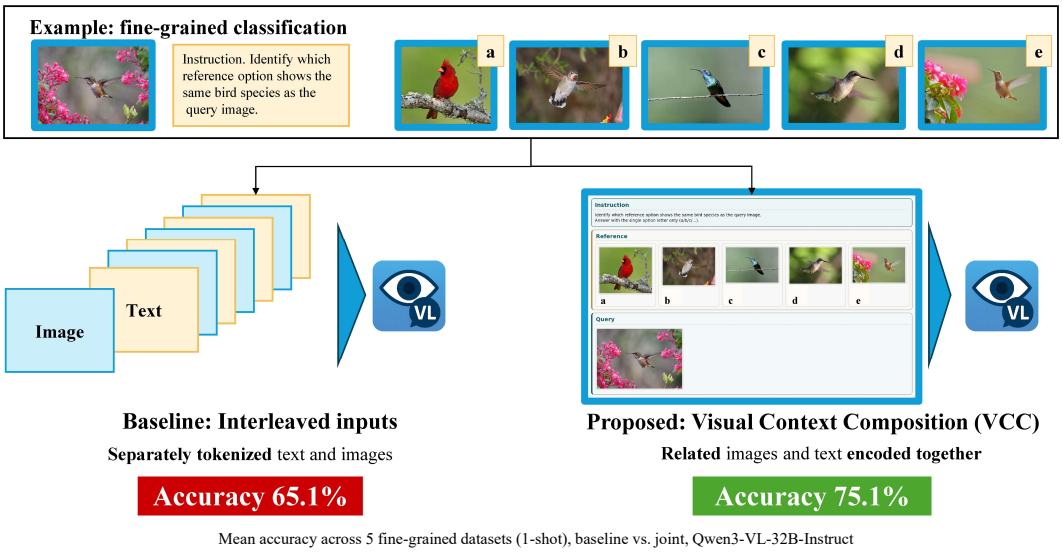


Figure 1: **Overview of Visual Context Composition (VCC).** The illustrated *Joint* configuration encodes the instruction, labeled references, and query in one image. The main fine-grained evaluation in Table 2 uses *Example*.

icant. Separate Qwen3-VL-8B analyses show reference-label readout and predictive contributions through other visual tokens. Their connection to benchmark gains remains unresolved.

Our contributions are:

1. We introduce VCC, a training-free method that changes how task information passes through visual encoding by controlling rendering scope and image grouping. Configurations selected for each benchmark family improve average scores across five VLMs.
2. We show that preferred image groupings vary across benchmark tasks and that grouping changes accuracy even with rendered content fixed.
3. We investigate the hypothesis that joint visual encoding represents task-relevant relationships. Reference features encode assigned labels, and separate token-removal experiments show predictive contributions through other visual tokens.

2 Related Work

Multi-image context. VLMs trained on interleaved image–text sequences can generate outputs using context from multiple images and text (Alayrac et al., 2022; Awadalla et al., 2023; Laurençon et al., 2023; Sun et al., 2024). Multi-image instruction tuning and training across image and video tasks further develop these capabilities (Jiang et al., 2024a; Zhao et al., 2024b; Li et al., 2025b;a).

JUDO trains a VLM to identify and explain anomalies by comparing query images with normal references (Kang et al., 2026). VCC evaluates how rendering and image grouping affect outputs and representations before decoding, with model weights fixed.

Evaluating the use of multimodal context. Multi-image evaluation tests whether VLMs can relate evidence across images and text. Controlled tasks probe compositional relationships through paired photographs and captions (Suhr et al., 2019; Thrush et al., 2022), while broader benchmarks cover diverse image relationships and interleaved reasoning (Meng et al., 2025; Xia et al., 2025; Du et al., 2025; Wang et al., 2025). Demonstration-based evaluations test whether VLMs use the supplied examples to generate outputs (Zong et al., 2025; Chen et al., 2025).

Evaluation results also depend on how examples and images are supplied. Studies examine the number of demonstrations and whether multiple queries are answered together (Jiang et al., 2024b).

MIRB compares separate images with a merged image and reports lower accuracy after merging for most tested VLMs, with exceptions in the LLaVA family (Zhao et al., 2024a). CollagePrompt examines the cost–accuracy trade-off of recognizing multiple images in one collage and shows that their arrangement affects accuracy (Xu et al., 2025).

Rendered language and structured visual evidence. Document and screenshot models learn to interpret text together with its visual surroundings (Huang et al., 2022; Kim et al., 2022; Lee et al., 2023). Pix2Struct renders prompts into image headers, whereas DePlot converts charts into text tables for language-model reasoning (Liu et al., 2023).

Recent work studies text rendered as pixels more directly. PixelWorld evaluates a common pixel representation for textual and visual inputs across tasks (Lyu et al., 2025). Glyph trains VLMs on compressed visual pages, while Text-Printed Image renders descriptions as synthetic images for task-specific training (Cheng et al., 2025; Yamabe et al., 2025). Performance depends on the task and rendering conditions: Sun et al. (2026) report sensitivity to font and resolution, as well as shorter reasoning outputs under image input. VCC separately controls rendering scope and image grouping, testing joint versus separate encoding of the same rendered evidence at fixed scale.

Interventions on multimodal representations. Multimodal interpretability uses attribution, probes, and interventions to examine how representations contribute to model outputs (Dang et al., 2024; Lin et al., 2025). Causal tracing and attention analyses study information storage and transfer, including the contribution of visual information to question answering (Palit et al., 2023; Basu et al., 2024; Zhang et al., 2025; Yu & Ananiadou, 2024). Analyses of factual recall further relate failures in selected VLMs to the late emergence of visual entity representations (Venhoff et al., 2025).

These methods distinguish the information available in a representation from its use in generating outputs. Probe accuracy depends partly on the learned readout (Hewitt & Liang, 2019); projection methods can suppress selected linear readouts without certifying removal of all relevant information (Ravfogel et al., 2020). VCC combines probes with token removal to measure the presence of contextual information and its contribution to predictions before language decoding, as examined in Section 4.5.

3 METHOD

3.1 PROBLEM SETUP

We consider tasks that require relating evidence distributed across images and text, such as answering a question using reference images and their descriptions. The input contains both the evidence and its supplied associations. We change how this information reaches the decoder while keeping model weights fixed.

Let x denote an ordered collection of image and text elements, with image subsequence x_{img} and text subsequence x_{txt}. The visual encoder maps the images to visual representations $E_\theta(x_{\mathrm{img}})$. The language decoder D_θ receives these representations together with the text. Suppressing sequence-position notation, the baseline output is $y_{\mathrm{baseline}} = D_\theta(E_\theta(x_{\mathrm{img}}), x_{\mathrm{txt}})$. Text token embeddings enter the language decoder without visual encoding. Rendering that text as part of an image instead routes it through visual encoding. Image grouping determines which elements undergo visual encoding together.

3.2 HYPOTHESIS

In architectures that encode input images separately, such as Qwen3-VL, visual attention allows information exchange within an image before its representations reach the decoder (Qwen Team, 2025). Figure 2 illustrates how images and rendered text placed in the same image undergo this shared visual computation. We hypothesize that joint visual encoding represents relationships among task images and text that the language decoder can use to produce accurate outputs.

These relationships may include a reference image and its assigned label, or a query and its references. Their representations could support retrieval and association of the supplied evidence.

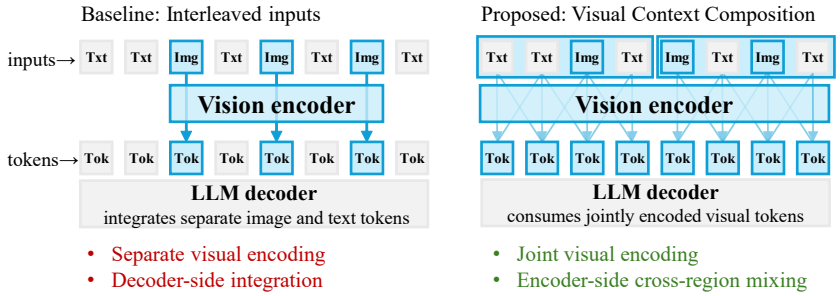


Figure 2: **Information routes in VCC.** Images are encoded separately; elements sharing an image can interact within the visual encoder. External text enters the language decoder directly.

Table 1: **Contributions and supporting evidence.**

Contr.	Question	Finding	Evidence
C1	Q1. Does VCC improve task performance?	Higher benchmark scores; grouping selection transfers across tasks within each benchmark family	Sec. 4.3; Tabs. 2, 3, 4
C2	Q2. How does image grouping affect task performance?	Preferred groupings vary across tasks, and grouping affects accuracy even with rendered content fixed.	Tabs. 5, 6
C3	Q3. How does VCC change token representations?	Label readout; predictive benefit through other tokens	Sec. 4.5; Fig. 4; Tab. 7

3.3 Visual Context Composition

An input configuration specifies rendering scope, image grouping, and rendering settings. The scope selects the elements included in composed images. It can include instructions, descriptions, labels, questions, and source images. Elements outside this scope remain as separate images or decoder text. Rendering text alongside source images sends the text through the visual encoder and changes the images' visual context.

Image grouping determines which selected elements share an input image and are therefore encoded together. A group can contain an image with its description, a demonstration with its answer, or several examples with their common label. VCC preserves supplied image–text associations and varies the surrounding context. Groups may share an image or form separate input images.

Let S specify the elements selected for visual composition and Π assign them to output images. The image-building function $\Phi_{S,\Pi}$ returns the composed images together with any source images outside S. The text function τ_S retains unrendered text and supplies deterministic helper text, such as "Answer:", to specify the response format. With the model weights θ fixed, the output is

$$y_{\mathrm{VCC}} = D_\theta\Big(E_\theta\big(\Phi_{S,\Pi}(x_{\mathrm{img}}, x_{\mathrm{txt}})\big), \tau_S(x_{\mathrm{txt}})\Big). \tag{1}$$

The layout assigns coordinates and sizes to the elements in S and marks the image groups specified by Π. Headings and spacing distinguish task roles. Text is then drawn alongside the source images at these locations. Finally, VCC exports the designated image units, retaining the canvas or extracting its regions; source images outside S remain unchanged.

VCC is deterministic for a given configuration and task instance, without using the query answer. Section 4 defines the evaluated groupings.

4 Experiments

Table 1 relates the contributions to their questions and supporting sections.

4.1 BENCHMARKS AND MODELS

We group the evaluations into four benchmark families: fine-grained classification, VL-ICL, TrueMICL, and MuirBench. We use five VLMs: GPT-5.4 (OpenAI, 2026), Qwen3-VL-32B-Instruct (Qwen Team, 2025), Qwen3.5-27B (Qwen Team, 2026), Gemma4-31B-IT (Abd et al., 2026), and GLM-4.6V-Flash (Z.ai, 2025).

For fine-grained classification, we use Oxford Flowers 102 (Nilsback & Zisserman, 2008), CUB-200-2011 (Wah et al., 2011), Food-101 (Bossard et al., 2014), Stanford Cars (Krause et al., 2013), and Stanford Dogs (Khosla et al., 2011), with 500 queries per dataset. Five-way, one-shot episodes follow the support–query structure of few-shot learning (Vinyals et al., 2016; Snell et al., 2017).

We evaluate on four subsets of VL-ICL Bench (Zong et al., 2025): Open-MI, CLEVR, TextOCR, and operator induction, at zero, one, two, and four shots. Text-to-image tasks are excluded. We use all seven task subsets of TrueMICL (Chen et al., 2025) at four shots and all 12 task categories of MuirBench (Wang et al., 2025).

To study these processing choices with less reliance on prior object knowledge, we use Concentric, a synthetic five-way classification task. Queries are matched to labeled references by the outer-to-inner color sequences of three to five concentric square layers. The dataset has two query formats: ImageQuery presents the query as an image, as illustrated in Figure 4, and TextQuery presents a textual description of its color sequence.

4.2 IMAGE GROUPING

We compare the three image groupings illustrated in Figure 3: *Joint* combines all selected context; *Role* separates task roles; *Example* separates reference examples or demonstrations, each with its label or answer, and keeps the instruction and query separate.

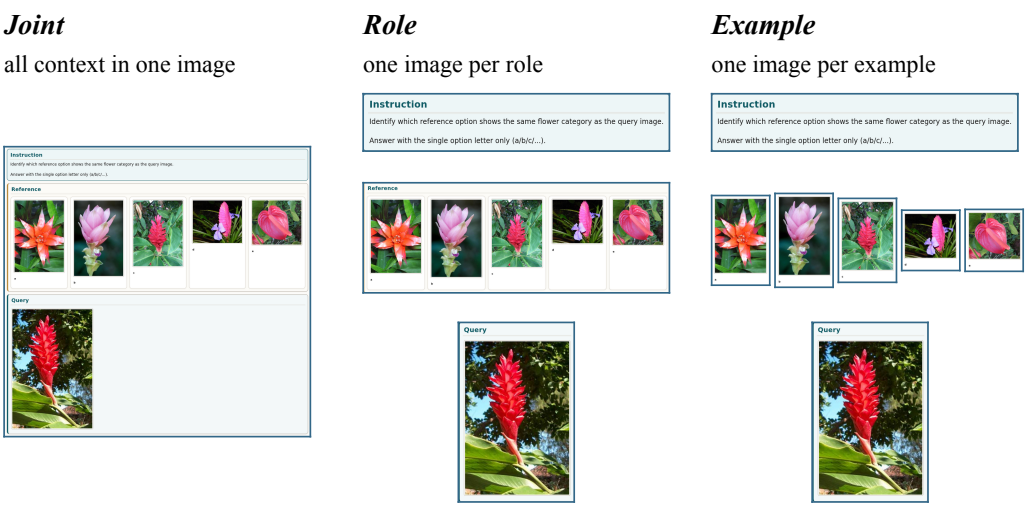


Figure 3: ***Joint*, *Role*, and *Example* groupings.** All configurations use the same instance. Borders mark separate input images; panel sizes are rescaled for display.

For fine-grained classification, VL-ICL, and TrueMICL, *Role* separates the instruction, reference set, and query. For MuirBench, it separates the question, visual evidence, and answer choices; *Example* is excluded because MuirBench has no labeled references or demonstrations.

We compare VCC with separate image–text inputs, explicitly specifying image–label associations in the fine-grained baseline. Models and task instances are fixed. We report accuracy for fine-grained classification, TrueMICL, and MuirBench, and Peak accuracy and ICL efficiency (Eff.) for VL-ICL (Zong et al., 2025). Peak is computed per task as the maximum accuracy over shot counts, then averaged across tasks. Eff. integrates accuracy above the zero-shot level and divides by the maximum possible area: the largest shot count times the full accuracy range. For VL-ICL, fine-grained classification, and TrueMICL, we use three support seeds and average across seeds and

Table 2: **Benchmark performance of VCC.** Base.: baseline score (%); ΔVCC: gain (pp). VCC uses *Example* for fine-grained classification and *Role* elsewhere. VL-ICL uses 0/1/2/4 shots; means weight models equally.

Model	Fine-grained		VL-ICL				TrueMICL		MuirBench	
	Accuracy		Peak		Eff.		Accuracy		Accuracy	
	Base.	ΔVCC	Base.	ΔVCC	Base.	ΔVCC	Base.	ΔVCC	Base.	ΔVCC
GPT-5.4	81.40	+2.57	82.72	+2.59	40.91	+5.26	92.95	+1.62	68.74	+0.27
Qwen3-VL-32B	65.12	+13.64	89.49	+3.76	41.47	+6.44	91.48	+0.90	71.12	+0.77
Qwen3.5-27B	58.93	+8.44	53.18	+39.49	16.75	+19.53	79.71	+11.71	68.12	+6.92
Gemma4-31B-IT	79.60	+1.48	85.58	+6.88	39.31	+10.54	97.14	+1.05	70.77	+5.00
GLM-4.6V-Flash	73.32	+0.72	78.94	+10.21	32.13	+14.06	73.14	+6.86	65.15	+1.46
Mean ΔVCC	-	+5.37	-	+12.59	-	+11.17	-	+4.43	-	+2.89

Table 3: **Grouping selection on one task and evaluation on the rest.** Mean accuracy gains over Baseline (pp), averaged over source splits. Positive counts splits with positive mean gains. Pointwise 95% intervals use 10,000 cluster bootstrap draws, repeating selection in each draw and jointly resampling queries that share an image.

Family	Target tasks	Mean gain	95% CI	Positive
Fine-grained	4	+4.77	[+4.15, +5.42]	5/5
VL-ICL	3	+14.66	[+13.84, +15.60]	4/4
TrueMICL	6	+3.35	[+2.33, +4.48]	6/7
MuirBench	11	+2.35	[+1.05, +3.46]	12/12

tasks. Unparsed answers count as errors. Appendix B reports parsing coverage, exclusions, and experimental settings.

4.3 Task performance

Selecting a grouping. We select each family's highest-mean configuration from the comparisons in Table 5: *Example* for fine-grained classification and *Role* elsewhere. One configuration is shared across the family's models, tasks, and queries. Selection uses results from the full family; we test transfer to tasks excluded from selection separately below.

Table 2 compares these selected configurations with Baseline. All five models have higher point estimates for accuracy on fine-grained classification, TrueMICL, and MuirBench, and for Peak accuracy and ICL efficiency on VL-ICL. Appendix D reports comparisons and uncertainty estimates.

Transfer of grouping selection. Can a grouping selected on one task improve performance on other tasks? Within each family, we select the highest-accuracy configuration on one dataset, task, or MuirBench category and evaluate it on all remaining members. Candidates are *Joint*, *Role*, and *Example*, except that MuirBench excludes *Example*. Selection and evaluation average over the same five models; target-task results are excluded from selection. VL-ICL uses mean accuracy over one, two, and four shots. We repeat this procedure for every source member and average the target gains over Baseline across splits.

Table 3 shows positive mean gains in every family, with pointwise intervals above zero. All source splits improve the target mean except one in TrueMICL. This post-hoc analysis tests reuse within the existing evaluation collection. Two-source and other transfer settings appear in Appendix D.5.

Image size and total pixels. We compare Baseline, *Role*, and *Example* across five VLMs on the five fine-grained classification datasets. The image-size control preserves pixel-identical reference and query crops across configurations within each VLM. The total-pixel control matches the padded *Role* budget within 1%, including whitespace and rendered text. Each control has its own Baseline.

Table 4: **Image-size and total-pixel controls.** Accuracy (%) averaged equally over five VLMs and five fine-grained datasets; Δ denotes the difference from Baseline (pp).

Control	Base.	VCC-*Role*	Δ	VCC-*Example*	Δ
Image size	69.28	75.15	+5.87	72.13	+2.85
Total pixels	71.04	75.15	+4.11	74.35	+3.31

Table 5: **Task dependence of image grouping.** Base.: accuracy (%); VCC columns: changes (pp). Scores average common shot settings (1/2/4 for VL-ICL), then five models equally. Bold marks the highest estimate, without a significance claim.

Benchmark	Base.	VCC-*Joint*	VCC-*Role*	VCC-*Example*
Fine-grained	71.67	+3.23	−3.09	**+5.37**
VL-ICL	69.96	+14.32	**+15.40**	+15.08
TrueMICL	86.89	+2.06	**+4.43**	+3.23
MuirBench	68.78	+1.02	**+2.89**	–

Table 4 shows positive model-averaged gains for both configurations. Under both controls, *Example* mean-difference intervals are positive for Qwen3-VL-32B and GPT-5.4, span zero for Qwen3.5-27B and GLM-4.6V-Flash, and are negative for Gemma4-31B-IT. For Gemma, enforcing the controls requires a common $7/8$ image scale and disables the model's default automatic resizing. Departing from its standard preprocessing may contribute to the accuracy decreases, although this effect has not been isolated. Rendering, image count, and placement still differ, so these comparisons do not isolate grouping. Appendix D.1.1 reports protocols and paired comparisons.

4.4 Effects of input composition

The highest-scoring configuration varies across benchmark families, and grouping can affect accuracy even when rendered evidence is fixed. We establish these findings through benchmark comparisons and a controlled experiment on Concentric.

Grouping on benchmark tasks. We compare *Joint*, *Role*, and *Example* on the same five models to test whether a single configuration performs best across tasks. Table 5 averages the per-condition accuracies, first over shot settings and then over models. *Example* achieves the highest fine-grained average, while *Role* has the highest average on the other three families. *Role* has lower accuracy than Baseline on fine-grained classification. These averages favor different configurations for different benchmark families. Resizing and whitespace also differ across configurations, leaving open whether grouping affects accuracy when the rendered evidence is held fixed.

Grouping with rendered evidence fixed. To test whether grouping affects accuracy without changes in resizing or whitespace, we compare partitions of the same rendered canvas on Concentric. Here, I, R, and Q denote the instruction, labeled reference set, and query; bars separate input images. We render all three roles once and partition or rejoin the canvas without resizing or discarding pixels. Content, scale, total pixels, visual-token counts, and external text are fixed. Joining I and Q changes the sequence, so $IQ|R$ uses the order-matched $I|Q|R$ control; other groupings use $I|R|Q$. Image-local positions, decoder image delimiters, and encoder computation can still change.

Table 6 shows that, for TextQuery, joining instructions and references improves accuracy by 2.33 percentage points, whereas joining references and query reduces it by 3.17 points. Both effects survive Holm correction; the remaining contrasts, including all ImageQuery contrasts, are not significant. Thus, with rendered evidence fixed, the effect of grouping depends on which task roles share an image. The full-grouping comparison alone misses these opposing effects. This controlled result establishes a role-dependent grouping effect on Concentric TextQuery; it does not identify the cause of the benchmark rankings above.

Table 6: **Grouping with fixed rendered evidence.** Qwen3-VL-32B, one-shot; 200 queries and three support seeds per format. Effects (pp) are relative to $I|R|Q$, except $IQ|R$ versus $I|Q|R$. Intervals resample query clusters; Holm correction covers all eight contrasts.

Query	Grouping	Effect	95% CI	Holm p
Image	$IR\|Q$	-0.17	[-2.33, 1.67]	1.0000
	$I\|RQ$	+0.17	[-2.33, 2.50]	1.0000
	$IQ\|R$	+0.33	[-0.83, 1.67]	1.0000
	IRQ	+0.67	[-1.83, 3.17]	1.0000
Text	$IR\|Q$	+2.33	[1.00, 3.83]	0.0258
	$I\|RQ$	-3.17	[-5.33, -1.17]	0.0258
	$IQ\|R$	+0.17	[-1.00, 1.17]	1.0000
	IRQ	+0.50	[-0.83, 2.00]	1.0000

4.5 Visual Representations

We examine two aspects of the hypothesis in Section 3: whether visual features encode associations supplied in the current input, and whether information propagated during visual encoding contributes to predictions. We use readout probes for the first question and region masking with token removal for the second.

Label readout from visual representations. We test whether a reference image's features encode its assigned label, which must be obtained from the surrounding context. We also test whether the query answer can be read from query-region features using a linear probe. Using Qwen3-VL-8B on Concentric, we compare reference-label and query-answer readout on the same *Joint* inputs, holding source-image pixels fixed while permuting the assigned labels. Each probe uses mean features from the model's final vision merger, labeled "Projection" in Figure 4, with identical normalization and training procedures (Hewitt & Liang, 2019). Positions overlapping label glyphs or input patches that change across permutations are excluded. Both probes use the same rendered training images. Because they predict different targets, higher reference-label accuracy would not establish that reference features contain more task information than query features.

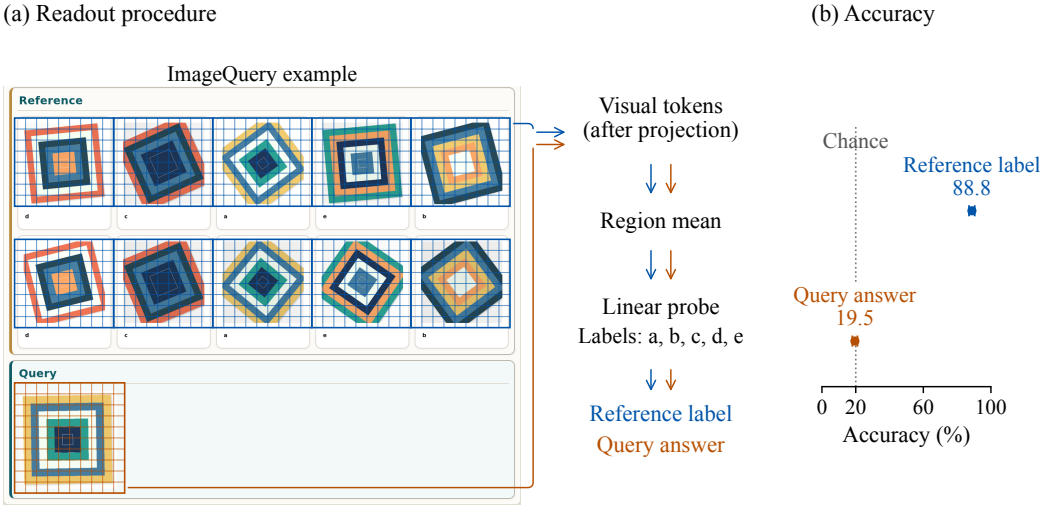


Figure 4: **Reference-label and query-answer readout.** *Joint* Concentric ImageQuery, Qwen3-VL-8B, two-shot. Error bars: 95% scene-bootstrap intervals on held-out scenes; chance: 20%.

Figure 4 shows that assigned labels can be read from reference-image features above chance on unseen ImageQuery scenes, while query-answer readout remains near chance. Appendix C.4 reports the same pattern for TextQuery. Because source-image pixels are fixed and label-overlapping

Table 7: **Predictive benefit through retained visual tokens.** *Joint* Concentric, Qwen3-VL-8B, 200 queries, 1/2-shot, three seeds (1,200 outputs per condition). Accuracy (%); Gain (pp): Full→Drop minus Mask→Drop.

	ImageQuery			TextQuery		
Removed region	Mask→Drop	Full→Drop	Gain (pp)	Mask→Drop	Full→Drop	Gain (pp)
Instruction	86.00	89.25	+3.25	69.67	73.67	+4.00
Reference image	18.25	30.00	+11.75	21.25	26.33	+5.08
Class label text	85.00	86.25	+1.25	68.33	73.50	+5.17
Query (image / text)	18.00	27.25	+9.25	21.00	65.42	+44.42

patches are excluded, the reference-label readout shows that joint visual encoding incorporates surrounding label information into the tokens representing the reference images.

Predictive benefit through other tokens. We next ask whether the contextual information found by readout helps the model produce correct answers. Appendix Figure 7 illustrates our comparison on *Joint* Concentric inputs with Qwen3-VL-8B, using query text as the target region.

In Full→Drop, we encode the complete image, then remove the target region's tokens before passing visual features to the decoder. In Mask→Drop, we first replace the target pixels with background, encode the masked image, and remove the same tokens. Both conditions retain identical token positions and counts across all visual paths, including DeepStack, with queries, support sets, and decoding settings matched. Only Full→Drop allows the encoder to see the target content and propagate its information to other tokens. Its accuracy advantage therefore measures whether this propagated information helps the decoder answer correctly.

Table 7 shows positive accuracy differences, largest for query and reference images in ImageQuery, and query text in TextQuery. Information propagated from query or reference regions to other tokens within the vision encoder therefore improves answer accuracy.

5 Discussion

Cross-model gains and within-family transfer support composition as a reusable, training-free intervention. Gains under size and pixel controls in two models show benefits beyond increased resolution.

Benchmark rankings and opposing fixed-rendering effects favor task-dependent grouping: joining more evidence need not help, because the roles sharing an image matter.

Readout and token removal show that visual encoding incorporates surrounding information into image tokens and that propagated information supports correct answers. Composition can thus shape the contextual evidence available before decoding.

6 Limitations

Gains and preferred configurations depend on the model, task, preprocessing, and reference count. Within-family transfer does not establish a selection rule for unseen tasks or rendering conditions.

Fixed-pixel effects are significant only for synthetic TextQuery in one model; representation analyses use another. Whether information propagation explains natural-task gains, and where reference–query matching occurs, remain unresolved.

7 Conclusion

VCC controls rendering and grouping with fixed model weights, improving average scores across five VLMs and four benchmark families. Experiments establish task-dependent grouping effects and show that context propagated during visual encoding contributes to correct answers.

ETHICS STATEMENT

The experiments use existing benchmark datasets and procedurally generated synthetic images. Representative benchmark images are reproduced to document the evaluated input configurations. Dataset sources are cited; their published licenses and research-use conditions, including the scientific fair use provision distributed with Food-101, are recorded in the accompanying third-party notices. The study involves no recruitment of human participants or collection of new human-subject data. The image-instruction and query-conflict studies characterize response behavior under specified inputs; they do not assess adversarial robustness or deployment safety.

AI USE STATEMENT

Generative AI was used to assist manuscript drafting and revision, hypothesis refinement, feedback on experimental methodology, interpretation of results, implementation, source and result auditing, figure-generation code, and build validation. Scientific claims and numerical tables are tied to saved experimental results; the accompanying audit records verification procedures and their limits. The authors take responsibility for the final text, claims, code, and figures.

REPRODUCIBILITY STATEMENT

Appendix B specifies datasets, model revisions, rendering, decoding, and aggregation. Appendices C.1–C.4 describe the grouping controls, token interventions, and probe training; Appendix D reports detailed benchmark results. The anonymous supplementary archive contains experimental settings, reported aggregates, dataset specifications, selected synthetic images, and representative input/output records. Its manifest links these artifacts to the paper and records omitted images and external data dependencies.

REFERENCES

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Ju-yeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya O. Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adámek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean-Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle K. Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, and Arthur Brazinskas. Gemma 4 Technical Report. *CoRR*, abs/2607.02770, July 2026. doi: 10.48550/ARXIV.2607.02770. URL `https://doi.org/10.48550/ARXIV.2607.02770`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information*

Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022., 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *CoRR*, abs/2308.01390, 2023. doi: 10.48550/ARXIV.2308.01390. URL https://doi.org/10.48550/ARXIV.2308.01390.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding Information Storage and Transfer in Multi-Modal Large Language Models. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0dfe31d6e703e138d46a7d2fced38b7c-Abstract-Conference.html.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - Mining Discriminative Components with Random Forests. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://doi.org/10.1007/978-3-319-10599-4_29.

Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. True Multimodal In-Context Learning Needs Attention to the Visual Context. *CoRR*, abs/2507.15807, July 2025. doi: 10.48550/ARXIV.2507.15807. URL https://doi.org/10.48550/ARXIV.2507.15807.

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, Yushi Bai, Jie Tang, Hongning Wang, and Minlie Huang. Glyph: Scaling Context Windows via Visual-Text Compression. *CoRR*, abs/2510.17800, October 2025. doi: 10.48550/ARXIV.2510.17800. URL https://doi.org/10.48550/ARXIV.2510.17800.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey. *CoRR*, abs/2412.02104, 2024. doi: 10.48550/ARXIV.2412.02104. URL https://doi.org/10.48550/ARXIV.2412.02104.

Hang Du, Jiayang Zhang, Guoshun Nan, Wendi Deng, Zhenyan Chen, Chenyang Zhang, Wang Xiao, Shan Huang, Yuqi Pan, Tao Qi, and Sicong Leng. From Easy to Hard: The MIR Benchmark for Progressive Interleaved Multi-Image Reasoning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 859–869. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00088. URL https://doi.org/10.1109/ICCV51701.2025.00088.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1275. URL https://doi.org/10.18653/V1/D19-1275.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 4083–4091. ACM,

2022. doi: 10.1145/3503161.3548112. URL https://doi.org/10.1145/3503161.3548112.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved Multi-Image Instruction Tuning. *Trans. Mach. Learn. Res.*, 2024a. URL https://openreview.net/forum?id=skLtdUVaJa.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H. Chen, and Andrew Y. Ng. Many-Shot In-Context Learning in Multimodal Foundation Models. *CoRR*, abs/2405.09798, 2024b. doi: 10.48550/ARXIV.2405.09798. URL https://doi.org/10.48550/ARXIV.2405.09798.

Hyunju Kang, Woohyun Lee, Jaewon Kim, and Hogun Park. JUDO: A Juxtaposed Domain-Oriented Multimodal Reasoner for Industrial Anomaly QA. *CoRR*, abs/2605.20284, May 2026. doi: 10.48550/ARXIV.2605.20284. URL https://doi.org/10.48550/ARXIV.2605.20284.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. URL http://vision.stanford.edu/aditya86/ImageNetDogs/main.html.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-Free Document Understanding Transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pp. 498–517. Springer, 2022. doi: 10.1007/978-3-031-19815-1_29. URL https://doi.org/10.1007/978-3-031-19815-1_29.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pp. 554–561. IEEE Computer Society, 2013. doi: 10.1109/ICCVW.2013.77. URL https://doi.org/10.1109/ICCVW.2013.77.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/e2cfb719f58585f779d0a4f9f07bd618-Abstract-Datasets_and_Benchmarks.html.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA.*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18893–18912. PMLR, 2023. URL https://proceedings.mlr.press/v202/lee23g.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Trans. Mach. Learn. Res.*, 2025a. URL https://openreview.net/forum?id=zKv8qULV6n.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=oSQiao9GqB.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, Soheil Feizi, and Lifu Huang. A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models. *CoRR*, abs/2502.17516, February 2025. doi: 10.48550/ARXIV.2502.17516. URL https://doi.org/10.48550/ARXIV.2502.17516.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023.*, volume ACL 2023 of *Findings of ACL*, pp. 10381–10399. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.660. URL https://doi.org/10.18653/V1/2023.FINDINGS-ACL.660.

Zhiheng Lyu, Xueguang Ma, and Wenhu Chen. PixelWorld: Towards Perceiving Everything as Pixels. *Trans. Mach. Learn. Res.*, 2025. URL https://openreview.net/forum?id=uY5eDN2bML.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=WsgEWL8i0K.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. DeepStack: Deeply Stacking Visual Tokens is Surprisingly Simple and Effective for LMMs. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/29cd7f8331d13ede6dc6d6ef3dfacb70-Abstract-Conference.html.

Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pp. 722–729. IEEE Computer Society, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

OpenAI. GPT-5.4. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.4. Accessed September 14, 2026.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pp. 2848–2853. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00307. URL https://doi.org/10.1109/ICCVW60793.2023.00307.

Qwen Team. Qwen3-VL Technical Report. *CoRR*, abs/2511.21631, November 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/ARXIV.2511.21631.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.647. URL https://doi.org/10.18653/V1/2020.ACL-MAIN.647.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4077–4087, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6418–6428. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1644. URL https://doi.org/10.18653/V1/P19-1644.

Kaiser Sun, Xiaochuang Yuan, Hongjun Liu, Chen Zhao, Cheng Zhang, Mark Dredze, and Fan Bai. Reading, Not Thinking: Understanding and Bridging the Modality Gap When Text Becomes Pixels in Multimodal LLMs. *CoRR*, abs/2603.09095, March 2026. doi: 10.48550/ARXIV.2603.09095. URL https://doi.org/10.48550/ARXIV.2603.09095.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative Multimodal Models are In-Context Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14398–14409. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01365. URL https://doi.org/10.1109/CVPR52733.2024.01365.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5228–5238. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00517. URL https://doi.org/10.1109/CVPR52688.2022.00517.

Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. Too Late to Recall: Explaining the Two-Hop Problem in Multimodal Knowledge Retrieval. *CoRR*, abs/2512.03276, December 2025. doi: 10.48550/ARXIV.2512.03276. URL https://doi.org/10.48550/ARXIV.2512.03276.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TrVYEZtSQH.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. MMIE: Massive Multimodal Interleaved Comprehension Benchmark for Large Vision-Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April*

24-28, 2025. OpenReview.net, 2025. URL https://openreview.net/forum?id=HnhNRrLPwm.

Siyu Xu, Yunke Wang, Daochang Liu, Bo Du, and Chang Xu. CollagePrompt: A Benchmark for Budget-Friendly Visual Recognition with GPT-4V. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pp. 6396–6418. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.358. URL https://doi.org/10.18653/V1/2025.FINDINGS-NAACL.358.

Shojiro Yamabe, Futa Waseda, Daiki Shiono, and Tsubasa Takahashi. Text-Printed Image: Bridging the Image-Text Modality Gap for Text-centric Training of Large Vision-Language Models. *CoRR*, abs/2512.03463, December 2025. doi: 10.48550/ARXIV.2512.03463. URL https://doi.org/10.48550/ARXIV.2512.03463.

Zeping Yu and Sophia Ananiadou. Understanding Multimodal LLMs: the Mechanistic Interpretability of LLaVA in Visual Question Answering. *CoRR*, abs/2411.10950, 2024. doi: 10.48550/ARXIV.2411.10950. URL https://doi.org/10.48550/ARXIV.2411.10950.

Z.ai. GLM-4.6V-Flash. Hugging Face model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V-Flash. Accessed September 14, 2026.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19781–19791. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01842. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy M. Hospedales. Benchmarking Multi-Image Understanding in Vision and Language Models: Perception, Knowledge, Reasoning, and Multi-Hop Reasoning. *CoRR*, abs/2406.12742, 2024a. doi: 10.48550/ARXIV.2406.12742. URL https://doi.org/10.48550/ARXIV.2406.12742.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=5KojubHBr8.

Yongshuo Zong, Ondrej Bohdal, and Timothy M. Hospedales. VL-ICL Bench: The Devil in the Details of Multimodal In-Context Learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=cpGPPLLYYx.

Table 8: **Zero-shot Concentric responses.** Each row contains 200 queries, classified exclusively as Option selected or Abstention. Accuracy uses the task parser.

Query	Input	Option selected	Abstention	Acc. (%)
Image	Baseline	200	0	24.0
Image	*Joint*	31	169	3.0
Text	Baseline	200	0	27.5
Text	*Joint*	7	193	2.0

A Response behavior and additional tasks

These studies test how composition affects answer selection, response format, counting, and the use of additional references. *Joint* renders the instruction, references, and query together. *Role* submits the instruction, reference set, and query as separate images. *Example* groups each reference with its label. Their rendering settings are specified for each study; their numerical results are compared only within that study.

A.1 Zero-shot abstention and non-choice responses

The Concentric response studies use Qwen3-VL-8B-Instruct through vLLM, temperature 0, and a 2,048-token output limit. Unless otherwise specified, each condition has 200 queries and one reference per class.

Joint exhibits more abstention behavior than Baseline when reference examples are absent (Table 8). In this zero-shot condition, the reference section is empty while the instruction still requests a match to references a through e (Figure 5). We classify responses into two mutually exclusive categories. Option selected means that the response gives one of the options a through e, including a guess preceded by a statement of uncertainty. Abstention means that the response leaves the option unanswered, either stating that the references are unavailable or describing the query without assigning an option. We checked responses without a parsed option for selections missed by the parser.

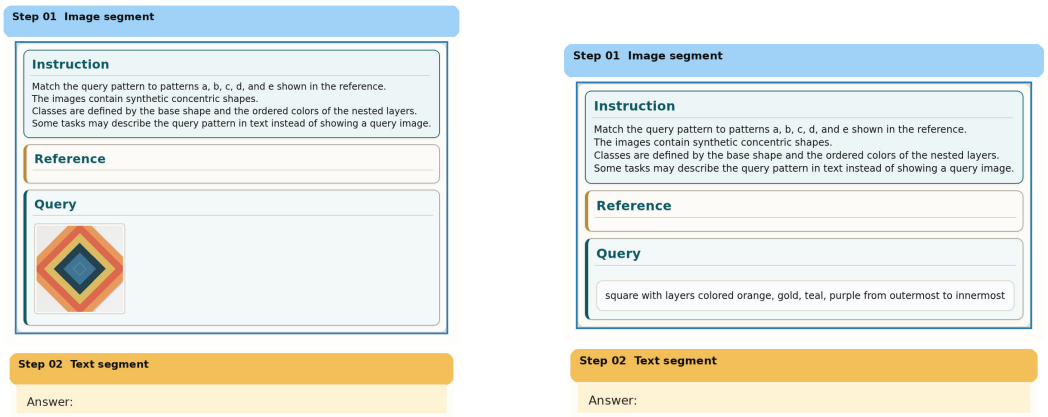


Figure 5: **Zero-shot Concentric inputs.** *Joint* inputs for ImageQuery (left) and TextQuery (right), without references. Numbered headings mark the order of image and external text inputs.

Joint frequently abstains in both query formats, whereas Baseline selects an option in every response. All responses are nonempty and terminate within the output limit. These results describe abstention when the reference-to-label mapping is absent; the experiment does not assess uncertainty calibration.

Table 9: **LocalizedCounting performance.** Accuracy (%) on 200 queries. The difference is *Joint* minus Baseline (pp), with a paired 95% bootstrap interval.

Input	Qwen3-VL-32B
Baseline	42.5
Joint	65.5
Difference [95% CI]	23.0 [14.0, 31.5]

A.2 Reference-conditioned counting

Visual composition also affects a task that requires a numerical count. Each LocalizedCounting query contains multiple synthetic objects and one reference symbol. The target is the number of objects matching the reference's base shape and ordered colors of nested layers. Thus, VLMs must identify matching objects and aggregate their count. The same 200 scenes and references are used in all conditions, with seed 0, temperature 0, a 2,048-token output limit, and one reference. Baseline states the task in text; *Joint* renders the instruction, reference, and query together and supplies "Answer with the number only." outside the image.



Figure 6: **LocalizedCounting input.** *Joint* input, including the external answer instruction. Numbered headings mark input order.

Table 9 shows that *Joint* improves reference-conditioned counting accuracy for Qwen3-VL-32B.

Scoring extracts an integer from the response, prioritizing an answer-marked or standalone number and otherwise taking the final integer, and compares it with the number of matching objects in the scene. Neither condition produces empty responses. Intervals use 10,000 paired query bootstrap draws with seed 0 and are pointwise, without multiplicity correction.

A.3 Following instructions rendered in the image

Rendered instructions can specify both a single-letter response and a structured JSON response. In the image-only instruction conditions, the requested output format appears in the *Joint* instruction and is not repeated as an external format instruction. Every response follows the requested format, but some answers are incorrect (Table 10). Format compliance therefore does not imply task correctness.

No response contains additional content under the scoring rule, which accepts a JSON code fence. In the accompanying conflict conditions, an external JSON instruction overrides a rendered letter

Table 10: **Following image-only format instructions.** Each row contains 200 Concentric queries. JSON requires a single `answer` field containing an option letter; a code fence around that object is accepted.

Query	Format	Adherence (%)	Parse success (%)	Acc. (%)
Image	Letter	100.0	100.0	89.0
Image	JSON	100.0	100.0	88.0
Text	Letter	100.0	100.0	86.0
Text	JSON	100.0	100.0	86.5

Table 11: **Answer selection under conflicting queries.** Each condition uses 200 queries. Matching image and text queries identify the same target; conflicting queries identify different targets.

Queries	Response category	Count	Rate (%)
Matching	Common target	181	90.5
	Other response	19	9.5
Conflicting	Image target	105	52.5
	Text target	59	29.5
	Neither target	36	18.0

instruction in all cases, whereas a rendered JSON instruction is followed in 72% of ImageQuery and 100% of TextQuery cases despite an external letter instruction. The asymmetric result depends on the requested formats.

A.4 Conflicting image and text queries

When the image query and an external textual query identify different reference classes, responses select the image target more often than the text target (Table 11). The reference set and letter-only output instruction remain fixed. The alternative text query describes another class in the same episode.

Among the 36 conflicting-query responses selecting neither target, 30 select other option letters, four complete responses state that no reference matches, and two reach the 2,048-token limit without a parsed option. This result is specific to the tested image queries and textual descriptions; differences in their interpretation may affect the ordering.

A.5 Comparison with image captioning

The evaluated caption-only pipeline has lower accuracy than visual composition on these fine-grained tasks. Qwen3-VL-32B serves as both the fixed captioner, with an 80-word limit per image, and the classifier receiving only those captions. In the paired comparison in Table 12, *Joint* and *Role* outperform caption-only input on all five datasets. All ten paired differences remain significant after Holm correction.

Four caption-only outputs state that no option applies and remain errors. On Food, 398 captions mention a dataset class name and 235 name the depicted image's class. The captioning pipeline can therefore contribute category information through its own recognition step. Its lower downstream accuracy remains specific to this captioner, description budget, and task setting.

A.6 Dependence on the number of references

We compare Baseline and *Example* with one, two, and four references per class on CUB and Flowers using Qwen3-VL-32B and GPT-5.4. All conditions use the same 500 queries per dataset, and larger reference sets include the smaller sets.

Tables 13 and 14 show that the effect of additional references depends on the model and dataset. On Flowers, *Example* is already near the accuracy ceiling at one shot. Qwen3-VL-32B's Baseline catches up as references are added, while GPT-5.4's Baseline declines. On CUB, both configura-

Table 12: **Image inputs versus caption-only inputs.** Accuracy (%) on 500 queries per dataset, with one reference per class. All comparisons use the same queries.

Dataset	Caption only	*Role*	*Joint*
Flowers	78.8	97.2	94.4
CUB	52.0	71.2	70.4
Food	69.4	82.8	79.6
Cars	45.8	69.2	64.2
Dogs	52.4	67.4	66.4

Table 13: **Reference-count scaling with *Example*.** Mean accuracy (%). Differences from Baseline (pp) have pointwise 95% query-cluster bootstrap intervals. Tokens: mean input-token count. B: Baseline; E: *Example*.

Dataset	Shots	B	E	Δ	95% CI	Tokens B	Tokens E
Qwen3-VL-32B							
CUB	1	62.00	74.60	+12.60	[9.60, 15.53]	1,150	1,113
CUB	2	75.00	74.33	-0.67	[-3.07, 1.73]	2,052	1,760
CUB	4	80.73	77.53	-3.20	[-5.60, -0.93]	3,872	3,046
Flowers	1	84.73	97.73	+13.00	[11.07, 15.00]	2,082	1,312
Flowers	2	98.33	98.40	+0.07	[-0.53, 0.60]	3,772	1,956
Flowers	4	98.87	98.60	-0.27	[-0.67, 0.07]	7,119	3,263
GPT-5.4							
CUB	1	72.67	70.67	-2.00	[-4.47, 0.40]	1,471	1,627
CUB	2	72.33	72.00	-0.33	[-2.93, 2.27]	2,633	2,582
CUB	4	76.60	76.87	+0.27	[-2.20, 2.73]	4,979	4,493
Flowers	1	97.00	98.80	+1.80	[0.67, 2.87]	2,547	1,929
Flowers	2	94.60	99.13	+4.53	[3.33, 5.80]	4,619	2,888
Flowers	4	90.67	99.27	+8.60	[7.13, 10.07]	8,738	4,841

tions improve from two to four shots in both models. Qwen3-VL-32B's Baseline has higher mean accuracy at four shots, although this difference does not pass Holm correction; GPT-5.4 shows no significant difference between configurations at any shot count.

At two and four shots, *Example* uses fewer input tokens than Baseline on both datasets. This measures input length, not runtime.

A.7 Splitting a rendered reference panel on CUB

We compare *Role* and *Example* on CUB using Qwen3-VL-32B-Instruct, with 500 shared queries and five-way one-shot episodes. Each instruction, labeled reference panel, and query is rendered once. *Role* receives the reference panel as one image; *Example* receives five strips cut between the labeled examples. Instruction and query images, reference order, and generation settings are identical.

Table 15 shows that splitting the reference panel improves accuracy by 3.00 percentage points. The split preserves all source pixels, including labels, borders, and whitespace. Automatic resizing makes the processed pixel and visual-token counts approximately 1.2% higher for *Example*; image-local positions and encoder computation also change. The result therefore supports splitting a shared rendered panel under standard preprocessing, rather than isolating grouping at identical processed inputs.

Table 14: **Accuracy changes with additional references.** Within-condition changes (pp) with pointwise 95% query-cluster bootstrap intervals. These exploratory comparisons are not adjusted for multiplicity.

Dataset	Construction	Shots	Change	95% CI
		Qwen3-VL-32B		
Flowers	Baseline	$1 \rightarrow 2$	+13.60	[11.67, 15.53]
Flowers	Baseline	$2 \rightarrow 4$	+0.53	[0.07, 1.07]
Flowers	Baseline	$1 \rightarrow 4$	+14.13	[12.27, 16.07]
Flowers	*Example*	$1 \rightarrow 2$	+0.67	[0.13, 1.33]
Flowers	*Example*	$2 \rightarrow 4$	+0.20	[-0.53, 0.93]
Flowers	*Example*	$1 \rightarrow 4$	+0.87	[0.00, 1.80]
CUB	Baseline	$1 \rightarrow 2$	+13.00	[10.40, 15.60]
CUB	Baseline	$2 \rightarrow 4$	+5.73	[3.40, 8.13]
CUB	Baseline	$1 \rightarrow 4$	+18.73	[15.93, 21.47]
CUB	*Example*	$1 \rightarrow 2$	-0.27	[-2.60, 2.13]
CUB	*Example*	$2 \rightarrow 4$	+3.20	[0.60, 5.80]
CUB	*Example*	$1 \rightarrow 4$	+2.93	[0.33, 5.53]
		GPT-5.4		
Flowers	Baseline	$1 \rightarrow 2$	-2.40	[-3.73, -1.07]
Flowers	Baseline	$2 \rightarrow 4$	-3.93	[-5.60, -2.27]
Flowers	Baseline	$1 \rightarrow 4$	-6.33	[-8.00, -4.60]
Flowers	*Example*	$1 \rightarrow 2$	+0.33	[-0.13, 0.87]
Flowers	*Example*	$2 \rightarrow 4$	+0.13	[-0.47, 0.67]
Flowers	*Example*	$1 \rightarrow 4$	+0.47	[-0.20, 1.20]
CUB	Baseline	$1 \rightarrow 2$	-0.33	[-2.93, 2.27]
CUB	Baseline	$2 \rightarrow 4$	+4.27	[1.40, 7.13]
CUB	Baseline	$1 \rightarrow 4$	+3.93	[0.73, 7.07]
CUB	*Example*	$1 \rightarrow 2$	+1.33	[-1.47, 4.20]
CUB	*Example*	$2 \rightarrow 4$	+4.87	[2.20, 7.60]
CUB	*Example*	$1 \rightarrow 4$	+6.20	[3.13, 9.20]

Table 15: **CUB accuracy with a shared reference raster.** Mean accuracy (%); Δ: *Example* minus *Role* (pp). The 95% interval uses query-cluster bootstrap resampling; p is from a paired two-sided sign-flip test.

Role	*Example*	Δ	95% CI	p
70.87	73.87	+3.00	[0.80, 5.20]	0.0090

Table 17: **Main-benchmark inference settings.** Decoding settings are shared across models, tasks, and input configurations. Model revisions are listed in Table 16.

Setting	Value
Output-token limit	2,048
Temperature	Unspecified
Thinking	Off
GPT-5.4	Azure deployment; image detail `original`
Open-weight models	vLLM; model-specific image processors
Image preprocessing	Gemma uses `max_soft_tokens=1120`; other open-weight models use their processor defaults. The separate resolution controls are detailed in Appendix D.1.1.
Support seeds	0, 1, 2 for positive-shot evaluation; 0 for zero-shot evaluation

B Experimental settings and reproducibility

Benchmark subsets. Fine-grained classification uses 500 query images from each of five datasets: Oxford Flowers 102 (Nilsback & Zisserman, 2008), CUB-200-2011 (Wah et al., 2011), Food-101 (Bossard et al., 2014), Stanford Cars (Krause et al., 2013), and Stanford Dogs (Khosla et al., 2011). Episodes have five candidate classes and one reference per class. VL-ICL evaluates Open-MI, CLEVR, and TextOCR with 200 queries each and operator induction with 60, at one, two, and four shots. TrueMICL evaluates classification (50 queries), CLEVR, clock, operator induction, palindrome, shapes count, and sudoku (100 each), at four shots. MuirBench evaluates its 12 categories, totaling 2,600 queries. Fine-grained classification, VL-ICL, and TrueMICL assign equal weight to task cells and use support seeds 0, 1, and 2. MuirBench weights task categories by their eligible query counts.

Models and decoding. Table 16 lists the evaluated model revisions. GPT-5.4 uses the Azure deployment with image detail set to `original`; its deployment name is not an immutable weight hash. All main-benchmark conditions leave temperature unspecified, disable thinking, and permit at most 2,048 generated tokens, with task-specific answer parsing. These decoding settings are shared by Baseline and all VCC configurations. Mechanistic experiments use Hugging Face Transformers, disable thinking, permit at most 16 generated tokens, and score forced-choice candidates a–e. They use Qwen3-VL-8B-Instruct revision `0c351dd01ed87e9c1b53cbc748cba10e6187ff3b`.

Table 16: **Evaluated model revisions.**

Model	Revision
GPT-5.4	`azure-deployment:gpt-5.4`
Qwen3-VL-32B	`0cfaf48183f594c314753d30a4c4974bc75f3ccb`
Qwen3.5-27B	`fc05daec18b0a78c049392ed2e771dde82bdf654`
Gemma4-31B-IT	`842da3794eaa0b77d5f08bae87a17459d91ff475`
GLM-4.6V-Flash	`411bb4d77144a3f03accbf4b780f5acb8b7cde4e`

Rendering. Table 18 specifies the compact label-grid renderer and its task-dependent presets. *Role* exports instruction, reference, and query regions fitted to their content; *Example* additionally exports each labeled reference separately. MuirBench instead exports question, visual-evidence, and answer-choice panels. The Flowers example in Figure 3 uses one *Joint* image, three *Role* images, or seven *Example* images. Baseline retains separate source images and text. The synthetic intervention uses the separate five-column token-grid layout described in Appendix C.3.

Eligible comparisons. Pairwise comparisons use queries with recorded responses in both conditions after provider-policy and input-configuration exclusions. Inputs must match the specified composition and image-detail setting: individual image crops do not constitute *Role* panels, and GPT inputs with image detail `high` are excluded from the `original` condition. Eligibility does

Table 18: **Compact label-grid rendering settings.** Dimensions below are CSS pixels before model preprocessing. The canvas adapts to its content. Presets place labels below their images.

Setting	Value
Font	DejaVu Sans; body 24 px, labels 16 px, titles 32 px
Line height	32 px
Canvas / panel padding	16 px / 16 px
Panel / example gaps	16 px / 16 px
Canvas dimensions	Adaptive width and height; initial browser viewport 1920×1080
Image dimensions	Aspect ratio preserved; reference and query image heights have no fixed cap in the template
Fine-grained	Five reference columns for all configurations
VL-ICL, *Joint*	Two columns for Open-MI; four for CLEVR, TextOCR, and operator induction
TrueMICL, *Joint*	One column for clock, shapes count, and sudoku; four for the other tasks
Role / *Example*	Five-column source layout; content-fitted instruction and query regions. *Example* also fits each reference card to its content.
Query width exception	CLEVR *Joint* for Qwen3-VL-32B and Gemma uses a 40% minimum query-image width; the compact template default is 20%.

Table 19: **Response availability and parsing in the main comparison.** Counts precede paired exclusions. Policy: provider-policy exclusions; empty parse: no answer returned by the task parser; limit: saved output-limit flag.

Benchmark	Records	Policy	Empty parse	Limit
Fine-grained	75,000	6	0	163
VL-ICL	66,000	18	1,663	4,721
TrueMICL	19,500	0	0	53
MuirBench	26,000	36	39	31
Total	186,500	60	1,702	4,968

not require a correctly formatted answer. Empty or malformed parsed answers remain in the accuracy denominator and count as errors, as in the main evaluation.

Response availability and parsing. For the five models and Baseline/VCC configurations in Table 2, including VL-ICL zero-shot inputs, the saved results contain 186,500 records before paired exclusions. Provider content policies exclude 60 records; the remaining 186,440 responses are nonempty, with no recorded request failures or configured skips. The task parsers return nonempty answers for 184,738 responses (99.09%). Empty or malformed parsed answers remain in the accuracy denominator and count as errors. Nonempty parsing does not establish format compliance: in fine-grained classification, 73,598 of 74,994 scored responses (98.14%) yield a single valid option, while 1,396 yield another string. Table 19 gives the breakdown. Token-limit flags can overlap parsing failures; reaching the limit does not itself exclude a response.

Fine-grained baseline prompt. Baseline states the association explicitly: "Each reference image immediately precedes its class label. Match the query to exactly one reference class." The reference images and labels follow in sequence, with the query and the task's answer instruction.

Accuracy aggregation. Baseline supplies images and text separately, with explicit image–label associations for fine-grained classification. Comparisons keep the model and task instance fixed. We report accuracy in percent and differences in percentage points. The accuracy analysis averages paired task/seed/shot cells equally for fine-grained classification, VL-ICL, and TrueMICL. MuirBench averages over eligible query pairs, weighting task categories by their sample counts. Grouping comparisons follow this benchmark-specific accuracy aggregation; the VL-ICL grouping ablation uses shot-averaged accuracy. Each comparison uses queries eligible in both conditions, so its baseline accuracy can differ from a separately aggregated baseline condition. Model-level gains are computed from unrounded values and averaged with equal model weight.

VL-ICL Peak and efficiency. For task t, let $a_t(k)$ be accuracy at k shots, averaged over support seeds. Peak is $\max_k a_t(k)$. Following the full-area normalization in Zong et al. (2025), efficiency is

$$\mathrm{Eff}_t = \frac{1}{K} \int_0^K [a_t(k) - a_t(0)]\, dk, \tag{2}$$

where accuracies are fractions and K is the largest evaluated shot count. We use linear interpolation between evaluated shot counts and multiply both metrics by 100 for reporting. Negative areas are retained. The denominator is the full accuracy range times K, without division by the remaining headroom above zero-shot accuracy. Task metrics are averaged equally; gains are VCC minus Baseline after metric computation. The curves use zero, one, two, and four shots. Zero-shot inputs are evaluated with the same model revisions and input configurations as the few-shot conditions. Each task uses the intersection of eligible queries across both configurations, all shot counts, and support seeds. Zero-shot evaluation uses seed 0; positive-shot accuracies average three support seeds. Pointwise percentile intervals use 10,000 query-cluster bootstrap draws within each task, with shared resampling indices across configurations, shots, and seeds. Peak is recomputed in every draw. These intervals are descriptive and are not multiplicity-adjusted.

Statistics. The accuracy comparisons use query-cluster sign-flip tests and query-cluster bootstrap intervals. VL-ICL statistics in Table 38 refer to shot-averaged accuracy, not the Peak or Eff. columns of Table 2. Resampling preserves repetitions of a query across seeds and shots and recomputes the benchmark-specific macro or micro average. The sign-flip tests use 5,000 Monte Carlo draws with the correction $(b + 1)/(5{,}000 + 1)$, where b counts null statistics at least as extreme as the observed statistic. Accuracy intervals use 5,000 bootstrap draws. Repeated appearances of a query are grouped for uncertainty estimation, so increasing support seeds does not turn those repetitions into independent queries. The main comparisons cluster repetitions by query ID; the transfer analysis in Appendix D.5 additionally joins query IDs that share an image-file hash. Holm correction is applied separately to each benchmark and adopted contrast, across the five evaluated models (five hypotheses per family). The contrasts are *Example* versus Baseline for fine-grained classification and *Role* versus Baseline elsewhere. These are exploratory comparisons of configurations selected using the full benchmark family; the reported p-values do not adjust for that configuration selection. These query-based estimates condition on the evaluated tasks, models, and support-seed schedule. Individual task/seed comparisons retain their paired counts and discordant outcomes in the accompanying table data. Detailed intervals, sample counts, and exclusions are in Appendix D. Pointwise 95% intervals are not simultaneous intervals. A non-significant estimate does not establish equivalence or absence of an effect. Grouping comparisons are descriptive. Mechanistic analyses use a separate Qwen3-VL-8B model and their own matched interventions and scene/query-level uncertainty estimates.

B.1 INPUT BUDGETS AND COMPUTATIONAL COST

Main versus controlled inputs. Table 20 compares the main fine-grained inputs with the image-size and total-pixel controls. Each entry first sums image pixels or input tokens within a request, then averages over queries and the five datasets. The comparison uses one reference per class and the same reference assignment across settings. All conditions share 500 eligible queries per dataset, except GPT-5.4 on Food, where their intersection contains 498 queries. Pixels count the submitted images before model preprocessing, including rendered text and whitespace. Input tokens are reported by the inference service and include visual tokens, external text, and image delimiters.

Table 20: **Input budgets in the main and controlled fine-grained experiments.** Mean total submitted pixels (MP: 10^6 pixels) and total input tokens per request. B: Baseline; R: Role; E: Example. All settings use the same eligible queries and reference assignment; dataset means receive equal weight.

Setting	Submitted pixels (MP)			Input tokens		
	B	R	E	B	R	E
GPT-5.4						
Main	1.723	1.201	1.232	2,135.4	1,626.9	1,780.4
Size control	0.754	1.705	1.388	1,100.1	2,121.8	1,843.7
Pixel control	1.705	1.705	1.705	2,122.2	2,121.8	2,175.2
Qwen3-VL-32B						
Main	1.723	1.201	1.232	1,787.2	1,177.6	1,216.5
Size control	0.754	1.705	1.388	816.2	1,680.8	1,379.6
Pixel control	1.705	1.705	1.705	1,745.1	1,680.8	1,689.0
Qwen3.5-27B						
Main	1.723	1.201	1.232	1,804.2	1,183.6	1,226.5
Size control	0.754	1.705	1.388	833.2	1,687.8	1,390.6
Pixel control	1.705	1.705	1.705	1,762.1	1,687.8	1,700.0
Gemma4-31B-IT						
Main	1.723	1.201	1.232	6,580.3	3,243.3	7,632.9
Size control	0.611	1.356	1.111	348.3	609.7	511.2
Pixel control	1.356	1.356	1.356	671.7	609.7	617.6
GLM-4.6V-Flash						
Main	1.723	1.201	1.232	2,253.7	1,554.2	1,610.7
Size control	0.777	1.674	1.315	1,073.5	2,152.6	1,703.1
Pixel control	1.674	1.674	1.674	2,216.9	2,152.6	2,160.6

The two controls constrain different quantities. Matching source-image sizes leaves room for additional rendered text and whitespace, so total pixels and input tokens can increase. Matching total pixels brings image budgets together but does not match input-token counts exactly. For Qwen3-VL-32B, *Example* changes input tokens relative to Baseline by -31.9% in the main setting, $+69.0\%$ under the size control, and -3.2% under the pixel control. Gemma's controls disable automatic resizing and use a common image scale, as described in Appendix D.1.1; its large token-count change across settings includes this preprocessing change.

Token usage across benchmarks. Table 21 reports input and output tokens for the adopted main-benchmark configurations. Means use paired eligible responses and the benchmark's accuracy-aggregation weights; incorrect and output-limited answers remain included. VL-ICL uses the common one-, two-, and four-shot settings. Token counts are compared within each model because tokenization differs across models.

Table 21: **Input and output token usage.** Mean tokens per response for Baseline (B) and the adopted VCC configuration (V). Δ is the relative change (%). VL-ICL averages one, two, and four shots.

	Input tokens			Output tokens		
Model	B	V	Δ	B	V	Δ
Fine-grained (Example)						
GPT-5.4	2,135.4	1,780.4	-16.6	5.0	5.0	0.0
Qwen3-VL-32B	1,787.2	1,216.5	-31.9	2.0	2.1	+3.5
Qwen3.5-27B	1,804.2	1,226.5	-32.0	162.4	207.9	+28.0
Gemma4-31B-IT	6,580.3	7,632.9	+16.0	4.8	2.2	-54.6
GLM-4.6V-Flash	2,253.7	1,610.7	-28.5	71.5	95.7	+33.8
VL-ICL (Role)						
GPT-5.4	1,463.6	1,588.4	+8.5	23.6	23.9	+1.3
Qwen3-VL-32B	1,241.7	1,202.5	-3.2	108.6	151.3	+39.3
Qwen3.5-27B	1,254.7	1,206.0	-3.9	1,520.5	854.9	-43.8
Gemma4-31B-IT	4,308.8	2,718.1	-36.9	27.5	44.6	+61.9
GLM-4.6V-Flash	1,546.4	1,578.6	+2.1	49.1	57.9	+17.9
TrueMICL (Role)						
GPT-5.4	1,440.7	1,953.1	+35.6	5.5	5.5	+0.1
Qwen3-VL-32B	1,124.2	1,463.3	+30.2	2.9	3.0	+2.3
Qwen3.5-27B	1,143.2	1,469.3	+28.5	93.7	125.5	+33.9
Gemma4-31B-IT	5,550.2	3,256.6	-41.3	2.8	2.8	+0.5
GLM-4.6V-Flash	1,401.1	1,941.0	+38.5	4.2	4.5	+6.8
MuirBench (Role)						
GPT-5.4	3,542.9	1,794.6	-49.3	5.0	5.0	0.0
Qwen3-VL-32B	3,627.3	1,422.2	-60.8	2.0	2.0	0.0
Qwen3.5-27B	3,636.5	1,430.2	-60.7	165.1	223.6	+35.4
Gemma4-31B-IT	4,810.3	3,278.5	-31.8	130.3	115.4	-11.5
GLM-4.6V-Flash	4,379.6	1,849.3	-57.8	27.1	39.7	+46.3

Input-token changes depend on both model and task. MuirBench inputs shrink for all five models, whereas TrueMICL inputs grow for four. Output lengths can change in the opposite direction: fewer input tokens therefore need not imply a lower total token charge, which also depends on input and output pricing.

Relation to computation. VCC adds image rendering and can change both the number of images and the extent of attention within each image. Equal total pixels or visual-token counts do not imply equal encoder computation; padding, image boundaries, and preprocessing also matter. The tables quantify input budgets and generated lengths, not FLOPs, peak memory, or a matched latency comparison. Timing records from runs with different serving and concurrency conditions do not establish a speed advantage.

Table 22: **Rendering and grouping on Concentric (original configurations).** Qwen3-VL-32B, one-shot, 200 queries and three support seeds. Base.: accuracy (%); other columns: changes (pp). $I/R/Q$: instruction/references/query; bars separate images. Unlisted roles retain baseline inputs. Rendered pixels are not held fixed across configurations.

Query	Base.	I	R	Q	$I\|R$	IR	$I\|Q$	IQ	$R\|Q$	RQ	$I\|R\|Q$	$IR\|Q$	$I\|RQ$	$IQ\|R$	IRQ
Image	61.0	+18.2	+25.5	−9.7	+31.8	+33.0	+15.5	+28.0	+29.0	+34.5	+34.5	+34.0	+36.5	+36.5	+35.5
Text	69.0	−2.3	+21.8	−20.8	+19.3	+18.5	+9.0	+3.3	+23.7	+23.8	+24.0	+24.0	+25.0	+22.2	+29.3

C Evidence supporting the main-text claims

C.1 Query rendering depends on reference composition

Rendering a query has different effects depending on whether the references are also rendered. We report all role subsets and partitions to test this dependence.

For this analysis, I, R, Q denote the instruction, reference set (images and labels), and query. Concatenation denotes a shared image, and a vertical bar denotes separate images: IRQ uses one image, $I|R|Q$ separates the three roles, and $I|RQ$ places references and query together. Unlisted roles remain as separate images or decoder text. We enumerate each subset of roles to render and each partition of the selected roles: one empty subset, three single-role conditions, six two-role conditions, and five three-role conditions, giving 15 configurations. The empty subset (empty) retains the baseline source images and separate text. Qwen3-VL-32B evaluates all configurations at one and two shots; GPT-5.4 evaluates six at two shots. Rendering an image query here means placing the image and its associated text in the query template. These synthetic partitions use a different renderer from the compact real-task configurations. Configurations that render I also expand the instruction to describe the synthetic shapes and their class-defining shape and color sequence. The original 15-configuration comparison therefore includes changes in instruction wording; the fixed-rendering comparison in Table 6 uses a common instruction across conditions. The query-rendering contrasts below keep I unrendered.

The analysis includes 216 runs and 43,200 outputs, aggregated into 72 cells of 600 outputs. All condition pairs use the same query identities and support-seed schedule. The baseline and rendered query conditions include the same task content; composition changes its presentation. The two query formats have different baseline difficulty, so their absolute accuracies are not a controlled comparison of query format alone. Rendering the query alone lowers Qwen3-VL-32B accuracy in both query formats at both shot counts ($\mathrm{empty} \rightarrow Q$). Adding query rendering when the references are already rendered improves accuracy in all four settings ($R \rightarrow R|Q$). The effect of query rendering therefore depends on the reference configuration. These are descriptive comparisons; no interaction significance test is reported.

Table 22 summarizes the one-shot Qwen3-VL-32B rendering and grouping comparisons. Rendering references improves accuracy over Baseline in both query formats. These are the original configurations, with varying rendered pixels; they are distinct from the fixed-pixel comparisons in Table 6.

C.2 Grouping with matched pixels and visual tokens

We compare all five partitions of the instruction (I), reference set (R), and query (Q): IRQ, $IR|Q$, $I|RQ$, $IQ|R$, and $I|R|Q$. We also include $I|Q|R$ as an order-matched split control for $IQ|R$. All conditions use Qwen3-VL-32B-Instruct, the same 200 queries per query format as the composition analysis, one reference per class, and support seeds 0, 1, and 2. The 1,200 episodes yield 7,200 outputs across six conditions. Reference images and labels remain together within R throughout.

Matched inputs. Each pair starts from the original IRQ canvas. Missing historical image files are restored from the saved episodes and original rendering configuration; the restoration reproduces the RGB pixels of a retained source in each query-format/seed cell. A minimal top padding of 0–31 pixels aligns cuts inside both inter-panel gaps; right and bottom padding then align the canvas dimensions to the 32-pixel visual-token grid. This translates the entire source uniformly without

Table 23: **Composition results for Qwen3-VL-32B-Instruct.** Accuracy (%) over 600 outputs per cell (three support seeds). Bold marks all tied maxima. Unselected roles retain baseline inputs.

Scope/partition	ImageQuery 1-shot	ImageQuery 2-shot	TextQuery 1-shot	TextQuery 2-shot
empty	61.00	84.67	69.00	89.00
I	79.17	90.50	66.67	83.83
R	86.50	97.00	90.83	93.17
Q	51.33	78.33	48.17	68.00
$I\|R$	92.83	97.17	88.33	89.50
IR	94.00	98.00	87.50	89.83
$I\|Q$	76.50	90.33	78.00	87.17
IQ	89.00	89.50	72.33	70.17
$R\|Q$	90.00	98.50	92.67	94.83
RQ	95.50	**99.50**	92.83	94.83
$I\|R\|Q$	95.50	99.00	93.00	93.83
$IR\|Q$	95.00	98.33	93.00	94.50
$I\|RQ$	**97.50**	**99.50**	94.00	94.50
$IQ\|R$	**97.50**	**99.50**	91.17	90.83
IRQ	96.50	98.67	**98.33**	**99.00**

Table 24: **Composition results for GPT-5.4.** Accuracy (%) over 600 outputs per cell (three support seeds). Bold marks all tied maxima. Unselected roles retain baseline inputs.

Scope/partition	ImageQuery 2-shot	TextQuery 2-shot
empty	28.17	91.67
R	91.67	99.33
$I\|R\|Q$	98.17	**100.00**
$IR\|Q$	97.00	**100.00**
$I\|RQ$	**98.33**	**100.00**
IRQ	97.33	**100.00**

resizing content or changing relative positions. Two cuts in the gaps between role panels produce three whole-width strips in I–R–Q order. All source pixels, including the background, shadows, and whitespace, are retained. Reassembling the strips exactly reconstructs the padded canvas for every episode. The submitted images remain within the processor's pixel bounds without further resizing. Actual processor patch tensors for one scene in each of the six query-format/seed cells, plus nonzero-top-padding cases where present, are identical between the single image and concatenated strips. Both conditions use the same external text, "Answer:"; the task instruction appears in the image.

Partial grouping and order. The $IR|Q$ and $I|RQ$ conditions concatenate adjacent strips without changing their pixels or order. For $IQ|R$, we concatenate the original I and Q strips and submit R second. Because this changes the sequence to I–Q–R, its grouping contrast uses $I|Q|R$ as the control. We separately compare $I|Q|R$ with $I|R|Q$ to measure the order effect. Every concatenation is checked for exact RGB identity with its source strips. Actual processor patch tensors also match exactly for all three joined pairs in each query-format/seed cell (18 checks). No content is resized, rerendered, or discarded. For the nonadjacent IQ pair, joining the strips changes their relative placement within an image; the order-matched contrast controls their sequence while image-local positions change with grouping.

Inference and statistics. All conditions use model revision `0cfaf48183f5`, vLLM, non-thinking mode, temperature 0, sampling seed 0, and a 2,048-token output limit. The task parser matches the composition analysis; all 7,200 answers are valid and finish without truncation or transport retries. Accuracy averages queries and support seeds equally. The four grouping contrasts per format are $IR|Q - I|R|Q$, $I|RQ - I|R|Q$, $IQ|R - I|Q|R$, and $IRQ - I|R|Q$. For each contrast, 10,000 bootstrap draws resample query clusters, retaining all three support seeds together. Two-sided sign-flip tests use 100,000 draws over the same clusters with finite-sample correction. Holm adjustment covers all eight grouping contrasts; the two order contrasts form a separate family. Both

Table 25: **Grouping with fixed pixels and visual-token counts.** Qwen3-VL-32B, one-shot. Accuracy (%) averages 200 queries and three support seeds per query format. I|Q|R controls input order for IQ|R.

Partition	ImageQuery	TextQuery
$I\|R\|Q$	96.83	96.17
$IR\|Q$	96.67	98.50
$I\|RQ$	97.00	93.00
$IQ\|R$	98.67	97.83
IRQ	97.50	96.67
$I\|Q\|R$	98.33	97.67

Table 26: **Paired effects of grouping and input order.** Effects are in pp. Query-cluster bootstrap intervals retain all three support seeds. Holm correction covers eight grouping contrasts; the two order contrasts form a separate family.

Query	Contrast	Δ	95% CI	Holm p
ImageQuery	$IR\|Q - I\|R\|Q$	-0.17	[-2.33, 1.67]	1.0000
ImageQuery	$I\|RQ - I\|R\|Q$	+0.17	[-2.33, 2.50]	1.0000
ImageQuery	$IQ\|R - I\|Q\|R$	+0.33	[-0.83, 1.67]	1.0000
ImageQuery	$IRQ - I\|R\|Q$	+0.67	[-1.83, 3.17]	1.0000
ImageQuery	$I\|Q\|R - I\|R\|Q$	+1.50	[0.17, 3.17]	0.2527
TextQuery	$IR\|Q - I\|R\|Q$	+2.33	[1.00, 3.83]	0.0258
TextQuery	$I\|RQ - I\|R\|Q$	-3.17	[-5.33, -1.17]	0.0258
TextQuery	$IQ\|R - I\|Q\|R$	+0.17	[-1.00, 1.17]	1.0000
TextQuery	$IRQ - I\|R\|Q$	+0.50	[-0.83, 2.00]	1.0000
TextQuery	$I\|Q\|R - I\|R\|Q$	+1.50	[-0.33, 3.33]	0.2527

procedures use seed 20260917. The initial full-canvas contrast was completed before the partial-grouping extension; all extension contrasts and the expanded correction families were fixed before its inference. The original 15-configuration table retains its original inputs and is separate from this matched comparison.

Results. Tables 25 and 26 give absolute accuracies and all paired contrasts, including the grouping effects reported in Table 6. The positive $IR|Q$ and negative $I|RQ$ effects on TextQuery have the same direction in all three support seeds. Neither order contrast is significant after Holm correction. The comparison preserves image–label associations inside R and does not test splitting them.

Input budgets. Mean visual-token counts are 4,620 for ImageQuery and 4,589.7 for TextQuery in every condition; total pixels likewise match exactly for each episode. The total decoder input adds 13, 15, or 17 tokens for one, two, or three submitted images, respectively, including the common external text and template. All 7,200 API input-token counts match these values. Image-local positional coordinates and the extent of within-image visual computation change with grouping. Equal visual-token counts therefore do not imply equal encoder computation.

C.3 Observed regions contribute through retained tokens

We extend the matched token-removal analysis in Section 4.5 to all evaluated regions and query formats. Alongside Full→Drop and Mask→Drop, we report Full and Mask, which retain all tokens from the complete and masked inputs, respectively.

Qwen3-VL supplies the decoder with final vision-merger features and three intermediate visual feature paths through DeepStack (Meng et al., 2024; Qwen Team, 2025). The intervention checks cover the visual paths as well as retained token positions.

Each accuracy pools 200 queries at one and two shots with support seeds 0, 1, and 2, giving 1,200 outputs per condition. Table 27 reports pointwise 95% query-cluster bootstrap intervals for Full→Drop minus Mask→Drop. Repeated appearances of a query across shots and seeds are clus-

tered together. The intervals are not multiplicity-adjusted. The table covers eight combinations of region and query format; query-text masking is included only for TextQuery, where the text specifies the query.

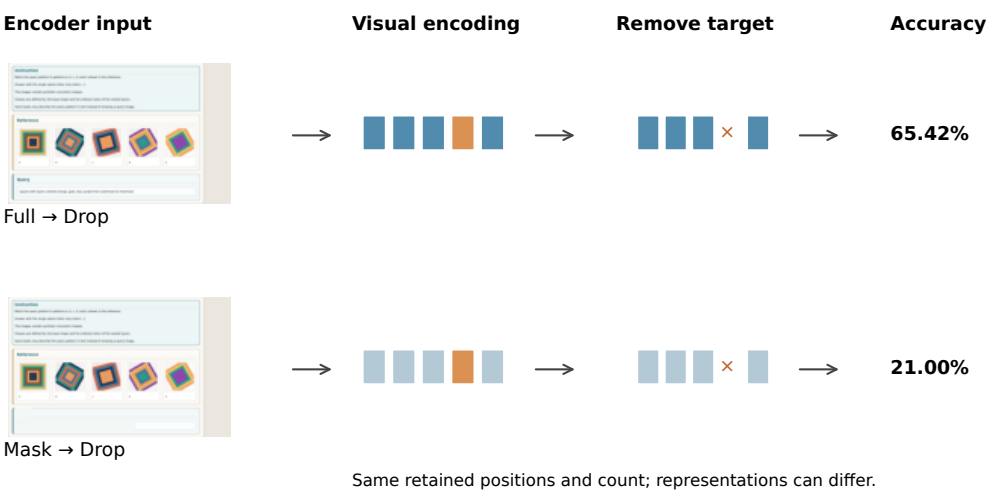


Figure 7: **Region masking and token removal.** *Joint* Concentric TextQuery, Qwen3-VL-8B, 1,200 outputs per condition. Masking before encoding and matched token removal test the benefit of observing query text through retained representations.

Table 27: **Region masking and token removal by target region.** *Joint* accuracy (%) over 1,200 outputs per condition. Query-text masking applies only to TextQuery. F→D and M→D denote Full/Mask followed by Drop; Δ_T is Full→Drop minus Mask→Drop (pp). Intervals are pointwise 95% query-cluster bootstrap intervals for Δ_T, without multiplicity adjustment.

Target	Full	Mask	F→D	M→D	Δ_T	95% CI
ImageQuery / class label text	89.50	89.00	86.25	85.00	1.25	[-1.25, 3.76]
ImageQuery / instruction text	89.50	88.50	89.25	86.00	3.25	[0.50, 6.25]
ImageQuery / query image	89.50	21.00	27.25	18.00	9.25	[5.00, 13.50]
ImageQuery / reference image	89.50	18.75	30.00	18.25	11.75	[5.50, 18.00]
TextQuery / class label text	84.00	81.83	73.50	68.33	5.17	[2.92, 7.42]
TextQuery / instruction text	84.00	78.83	73.67	69.67	4.00	[0.92, 7.17]
TextQuery / query text	84.00	23.08	65.42	21.00	44.42	[39.50, 49.25]
TextQuery / reference image	84.00	20.33	26.33	21.25	5.08	[1.67, 8.58]

Table 28: **Token-removal effects under *Joint* and *Role*.** Accuracy (%) and drop contrasts (pp), with 1,200 outputs per condition. F and M denote Full and Mask followed by Drop.

Target	VCC-*Joint*			VCC-*Role*		
	F	M	Δ	F	M	Δ
ImageQuery / query image	27.25	18.00	9.25	85.50	21.50	64.00
ImageQuery / reference image	30.00	18.25	11.75	31.75	22.25	9.50
TextQuery / instruction text	73.67	69.67	4.00	85.75	76.67	9.08
TextQuery / query text	65.42	21.00	44.42	83.75	22.83	60.92
TextQuery / reference image	26.33	21.25	5.08	34.00	21.58	12.42

Table 28 checks whether the retained-token benefit also occurs when the instruction, references, and query occupy different images. The query-image contrast is positive for both *Joint* and *Role*. It does not locate reference–query matching or establish that the same mechanism explains the two configurations.

The five *Role* conditions form a selected subset and are not an exhaustive partition comparison. Interventions preserve the same retained positions across all visual feature paths, with disjoint target and retained regions.

C.4 Reference-image representations contain assigned labels

This section details the probe training in Section 4.5 and reports the additional TextQuery results.

Both probes use the same *Joint* two-shot inputs and final vision-merger features from Qwen3-VL-8B. The projection in Figure 4 is this model component; no additional feature projection is applied. Each query format uses 80 training, 20 validation, and 40 test scenes, with five label permutations per scene. The synthetic classes, defined by their outer-to-inner color sequences, are disjoint across splits. The answer labels a through e are shared across all splits and permuted among the five classes within each scene. Each reference image contributes the mean of 72 tokens; the query contributes the mean of 100 tokens for ImageQuery or 132 for TextQuery. Regions are selected geometrically, excluding label-overlapping tokens and input patches changed by label permutation. No positions are selected by readout accuracy.

The probes use identical train-only centering and scalar RMS normalization, with linear classifiers fit by LBFGS and L2 chosen from $\{0.0001, 0.01, 1\}$ on validation data. Each input provides ten reference vectors and one query vector, giving 4,000 reference and 400 query training samples. Reference targets are the assigned labels; query targets are the correct answers. All target labels are balanced, with 20% chance accuracy. The test set is unused during selection, and no decoder forward pass is required.

Figure 4 reports the ImageQuery readout accuracies and uncertainty intervals. For TextQuery, reference-label accuracy is 88.75% (95% CI [87.05, 90.30]) and query-answer accuracy is 18.50% (95% CI [15.50, 22.00]). Intervals use 10,000 scene resamples, averaging label permutations and reference slots within each scene. The reference-label and query-answer targets differ, so these measurements compare the specified linear readouts rather than the total information in each region.

D FULL RESULTS BY BENCHMARK

Each subsection groups the configuration comparison, paired uncertainty estimates, and task-level results for one benchmark. Task tables retain all reported models and shot settings, including regressions. Fine-grained and TrueMICL use one and four shots, respectively; their fixed shot counts are stated in captions. Condition means and paired-comparison means use their stated sets of eligible queries and may differ.

D.1 FINE-GRAINED

Table 29: **Grouping results for fine-grained classification.** One-shot accuracy (%).

Model	Baseline	*Joint*	*Role*	*Example*
GPT-5.4	81.40	81.07	78.59	83.97
Qwen3-VL-32B	65.12	75.09	67.37	78.76
Qwen3.5-27B	58.93	78.45	62.20	67.37
Gemma4-31B-IT	79.60	74.52	76.03	81.08
GLM-4.6V-Flash	73.32	65.40	58.73	74.04

Table 30: **Uncertainty in fine-grained classification gains.** Paired gains for *Example* minus Baseline (pp), pointwise 95% intervals, and Holm-adjusted p values. The *Queries* column counts query clusters.

Model	Δ	95% CI	p	Queries
GPT-5.4	+2.57	[1.61, 3.49]	0.0010	2498
Qwen3-VL-32B	+13.64	[12.43, 14.91]	0.0010	2500
Qwen3.5-27B	+8.44	[6.91, 10.01]	0.0010	2500
Gemma4-31B-IT	+1.48	[0.53, 2.45]	0.0060	2500
GLM-4.6V-Flash	+0.72	[-0.39, 1.80]	0.1940	2500

Table 31: **Fine-grained classification results by dataset.** One-shot accuracy (%) and gains for *Example* minus Baseline (pp), averaged over paired support-seed cells.

Task	Model	Baseline	VCC	Δ
102Flowers	Gemma4-31B-IT	94.40	95.07	0.67
	GLM-4.6V-Flash	93.27	92.73	-0.53
	GPT-5.4	97.00	98.80	1.80
	Qwen3-VL-32B	84.73	97.73	13.00
	Qwen3.5-27B	74.67	91.67	17.00
CUB-Birds	Gemma4-31B-IT	72.87	73.93	1.07
	GLM-4.6V-Flash	61.40	62.27	0.87
	GPT-5.4	72.67	70.67	-2.00
	Qwen3-VL-32B	62.00	74.60	12.60
	Qwen3.5-27B	48.80	62.47	13.67
Food-101	Gemma4-31B-IT	86.53	83.27	-3.27
	GLM-4.6V-Flash	76.73	79.13	2.40
	GPT-5.4	86.21	86.81	0.60
	Qwen3-VL-32B	60.93	81.93	21.00
	Qwen3.5-27B	70.00	61.00	-9.00
Stanford Cars	Gemma4-31B-IT	78.47	83.33	4.87
	GLM-4.6V-Flash	73.53	71.53	-2.00
	GPT-5.4	78.53	87.27	8.73
	Qwen3-VL-32B	63.47	70.27	6.80
	Qwen3.5-27B	59.80	59.07	-0.73

Task	Model	Baseline	VCC	Δ
Stanford Dogs	Gemma4-31B-IT	65.73	69.80	4.07
	GLM-4.6V-Flash	61.67	64.53	2.87
	GPT-5.4	72.60	76.33	3.73
	Qwen3-VL-32B	54.47	69.27	14.80
	Qwen3.5-27B	41.40	62.67	21.27

D.1.1 IMAGE-SIZE AND TOTAL-PIXEL CONTROLS

Table 4 averages the controls over five VLMs and five datasets with equal weights. The results below report each VLM separately.

Inputs and matched quantities. The controls use the main evaluation's *Role*/*Example* layouts for five fine-grained datasets, with 500 queries per dataset, one reference per class, and seed 0. Query identities, references, and assigned labels match the main evaluation. In the image-size control, the reference and query image-content crops are pixel-identical across configurations within each VLM. Baseline submits six individual images with the baseline task prompt; *Role* submits three role panels; *Example* submits seven images. White padding on the right and bottom aligns dimensions to the processor grid: 32 pixels for the Qwen VLMs, 28 for GLM-4.6V-Flash, and 48 for Gemma4-31B-IT. Processor audits confirm that submitted dimensions are preserved for the local VLMs. GPT-5.4 uses the Qwen inputs with `original` image detail. Image count, rendered text, and content positions relative to the token grid remain different.

For Gemma4-31B-IT, all configurations use a common $7/8$ scale before grid padding. Automatic image resizing is disabled to preserve the matched content sizes, and the vision processor retains its 1,120-token padding capacity per image. A vLLM adapter sets image-placeholder counts to the processor's returned token counts without changing the vision encoder or its weights. These settings depart from the automatic resizing in Gemma's standard preprocessing pipeline. This departure may contribute to the observed accuracy decreases. Since preprocessing and input configuration have not been varied independently, these results do not identify the cause of the decreases.

The total-pixel control uses each query's padded *Role* input as the total-pixel target. Baseline and *Example* images are resized by an approximately common factor per input, followed by grid rounding and edge adjustments to keep total pixels within 1% of the target. Content aspect ratios are preserved with residual white padding. This matches the submitted budget, including whitespace and rendered text, while allowing content resolution and letter sizes to differ. Table 34 separates valid visual tokens supplied to the decoder from total input tokens. Gemma's encoder also processes padded positions, so its valid-token count does not measure encoder computation. Matching pixels does not equate information content or the length of external text.

Inference and response validity. All five conditions use the revisions in Table 16 and a 2,048-token output limit. Local VLMs use vLLM with temperature 0 and seed 0; GPT-5.4 uses the API. *Role* is shared between the two series, giving 12,500 evaluations per VLM before exclusions. Gemma uses four-way tensor parallelism for 2,457 evaluations and two-way tensor parallelism for the remaining 10,043. Inputs, weights, vision padding, and decoding settings are unchanged across this transition. A 26-input topology check yielded one differing answer; the saved results retain both execution configurations.

The response parser matches the main evaluation. Invalid answers remain incorrect: Qwen3.5-27B produces 1,224 invalid answers, GLM-4.6V-Flash eight, Gemma4-31B-IT four, and Qwen3-VL-32B none. Qwen3.5-27B reaches the output limit on 141 responses; all responses from the other local VLMs terminate before this limit. For GPT-5.4, three requests for the same Food-101 query are rejected by the provider. We exclude that query from all five conditions, leaving 499 paired queries for Food-101 and 500 for each other dataset. Scoring the rejected requests as incorrect on the full query set preserves the direction of the mean gains.

Paired comparisons. Each series reports *Role*–Baseline, *Example*–Baseline, and *Example*–*Role*. Exact McNemar tests receive Holm correction over 15 comparisons per VLM and control series.

Pointwise paired bootstrap intervals use 10,000 draws and seed 20260914, with resampling within datasets; mean differences weight datasets equally. These intervals are not adjusted for multiple comparisons. The shared *Role* condition means the two series are not independent replications. Tables 32–34 report accuracies, mean intervals, and budgets; Tables 35 and 36 report all paired tests.

Both configurations improve for Qwen3-VL-32B and GPT-5.4 under both controls, with mean-difference intervals above zero. For Qwen3.5-27B, only *Role* meets this criterion. GLM-4.6V-Flash has intervals spanning zero for both configurations relative to Baseline. Gemma4-31B-IT improves with *Role* under image-size control, but both configurations decrease under total-pixel control; *Example* also decreases under image-size control. Thus, positive model averages coexist with decreases in individual VLMs.

Table 32: **Control accuracies by model and dataset.** Accuracy (%). B: Baseline; E: *Example*; n: paired queries. *Role* is shared between controls.

Model	Dataset	n	B (size)	*Role*	E (size)	B (pixels)	E (pixels)
GPT-5.4	Flowers	500	96.60	98.60	98.80	90.60	99.20
GPT-5.4	CUB	500	66.40	69.60	72.00	70.40	73.80
GPT-5.4	Food	499	83.17	82.97	85.97	80.36	85.57
GPT-5.4	Cars	500	81.60	84.00	88.60	80.00	88.00
GPT-5.4	Dogs	500	70.20	75.40	75.40	70.80	76.80
Qwen3-VL-32B	Flowers	500	86.40	95.20	96.20	84.00	97.40
Qwen3-VL-32B	CUB	500	59.80	71.20	74.00	62.40	74.80
Qwen3-VL-32B	Food	500	65.20	81.80	81.40	65.80	80.80
Qwen3-VL-32B	Cars	500	59.80	68.00	71.60	56.60	72.40
Qwen3-VL-32B	Dogs	500	57.00	69.60	71.20	57.00	70.60
Qwen3.5-27B	Flowers	500	81.00	91.00	84.60	79.40	87.80
Qwen3.5-27B	CUB	500	48.80	64.60	54.60	43.80	60.80
Qwen3.5-27B	Food	500	69.80	80.00	56.20	75.20	58.60
Qwen3.5-27B	Cars	500	63.80	67.40	50.80	67.80	58.00
Qwen3.5-27B	Dogs	500	41.60	67.60	58.80	44.00	56.00
Gemma4-31B-IT	Flowers	500	87.80	89.20	89.00	92.80	91.80
Gemma4-31B-IT	CUB	500	56.20	58.60	51.20	65.00	56.60
Gemma4-31B-IT	Food	500	75.20	79.60	70.40	82.80	76.60
Gemma4-31B-IT	Cars	500	58.00	62.80	55.40	72.20	59.20
Gemma4-31B-IT	Dogs	500	58.20	62.80	52.40	69.60	62.00
GLM-4.6V-Flash	Flowers	500	94.00	91.60	92.80	94.20	93.40
GLM-4.6V-Flash	CUB	500	60.80	64.00	60.20	62.60	62.80
GLM-4.6V-Flash	Food	500	74.00	77.80	77.00	73.80	78.40
GLM-4.6V-Flash	Cars	500	73.40	62.60	73.00	74.00	72.40
GLM-4.6V-Flash	Dogs	500	63.20	62.80	61.60	60.80	65.00

Table 33: **Mean control differences by model.** Differences (pp) weight datasets equally. Intervals are pointwise 95% paired bootstrap intervals. B: Baseline; R: *Role*; E: *Example*.

Model	Control	Pair	Δ	95% CI
GPT-5.4	Size	R–B	+2.52	[1.08, 4.00]
GPT-5.4	Size	E–B	+4.56	[3.20, 5.96]
GPT-5.4	Size	E–R	+2.04	[0.88, 3.24]
GPT-5.4	Pixels	R–B	+3.68	[2.08, 5.32]
GPT-5.4	Pixels	E–B	+6.24	[4.76, 7.76]
GPT-5.4	Pixels	E–R	+2.56	[1.32, 3.80]
Qwen3-VL-32B	Size	R–B	+11.52	[9.56, 13.52]
Qwen3-VL-32B	Size	E–B	+13.24	[11.36, 15.20]
Qwen3-VL-32B	Size	E–R	+1.72	[0.24, 3.20]
Qwen3-VL-32B	Pixels	R–B	+12.00	[10.00, 14.00]
Qwen3-VL-32B	Pixels	E–B	+14.04	[12.08, 15.96]
Qwen3-VL-32B	Pixels	E–R	+2.04	[0.60, 3.52]
Qwen3.5-27B	Size	R–B	+13.12	[11.00, 15.28]

Model	Control	Pair	Δ	95% CI
Qwen3.5-27B	Size	E–B	-0.00	[-2.36, 2.40]
Qwen3.5-27B	Size	E–R	-13.12	[-15.12, -11.12]
Qwen3.5-27B	Pixels	R–B	+12.08	[9.96, 14.20]
Qwen3.5-27B	Pixels	E–B	+2.20	[-0.12, 4.52]
Qwen3.5-27B	Pixels	E–R	-9.88	[-11.92, -7.88]
Gemma4-31B-IT	Size	R–B	+3.52	[1.64, 5.40]
Gemma4-31B-IT	Size	E–B	-3.40	[-5.32, -1.48]
Gemma4-31B-IT	Size	E–R	-6.92	[-8.88, -5.00]
Gemma4-31B-IT	Pixels	R–B	-5.88	[-7.60, -4.08]
Gemma4-31B-IT	Pixels	E–B	-7.24	[-8.92, -5.52]
Gemma4-31B-IT	Pixels	E–R	-1.36	[-3.16, 0.44]
GLM-4.6V-Flash	Size	R–B	-1.32	[-3.24, 0.56]
GLM-4.6V-Flash	Size	E–B	-0.16	[-1.76, 1.44]
GLM-4.6V-Flash	Size	E–R	+1.16	[-0.72, 3.04]
GLM-4.6V-Flash	Pixels	R–B	-1.32	[-3.24, 0.60]
GLM-4.6V-Flash	Pixels	E–B	+1.32	[-0.32, 3.00]
GLM-4.6V-Flash	Pixels	E–R	+2.64	[0.80, 4.48]

Table 34: **Pixel and token counts under the controls.** Dataset means receive equal weight. Pixels include rendered text and whitespace; input tokens include external text. GPT-5.4 reports total input tokens but does not expose visual-token counts.

Model	Condition	Pixels	Visual tokens	Input tokens
GPT-5.4	B (size)	753,866.03	–	1,100.06
GPT-5.4	*Role*	1,704,831.33	–	2,121.79
GPT-5.4	E (size)	1,388,120.00	–	1,843.70
GPT-5.4	B (pixels)	1,705,118.31	–	2,122.23
GPT-5.4	E (pixels)	1,705,046.48	–	2,175.24
Qwen3-VL-32B	B (size)	753,842.59	736.17	816.17
Qwen3-VL-32B	*Role*	1,704,790.43	1,664.83	1,680.83
Qwen3-VL-32B	E (size)	1,388,095.49	1,355.56	1,379.56
Qwen3-VL-32B	B (pixels)	1,705,077.15	1,665.11	1,745.11
Qwen3-VL-32B	E (pixels)	1,705,005.06	1,665.04	1,689.04
Qwen3.5-27B	B (size)	753,842.59	736.17	833.17
Qwen3.5-27B	*Role*	1,704,790.43	1,664.83	1,687.83
Qwen3.5-27B	E (size)	1,388,095.49	1,355.56	1,390.56
Qwen3.5-27B	B (pixels)	1,705,077.15	1,665.11	1,762.11
Qwen3.5-27B	E (pixels)	1,705,005.06	1,665.04	1,700.04
Gemma4-31B-IT	B (size)	611,292.67	265.32	348.32
Gemma4-31B-IT	*Role*	1,356,462.49	588.74	609.74
Gemma4-31B-IT	E (size)	1,110,998.02	482.20	511.20
Gemma4-31B-IT	B (pixels)	1,356,370.33	588.70	671.70
Gemma4-31B-IT	E (pixels)	1,356,163.89	588.61	617.61
GLM-4.6V-Flash	B (size)	777,317.18	991.48	1,073.48
GLM-4.6V-Flash	*Role*	1,673,518.25	2,134.59	2,152.59
GLM-4.6V-Flash	E (size)	1,314,847.97	1,677.10	1,703.10
GLM-4.6V-Flash	B (pixels)	1,673,756.90	2,134.89	2,216.89
GLM-4.6V-Flash	E (pixels)	1,673,504.13	2,134.57	2,160.57

Table 35: **Paired comparisons under the image-size control.** Differences are in pp with pointwise 95% intervals. Exact McNemar tests receive Holm correction over 15 comparisons per model. Gain/loss counts changes in correct answers. B: Baseline; R: *Role*; E: *Example*.

Model	Dataset	Pair	Δ	95% CI	Gain/loss	p_{Holm}
GPT-5.4	Flowers	R–B	+2.00	[0.40, 3.80]	14/4	0.278
GPT-5.4	CUB	R–B	+3.20	[-0.60, 7.00]	55/39	0.728

Model	Dataset	Pair	Δ	95% CI	Gain/loss	p_{Holm}
GPT-5.4	Food	R–B	-0.20	[-3.21, 2.81]	29/30	1
GPT-5.4	Cars	R–B	+2.40	[-1.20, 6.00]	49/37	0.942
GPT-5.4	Dogs	R–B	+5.20	[1.40, 9.00]	62/36	0.123
GPT-5.4	Flowers	E–B	+2.20	[0.60, 3.80]	14/3	0.127
GPT-5.4	CUB	E–B	+5.60	[1.80, 9.20]	59/31	0.0541
GPT-5.4	Food	E–B	+2.81	[-0.40, 6.01]	39/25	0.724
GPT-5.4	Cars	E–B	+7.00	[3.80, 10.40]	54/19	0.000759
GPT-5.4	Dogs	E–B	+5.20	[1.60, 8.80]	57/31	0.0881
GPT-5.4	Flowers	E–R	+0.20	[0.00, 0.60]	1/0	1
GPT-5.4	CUB	E–R	+2.40	[-0.80, 5.60]	41/29	0.941
GPT-5.4	Food	E–R	+3.01	[0.20, 5.81]	33/18	0.391
GPT-5.4	Cars	E–R	+4.60	[1.60, 7.40]	40/17	0.0448
GPT-5.4	Dogs	E–R	+0.00	[-3.00, 3.00]	28/28	1
Qwen3-VL-32B	Flowers	R–B	+8.80	[5.60, 12.00]	58/14	1.65e-06
Qwen3-VL-32B	CUB	R–B	+11.40	[6.60, 16.20]	108/51	5.06e-05
Qwen3-VL-32B	Food	R–B	+16.60	[12.20, 21.00]	110/27	6.57e-12
Qwen3-VL-32B	Cars	R–B	+8.20	[3.60, 12.80]	91/50	0.0042
Qwen3-VL-32B	Dogs	R–B	+12.60	[8.00, 17.20]	105/42	1.88e-06
Qwen3-VL-32B	Flowers	E–B	+9.80	[6.60, 13.00]	59/10	1.65e-08
Qwen3-VL-32B	CUB	E–B	+14.20	[10.00, 18.40]	97/26	1.04e-09
Qwen3-VL-32B	Food	E–B	+16.20	[11.80, 20.60]	113/32	1.19e-10
Qwen3-VL-32B	Cars	E–B	+11.80	[7.20, 16.40]	102/43	8.4e-06
Qwen3-VL-32B	Dogs	E–B	+14.20	[9.40, 19.00]	115/44	1.82e-07
Qwen3-VL-32B	Flowers	E–R	+1.00	[-0.80, 2.80]	12/7	1
Qwen3-VL-32B	CUB	E–R	+2.80	[-0.60, 6.40]	47/33	0.583
Qwen3-VL-32B	Food	E–R	-0.40	[-3.40, 2.40]	27/29	1
Qwen3-VL-32B	Cars	E–R	+3.60	[-0.40, 7.60]	61/43	0.475
Qwen3-VL-32B	Dogs	E–R	+1.60	[-2.20, 5.40]	52/44	1
Qwen3.5-27B	Flowers	R–B	+10.00	[5.80, 14.20]	84/34	3.81e-05
Qwen3.5-27B	CUB	R–B	+15.80	[10.80, 21.00]	130/51	4.6e-08
Qwen3.5-27B	Food	R–B	+10.20	[5.80, 14.60]	92/41	8.06e-05
Qwen3.5-27B	Cars	R–B	+3.60	[-1.20, 8.60]	86/68	0.324
Qwen3.5-27B	Dogs	R–B	+26.00	[21.00, 31.00]	166/36	5.3e-20
Qwen3.5-27B	Flowers	E–B	+3.60	[-1.20, 8.40]	83/65	0.324
Qwen3.5-27B	CUB	E–B	+5.80	[0.40, 11.20]	110/81	0.127
Qwen3.5-27B	Food	E–B	-13.60	[-18.80, -8.40]	61/129	9.05e-06
Qwen3.5-27B	Cars	E–B	-13.00	[-18.40, -7.60]	66/131	3.81e-05
Qwen3.5-27B	Dogs	E–B	+17.20	[11.60, 22.80]	151/65	5.08e-08
Qwen3.5-27B	Flowers	E–R	-6.40	[-10.00, -2.80]	29/61	0.00389
Qwen3.5-27B	CUB	E–R	-10.00	[-14.80, -5.40]	48/98	0.000256
Qwen3.5-27B	Food	E–R	-23.80	[-28.20, -19.20]	20/139	6.02e-22
Qwen3.5-27B	Cars	E–R	-16.60	[-21.40, -11.80]	44/127	2.04e-09
Qwen3.5-27B	Dogs	E–R	-8.80	[-13.20, -4.40]	45/89	0.000901
Gemma4-31B-IT	Flowers	R–B	+1.40	[-1.60, 4.40]	32/25	1
Gemma4-31B-IT	CUB	R–B	+2.40	[-2.20, 7.20]	79/67	1
Gemma4-31B-IT	Food	R–B	+4.40	[0.60, 8.20]	58/36	0.298
Gemma4-31B-IT	Cars	R–B	+4.80	[0.20, 9.40]	83/59	0.372
Gemma4-31B-IT	Dogs	R–B	+4.60	[-0.20, 9.40]	89/66	0.461
Gemma4-31B-IT	Flowers	E–B	+1.20	[-1.20, 3.80]	23/17	1
Gemma4-31B-IT	CUB	E–B	-5.00	[-9.80, -0.20]	60/85	0.367
Gemma4-31B-IT	Food	E–B	-4.80	[-9.00, -0.60]	45/69	0.298
Gemma4-31B-IT	Cars	E–B	-2.60	[-7.60, 2.40]	72/85	1
Gemma4-31B-IT	Dogs	E–B	-5.80	[-10.60, -1.20]	61/90	0.246
Gemma4-31B-IT	Flowers	E–R	-0.20	[-2.80, 2.40]	22/23	1
Gemma4-31B-IT	CUB	E–R	-7.40	[-12.20, -2.80]	56/93	0.0397
Gemma4-31B-IT	Food	E–R	-9.20	[-13.20, -5.20]	31/77	0.000167
Gemma4-31B-IT	Cars	E–R	-7.40	[-12.40, -2.60]	59/96	0.0443
Gemma4-31B-IT	Dogs	E–R	-10.40	[-15.60, -5.40]	59/111	0.00114
GLM-4.6V-Flash	Flowers	R–B	-2.40	[-5.40, 0.60]	22/34	1
GLM-4.6V-Flash	CUB	R–B	+3.20	[-1.40, 8.00]	80/64	1
GLM-4.6V-Flash	Food	R–B	+3.80	[-0.40, 8.00]	67/48	1
GLM-4.6V-Flash	Cars	R–B	-10.80	[-15.60, -6.20]	48/102	0.000185
GLM-4.6V-Flash	Dogs	R–B	-0.40	[-5.20, 4.40]	72/74	1

Model	Dataset	Pair	Δ	95% CI	Gain/loss	p_{Holm}
GLM-4.6V-Flash	Flowers	E–B	-1.20	[-3.60, 1.20]	17/23	1
GLM-4.6V-Flash	CUB	E–B	-0.60	[-4.40, 3.20]	47/50	1
GLM-4.6V-Flash	Food	E–B	+3.00	[-0.80, 6.80]	55/40	1
GLM-4.6V-Flash	Cars	E–B	-0.40	[-4.20, 3.40]	45/47	1
GLM-4.6V-Flash	Dogs	E–B	-1.60	[-5.60, 2.60]	51/59	1
GLM-4.6V-Flash	Flowers	E–R	+1.20	[-1.40, 3.80]	26/20	1
GLM-4.6V-Flash	CUB	E–R	-3.80	[-8.60, 1.00]	64/83	1
GLM-4.6V-Flash	Food	E–R	-0.80	[-4.80, 3.20]	50/54	1
GLM-4.6V-Flash	Cars	E–R	+10.40	[5.80, 15.00]	96/44	0.000185
GLM-4.6V-Flash	Dogs	E–R	-1.20	[-6.00, 3.40]	69/75	1

Table 36: **Paired comparisons under the total-pixel control.** Differences are in pp with pointwise 95% intervals. Exact McNemar tests receive Holm correction over 15 comparisons per model. Gain/loss counts changes in correct answers. B: Baseline; R: *Role*; E: *Example*.

Model	Dataset	Pair	Δ	95% CI	Gain/loss	p_{Holm}
GPT-5.4	Flowers	R–B	+8.00	[5.40, 10.60]	45/5	5.89e-08
GPT-5.4	CUB	R–B	-0.80	[-4.60, 3.00]	43/47	1
GPT-5.4	Food	R–B	+2.61	[-0.80, 6.01]	47/34	0.728
GPT-5.4	Cars	R–B	+4.00	[0.00, 8.40]	67/47	0.523
GPT-5.4	Dogs	R–B	+4.60	[0.60, 8.60]	66/43	0.277
GPT-5.4	Flowers	E–B	+8.60	[6.20, 11.20]	44/1	3.92e-11
GPT-5.4	CUB	E–B	+3.40	[-0.20, 7.00]	51/34	0.523
GPT-5.4	Food	E–B	+5.21	[2.00, 8.62]	50/24	0.0371
GPT-5.4	Cars	E–B	+8.00	[4.20, 11.80]	68/28	0.00071
GPT-5.4	Dogs	E–B	+6.00	[2.40, 9.60]	59/29	0.0219
GPT-5.4	Flowers	E–R	+0.60	[-0.20, 1.60]	4/1	1
GPT-5.4	CUB	E–R	+4.20	[0.80, 7.60]	49/28	0.198
GPT-5.4	Food	E–R	+2.61	[-0.20, 5.41]	32/19	0.523
GPT-5.4	Cars	E–R	+4.00	[1.00, 7.00]	41/21	0.151
GPT-5.4	Dogs	E–R	+1.40	[-2.00, 4.80]	41/34	1
Qwen3-VL-32B	Flowers	R–B	+11.20	[7.60, 14.80]	73/17	2.14e-08
Qwen3-VL-32B	CUB	R–B	+8.80	[4.00, 13.60]	101/57	0.00348
Qwen3-VL-32B	Food	R–B	+16.00	[11.80, 20.20]	103/23	3.9e-12
Qwen3-VL-32B	Cars	R–B	+11.40	[6.60, 16.20]	104/47	2.82e-05
Qwen3-VL-32B	Dogs	R–B	+12.60	[8.00, 17.20]	105/42	1.88e-06
Qwen3-VL-32B	Flowers	E–B	+13.40	[10.20, 16.80]	73/6	1.5e-14
Qwen3-VL-32B	CUB	E–B	+12.40	[7.80, 17.00]	106/44	3.52e-06
Qwen3-VL-32B	Food	E–B	+15.00	[10.60, 19.40]	105/30	7.76e-10
Qwen3-VL-32B	Cars	E–B	+15.80	[11.20, 20.40]	116/37	1.31e-09
Qwen3-VL-32B	Dogs	E–B	+13.60	[8.80, 18.40]	116/48	1.12e-06
Qwen3-VL-32B	Flowers	E–R	+2.20	[0.80, 3.80]	13/2	0.0369
Qwen3-VL-32B	CUB	E–R	+3.60	[0.00, 7.20]	53/35	0.208
Qwen3-VL-32B	Food	E–R	-1.00	[-4.00, 2.00]	28/33	1
Qwen3-VL-32B	Cars	E–R	+4.40	[0.60, 8.20]	59/37	0.126
Qwen3-VL-32B	Dogs	E–R	+1.00	[-2.80, 4.80]	49/44	1
Qwen3.5-27B	Flowers	R–B	+11.60	[7.40, 15.80]	92/34	2.36e-06
Qwen3.5-27B	CUB	R–B	+20.80	[15.80, 25.80]	147/43	2.16e-13
Qwen3.5-27B	Food	R–B	+4.80	[0.80, 9.00]	66/42	0.106
Qwen3.5-27B	Cars	R–B	-0.40	[-5.20, 4.40]	74/76	0.935
Qwen3.5-27B	Dogs	R–B	+23.60	[18.40, 28.80]	162/44	6.36e-16
Qwen3.5-27B	Flowers	E–B	+8.40	[3.80, 13.00]	90/48	0.00308
Qwen3.5-27B	CUB	E–B	+17.00	[12.00, 22.00]	137/52	5.62e-09
Qwen3.5-27B	Food	E–B	-16.60	[-21.60, -11.80]	42/125	1.09e-09
Qwen3.5-27B	Cars	E–B	-9.80	[-15.20, -4.20]	71/120	0.00308
Qwen3.5-27B	Dogs	E–B	+12.00	[6.60, 17.40]	135/75	0.000334
Qwen3.5-27B	Flowers	E–R	-3.20	[-7.00, 0.40]	36/52	0.313
Qwen3.5-27B	CUB	E–R	-3.80	[-8.00, 0.60]	52/71	0.313
Qwen3.5-27B	Food	E–R	-21.40	[-26.00, -16.80]	25/132	1.36e-17
Qwen3.5-27B	Cars	E–R	-9.40	[-14.60, -4.20]	66/113	0.00308

Model	Dataset	Pair	Δ	95% CI	Gain/loss	p_{Holm}
Qwen3.5-27B	Dogs	E–R	-11.60	[-16.20, -7.00]	42/100	1.12e-05
Gemma4-31B-IT	Flowers	R–B	-3.60	[-6.20, -1.20]	13/31	0.086
Gemma4-31B-IT	CUB	R–B	-6.40	[-11.20, -1.60]	59/91	0.089
Gemma4-31B-IT	Food	R–B	-3.20	[-6.40, 0.00]	24/40	0.373
Gemma4-31B-IT	Cars	R–B	-9.40	[-13.40, -5.20]	36/83	0.000275
Gemma4-31B-IT	Dogs	R–B	-6.80	[-11.20, -2.60]	44/78	0.0266
Gemma4-31B-IT	Flowers	E–B	-1.00	[-3.20, 1.20]	14/19	1
Gemma4-31B-IT	CUB	E–B	-8.40	[-12.80, -4.00]	45/87	0.0042
Gemma4-31B-IT	Food	E–B	-6.20	[-9.60, -3.00]	22/53	0.0054
Gemma4-31B-IT	Cars	E–B	-13.00	[-17.20, -8.60]	33/98	1.69e-07
Gemma4-31B-IT	Dogs	E–B	-7.60	[-12.00, -3.40]	44/82	0.00996
Gemma4-31B-IT	Flowers	E–R	+2.60	[0.20, 5.00]	26/13	0.373
Gemma4-31B-IT	CUB	E–R	-2.00	[-6.80, 2.80]	70/80	1
Gemma4-31B-IT	Food	E–R	-3.00	[-6.40, 0.40]	32/47	0.573
Gemma4-31B-IT	Cars	E–R	-3.60	[-8.20, 1.00]	58/76	0.573
Gemma4-31B-IT	Dogs	E–R	-0.80	[-5.20, 3.80]	64/68	1
GLM-4.6V-Flash	Flowers	R–B	-2.60	[-5.60, 0.40]	21/34	1
GLM-4.6V-Flash	CUB	R–B	+1.40	[-3.20, 6.00]	74/67	1
GLM-4.6V-Flash	Food	R–B	+4.00	[-0.20, 8.00]	65/45	0.768
GLM-4.6V-Flash	Cars	R–B	-11.40	[-16.20, -6.60]	50/107	9.43e-05
GLM-4.6V-Flash	Dogs	R–B	+2.00	[-2.80, 6.80]	80/70	1
GLM-4.6V-Flash	Flowers	E–B	-0.80	[-3.20, 1.60]	17/21	1
GLM-4.6V-Flash	CUB	E–B	+0.20	[-3.60, 4.00]	47/46	1
GLM-4.6V-Flash	Food	E–B	+4.60	[0.60, 8.60]	64/41	0.407
GLM-4.6V-Flash	Cars	E–B	-1.60	[-5.60, 2.60]	52/60	1
GLM-4.6V-Flash	Dogs	E–B	+4.20	[0.00, 8.40]	69/48	0.768
GLM-4.6V-Flash	Flowers	E–R	+1.80	[-1.00, 4.60]	29/20	1
GLM-4.6V-Flash	CUB	E–R	-1.20	[-5.80, 3.40]	64/70	1
GLM-4.6V-Flash	Food	E–R	+0.60	[-3.40, 4.40]	52/49	1
GLM-4.6V-Flash	Cars	E–R	+9.80	[5.00, 14.60]	102/53	0.00143
GLM-4.6V-Flash	Dogs	E–R	+2.20	[-2.40, 7.00]	79/68	1

Table 39: **VL-ICL Peak accuracy and ICL efficiency.** Metrics use 0/1/2/4 shots. Scores are percentages; gains and intervals are in pp. Pointwise 95% query-cluster bootstrap intervals recompute Peak in each draw and are not multiplicity-adjusted.

Model	Metric	Base.	VCC	Gain	95% CI
GPT-5.4	Peak	82.72	85.31	+2.59	[1.22, 3.77]
GPT-5.4	Eff.	40.91	46.17	+5.26	[1.47, 9.17]
Qwen3-VL-32B	Peak	89.49	93.25	+3.76	[1.78, 5.60]
Qwen3-VL-32B	Eff.	41.47	47.91	+6.44	[2.99, 10.03]
Qwen3.5-27B	Peak	53.18	92.67	+39.49	[37.29, 40.88]
Qwen3.5-27B	Eff.	16.75	36.28	+19.53	[16.34, 22.68]
Gemma4-31B-IT	Peak	85.58	92.46	+6.88	[5.46, 8.00]
Gemma4-31B-IT	Eff.	39.31	49.86	+10.54	[6.81, 14.27]
GLM-4.6V-Flash	Peak	78.94	89.15	+10.21	[8.19, 12.33]
GLM-4.6V-Flash	Eff.	32.13	46.19	+14.06	[10.05, 18.10]

D.2 VL-ICL

Table 37: **Grouping results for VL-ICL.** Accuracy (%).

Model	Shot	Baseline	*Joint*	*Role*	*Example*
GPT-5.4	1	65.24	81.25	80.38	80.92
GPT-5.4	2	81.85	88.20	83.28	86.32
GPT-5.4	4	82.69	87.67	83.57	85.65
Qwen3-VL-32B	1	73.78	84.49	85.04	85.39
Qwen3-VL-32B	2	84.76	91.42	90.49	92.29
Qwen3-VL-32B	4	84.24	92.21	90.17	91.38
Qwen3.5-27B	1	39.01	75.71	78.35	78.36
Qwen3.5-27B	2	51.97	87.39	89.10	90.97
Qwen3.5-27B	4	51.68	87.00	91.50	93.62
Gemma4-31B-IT	1	63.00	82.26	82.57	79.62
Gemma4-31B-IT	2	84.50	89.54	91.63	90.88
Gemma4-31B-IT	4	85.33	89.71	92.29	91.29
GLM-4.6V-Flash	1	47.53	61.90	71.42	64.07
GLM-4.6V-Flash	2	77.47	81.29	84.79	82.19
GLM-4.6V-Flash	4	76.36	84.12	85.90	82.68

Table 38: **Uncertainty in VL-ICL accuracy gains.** Paired gains for *Role* minus Baseline (pp), pointwise 95% intervals, and Holm-adjusted p values for shot-averaged accuracy, not Peak or Eff. The *Queries* column counts query clusters.

Model	Δ	95% CI	p	Queries
GPT-5.4	+5.81	[4.61, 6.96]	0.0010	660
Qwen3-VL-32B	+7.64	[5.67, 9.63]	0.0010	660
Qwen3.5-27B	+38.76	[37.48, 40.07]	0.0010	660
Gemma4-31B-IT	+11.22	[10.07, 12.37]	0.0010	660
GLM-4.6V-Flash	+13.58	[12.01, 15.14]	0.0010	660

Table 40: **VL-ICL results by task.** Accuracy (%) and gains for *Role* minus Baseline (pp), averaged over paired support-seed cells.

Task	Model	Shot	Baseline	VCC	Δ
clevr	Gemma4-31B-IT	1	52.00	80.67	28.67
	Gemma4-31B-IT	2	58.83	82.83	24.00
	Gemma4-31B-IT	4	58.00	85.67	27.67
	GLM-4.6V-Flash	1	35.50	82.67	47.17
	GLM-4.6V-Flash	2	52.83	75.33	22.50
	GLM-4.6V-Flash	4	45.33	69.67	24.33
	GPT-5.4	1	40.17	62.67	22.50
	GPT-5.4	2	51.33	56.33	5.00
	GPT-5.4	4	53.50	56.33	2.83
	Qwen3-VL-32B	1	89.50	91.33	1.83
	Qwen3-VL-32B	2	78.33	83.33	5.00
	Qwen3-VL-32B	4	74.00	81.67	7.67
	Qwen3.5-27B	1	47.00	91.67	44.67
	Qwen3.5-27B	2	50.17	88.50	38.33
	Qwen3.5-27B	4	44.50	87.00	42.50
open mi	Gemma4-31B-IT	1	21.67	71.17	49.50
	Gemma4-31B-IT	2	95.50	98.83	3.33
	Gemma4-31B-IT	4	99.00	99.33	0.33
	GLM-4.6V-Flash	1	12.00	33.50	21.50
	GLM-4.6V-Flash	2	90.67	93.00	2.33
	GLM-4.6V-Flash	4	89.17	96.67	7.50
	GPT-5.4	1	53.93	85.64	31.72
	GPT-5.4	2	98.99	98.99	0.00
	GPT-5.4	4	99.67	100.00	0.33
	Qwen3-VL-32B	1	40.50	73.83	33.33
	Qwen3-VL-32B	2	97.50	98.67	1.17
	Qwen3-VL-32B	4	99.83	99.67	-0.17
	Qwen3.5-27B	1	17.33	46.00	28.67
	Qwen3.5-27B	2	67.17	85.00	17.83
	Qwen3.5-27B	4	70.33	94.50	24.17
operator induction	Gemma4-31B-IT	1	95.00	94.44	-0.56
	Gemma4-31B-IT	2	100.00	100.00	0.00
	Gemma4-31B-IT	4	100.00	100.00	0.00
	GLM-4.6V-Flash	1	71.11	91.67	20.56
	GLM-4.6V-Flash	2	88.89	93.33	4.44
	GLM-4.6V-Flash	4	92.78	97.78	5.00
	GPT-5.4	1	89.44	95.00	5.56
	GPT-5.4	2	99.44	100.00	0.56
	GPT-5.4	4	100.00	100.00	0.00
	Qwen3-VL-32B	1	86.11	95.00	8.89
	Qwen3-VL-32B	2	83.89	99.44	15.56
	Qwen3-VL-32B	4	82.78	100.00	17.22
	Qwen3.5-27B	1	8.89	92.22	83.33
	Qwen3.5-27B	2	7.22	98.89	91.67
	Qwen3.5-27B	4	8.89	100.00	91.11

Task	Model	Shot	Baseline	VCC	Δ
	Gemma4-31B-IT	1	83.33	84.00	0.67
	Gemma4-31B-IT	2	83.67	84.83	1.17
	Gemma4-31B-IT	4	84.33	84.17	-0.17
	GLM-4.6V-Flash	1	71.50	77.83	6.33
	GLM-4.6V-Flash	2	77.50	77.50	0.00
	GLM-4.6V-Flash	4	78.17	79.50	1.33
	GPT-5.4	1	77.63	78.30	0.67
textocr	GPT-5.4	2	77.63	77.80	0.16
	GPT-5.4	4	77.61	77.94	0.33
	Qwen3-VL-32B	1	79.00	80.00	1.00
	Qwen3-VL-32B	2	79.33	80.50	1.17
	Qwen3-VL-32B	4	80.33	79.33	-1.00
	Qwen3.5-27B	1	82.83	83.50	0.67
	Qwen3.5-27B	2	83.33	84.00	0.67
	Qwen3.5-27B	4	83.00	84.50	1.50

D.3 TRUEMICL

Table 41: **Grouping results for TrueMICL.** Four-shot accuracy (%).

Model	Baseline	*Joint*	*Role*	*Example*
GPT-5.4	92.95	92.95	94.57	95.43
Qwen3-VL-32B	91.48	91.10	92.38	92.38
Qwen3.5-27B	79.71	94.10	91.43	88.90
Gemma4-31B-IT	97.14	95.14	98.19	96.71
GLM-4.6V-Flash	73.14	71.43	80.00	77.14

Table 42: **Uncertainty in TrueMICL gains.** Paired gains for *Role* minus Baseline (pp), pointwise 95% intervals, and Holm-adjusted p values. The *Queries* column counts query clusters.

Model	Δ	95% CI	p	Queries
GPT-5.4	+1.62	[0.19, 3.10]	0.1380	650
Qwen3-VL-32B	+0.90	[-1.10, 2.95]	0.4241	650
Qwen3.5-27B	+11.71	[9.76, 13.76]	0.0010	650
Gemma4-31B-IT	+1.05	[-0.19, 2.38]	0.2967	650
GLM-4.6V-Flash	+6.86	[4.29, 9.48]	0.0010	650

Table 43: **TrueMICL results by task.** Four-shot accuracy (%) and gains for *Role* minus Baseline (pp), averaged over paired support-seed cells.

Task	Model	Baseline	VCC	Δ
classification	Gemma4-31B-IT	86.67	96.00	9.33
	GLM-4.6V-Flash	59.33	70.67	11.33
	GPT-5.4	75.33	88.67	13.33
	Qwen3-VL-32B	70.00	81.33	11.33
	Qwen3.5-27B	73.33	94.00	20.67
clevr	Gemma4-31B-IT	94.00	95.00	1.00
	GLM-4.6V-Flash	67.67	75.67	8.00
	GPT-5.4	85.33	88.67	3.33
	Qwen3-VL-32B	95.00	91.33	-3.67
	Qwen3.5-27B	97.33	98.33	1.00
clock	Gemma4-31B-IT	100.00	97.00	-3.00
	GLM-4.6V-Flash	54.67	57.00	2.33
	GPT-5.4	93.67	93.67	0.00
	Qwen3-VL-32B	99.00	84.67	-14.33
	Qwen3.5-27B	91.00	78.00	-13.00
operator induction	Gemma4-31B-IT	99.33	99.33	0.00
	GLM-4.6V-Flash	77.00	93.00	16.00
	GPT-5.4	98.00	98.67	0.67
	Qwen3-VL-32B	80.00	94.33	14.33
	Qwen3.5-27B	7.67	99.33	91.67
palindrome	Gemma4-31B-IT	100.00	100.00	0.00
	GLM-4.6V-Flash	67.33	81.33	14.00
	GPT-5.4	98.33	92.33	-6.00
	Qwen3-VL-32B	97.00	98.00	1.00
	Qwen3.5-27B	98.00	93.67	-4.33
shapes count	Gemma4-31B-IT	100.00	100.00	0.00
	GLM-4.6V-Flash	90.33	83.67	-6.67
	GPT-5.4	100.00	100.00	0.00
	Qwen3-VL-32B	99.33	97.00	-2.33
	Qwen3.5-27B	92.00	93.00	1.00

Task	Model	Baseline	VCC	Δ
sudoku	Gemma4-31B-IT	100.00	100.00	0.00
	GLM-4.6V-Flash	95.67	98.67	3.00
	GPT-5.4	100.00	100.00	0.00
	Qwen3-VL-32B	100.00	100.00	0.00
	Qwen3.5-27B	98.67	83.67	-15.00

Table 46: **Variation in VCC gains across MuirBench tasks.** Accuracy (%) weights five models equally; Δ is *Role* minus Baseline (pp). *Models improved* counts positive point estimates, not significant gains. Each Baseline uses valid pairs for its *Role* comparison. The main-table score instead weights tasks by valid sample counts within each model.

Task	Baseline	*Role*	Δ	Models improved
Action understanding	50.11	52.44	+2.33	4/5
Attribute similarity	58.16	58.57	+0.41	3/5
Cartoon understanding	54.64	59.05	+4.41	4/5
Counting	65.22	65.72	+0.50	2/5
Diagram understanding	82.56	88.24	+5.68	4/5
Difference spotting	64.36	68.63	+4.27	4/5
Geographic understanding	58.60	52.40	-6.20	1/5
Image text matching	82.57	83.12	+0.55	2/5
Ordering	38.75	45.94	+7.19	3/5
Scene understanding	75.38	81.61	+6.24	4/5
Visual grounding	46.67	49.52	+2.86	3/5
Visual retrieval	69.66	74.38	+4.73	2/5

D.4 MUIRBENCH

Table 44: **Grouping results for MuirBench.** Accuracy (%).

Model	Baseline	*Joint*	*Role*
GPT-5.4	68.74	69.28	69.01
Qwen3-VL-32B	71.12	71.23	71.88
Qwen3.5-27B	68.12	71.35	75.04
Gemma4-31B-IT	70.77	73.19	75.77
GLM-4.6V-Flash	65.15	63.96	66.62

Table 45: **Uncertainty in MuirBench gains.** Paired gains for *Role* minus Baseline (pp), pointwise 95% intervals, and Holm-adjusted p values. The *Queries* column counts query clusters.

Model	Δ	95% CI	p	Queries
GPT-5.4	+0.27	[-1.21, 1.75]	0.7644	2572
Qwen3-VL-32B	+0.77	[-0.65, 2.27]	0.6643	2600
Qwen3.5-27B	+6.92	[5.04, 8.81]	0.0010	2600
Gemma4-31B-IT	+5.00	[3.31, 6.73]	0.0010	2600
GLM-4.6V-Flash	+1.46	[-0.35, 3.23]	0.3599	2600

Table 47: **MuirBench results by task.** Accuracy (%) and gains for *Role* minus Baseline (pp) on queries eligible in both conditions.

Task	Model	Baseline	VCC	Δ
action understanding	Gemma4-31B-IT	48.17	54.88	6.71
	GLM-4.6V-Flash	46.34	51.22	4.88
	GPT-5.4	47.50	50.62	3.12
	Qwen3-VL-32B	53.66	56.71	3.05
	Qwen3.5-27B	54.88	48.78	-6.10
attribute similarity	Gemma4-31B-IT	54.08	61.22	7.14
	GLM-4.6V-Flash	58.67	60.71	2.04
	GPT-5.4	57.65	57.14	-0.51
	Qwen3-VL-32B	66.84	59.69	-7.14
	Qwen3.5-27B	53.57	54.08	0.51

Task	Model	Baseline	VCC	Δ
cartoon understanding	Gemma4-31B-IT	55.13	53.85	-1.28
	GLM-4.6V-Flash	44.87	52.56	7.69
	GPT-5.4	55.26	64.47	9.21
	Qwen3-VL-32B	61.54	62.82	1.28
	Qwen3.5-27B	56.41	61.54	5.13
counting	Gemma4-31B-IT	79.49	85.04	5.56
	GLM-4.6V-Flash	57.26	64.53	7.26
	GPT-5.4	54.31	47.84	-6.47
	Qwen3-VL-32B	50.43	50.43	0.00
	Qwen3.5-27B	84.62	80.77	-3.85
diagram understanding	Gemma4-31B-IT	71.36	89.45	18.09
	GLM-4.6V-Flash	84.42	83.92	-0.50
	GPT-5.4	83.67	86.68	3.02
	Qwen3-VL-32B	86.18	88.94	2.76
	Qwen3.5-27B	87.19	92.21	5.03
difference spotting	Gemma4-31B-IT	74.71	77.65	2.94
	GLM-4.6V-Flash	54.41	49.71	-4.71
	GPT-5.4	64.44	70.52	6.08
	Qwen3-VL-32B	63.82	66.76	2.94
	Qwen3.5-27B	64.41	78.53	14.12
geographic understanding	Gemma4-31B-IT	58.00	47.00	-11.00
	GLM-4.6V-Flash	61.00	45.00	-16.00
	GPT-5.4	54.00	50.00	-4.00
	Qwen3-VL-32B	72.00	69.00	-3.00
	Qwen3.5-27B	48.00	51.00	3.00
image text matching	Gemma4-31B-IT	84.05	84.70	0.65
	GLM-4.6V-Flash	78.02	76.94	-1.08
	GPT-5.4	84.84	83.08	-1.76
	Qwen3-VL-32B	87.07	85.13	-1.94
	Qwen3.5-27B	78.88	85.78	6.90
ordering	Gemma4-31B-IT	43.75	45.31	1.56
	GLM-4.6V-Flash	17.19	40.62	23.44
	GPT-5.4	37.50	37.50	0.00
	Qwen3-VL-32B	43.75	59.38	15.62
	Qwen3.5-27B	51.56	46.88	-4.69
scene understanding	Gemma4-31B-IT	75.81	88.17	12.37
	GLM-4.6V-Flash	67.74	83.33	15.59
	GPT-5.4	76.34	77.42	1.08
	Qwen3-VL-32B	69.35	73.12	3.76
	Qwen3.5-27B	87.63	86.02	-1.61
visual grounding	Gemma4-31B-IT	53.57	58.33	4.76
	GLM-4.6V-Flash	34.52	40.48	5.95
	GPT-5.4	39.29	46.43	7.14
	Qwen3-VL-32B	48.81	47.62	-1.19
	Qwen3.5-27B	57.14	54.76	-2.38
visual retrieval	Gemma4-31B-IT	77.40	74.32	-3.08
	GLM-4.6V-Flash	76.71	74.32	-2.40
	GPT-5.4	77.74	71.92	-5.82
	Qwen3-VL-32B	78.77	79.79	1.03
	Qwen3.5-27B	37.67	71.58	33.90

Table 48: **Grouping selection across models.** Constructions are selected on one or two models and evaluated on the rest. Gains and ranges are in pp. Pointwise 95% cluster bootstrap intervals cover the mean over splits, repeating selection in each draw. *Positive* counts splits with positive point estimates.

Scope	Select / evaluate	Mean gain [95% CI]	Range	Positive
Fine-grained	1 / 4	+4.28 [+3.67, +4.87]	[-0.84, +6.53]	4/5
Fine-grained	2 / 3	+3.19 [+2.56, +3.82]	[-4.44, +8.22]	8/10
VL-ICL	1 / 4	+14.68 [+13.81, +15.53]	[+8.83, +16.66]	5/5
VL-ICL	2 / 3	+14.55 [+13.70, +15.44]	[+8.22, +19.39]	10/10
TrueMICL	1 / 4	+3.36 [+2.18, +4.41]	[-1.02, +5.31]	4/5
TrueMICL	2 / 3	+3.33 [+1.95, +5.01]	[-1.37, +6.73]	8/10
MuirBench	1 / 4	+2.41 [+1.51, +3.65]	[+1.14, +3.41]	5/5
MuirBench	2 / 3	+2.89 [+1.74, +3.86]	[+0.83, +4.46]	10/10

D.5 Transfer of grouping selection across models and benchmarks

We evaluate whether a configuration selected on a small set of models or tasks improves accuracy on the remaining models or tasks. Each split uses the source results to select the configuration with the highest mean accuracy and applies that configuration to every target instance. Target results are excluded from selection. We enumerate every source subset of the specified size. Candidates are *Joint*, *Role*, and *Example*, except where MuirBench restricts the comparison to *Joint* and *Role*. Ties are resolved in that order, allowing a tolerance of 10^{-12} in accuracy to account for numerical roundoff.

Data and aggregation. The analysis uses the saved outputs from all five models and 28 tasks or categories. Within each model/task/shot/seed cell, we retain the query IDs that are eligible for Baseline and every candidate configuration. This removes 17 additional query occurrences across cells relative to the corresponding Baseline–*Joint* comparisons; repeated occurrences across seeds and shots are counted separately here. Fine-grained classification, VL-ICL, and TrueMICL give equal weight to tasks, support seeds, and evaluated shot settings. MuirBench weights categories by valid query counts within each model. Models and source/target benchmark families receive equal weight. VL-ICL uses mean accuracy over one, two, and four shots throughout this analysis. All source splits receive equal weight in the reported means.

Uncertainty and image reuse. We compute pointwise 95% percentile intervals from 10,000 cluster bootstrap draws, repeating configuration selection and target evaluation in each draw. Repetitions of a query across configurations, models, seeds, and shots share resampling weights. We also join queries that share an exact image-file hash, including transitive matches, and resample these clusters jointly within connected groups of tasks. The 6,410 query IDs form 4,455 clusters. Of these, 132 span multiple tasks or categories, including 112 spanning VL-ICL and TrueMICL through CLEVR or operator induction. All query image paths were resolved. This audit detects exact file copies; it does not exhaustively detect re-encoded images, semantic duplicates, or overlap with support examples. The intervals condition on the evaluated models and tasks and do not treat overlapping splits as independent observations.

Selection on one or two models. For each benchmark family, we select a configuration on one model and evaluate it on the other four, or select on two models and evaluate on the other three. These settings give five and ten source subsets, respectively. Table 48 shows positive mean gains for all four families in both settings. In fine-grained classification and TrueMICL, selection on Qwen3.5-27B alone favors *Joint*, which reduces mean accuracy on the other four models by 0.84 and 1.02 pp, respectively. The same families also have regressions for two of the ten two-model source subsets.

Selection on one or two tasks within a family. We select on one or two datasets, tasks, or categories and evaluate on all remaining members of the same family, using all five models on both sides. Fine-grained classification uses five datasets; VL-ICL and TrueMICL use four and seven tasks; MuirBench uses 12 categories of one benchmark. Table 49 extends the single-task selection

Table 49: **Grouping selection across tasks.** Constructions are selected on one or two tasks and evaluated on the rest within each family. Gains and ranges are in pp. Pointwise 95% cluster bootstrap intervals cover the mean over splits, repeating selection in each draw. *Positive* counts splits with positive point estimates.

Scope	Select / evaluate	Mean gain [95% CI]	Range	Positive
Fine-grained	1 / 4	+4.77 [+4.15, +5.42]	[+3.11, +5.83]	5/5
Fine-grained	2 / 3	+5.01 [+4.26, +5.88]	[+2.46, +6.99]	10/10
VL-ICL	1 / 3	+14.66 [+13.84, +15.60]	[+11.25, +19.77]	4/4
VL-ICL	2 / 2	+14.81 [+13.84, +15.69]	[+6.45, +22.93]	6/6
TrueMICL	1 / 6	+3.35 [+2.33, +4.48]	[-0.33, +5.43]	6/7
TrueMICL	2 / 5	+3.27 [+2.19, +4.36]	[-1.35, +6.33]	19/21
MuirBench	1 / 11	+2.35 [+1.05, +3.46]	[+0.87, +3.12]	12/12
MuirBench	2 / 10	+2.51 [+1.27, +3.62]	[-0.07, +3.75]	65/66

Table 50: **Grouping selection across benchmark families.** Constructions are selected on source families and evaluated on the rest. Four-family comparisons use *Joint* and *Role*; three-family comparisons exclude MuirBench and include *Example*. Gains and ranges are in pp. Pointwise 95% cluster bootstrap intervals cover the mean over splits, repeating selection in each draw. *Positive* counts splits with positive point estimates.

Scope	Select / evaluate	Mean gain [95% CI]	Range	Positive
Four families	1 / 3	+4.46 [+4.00, +4.96]	[+1.41, +5.80]	4/4
Four families	2 / 2	+4.02 [+3.54, +4.54]	[-0.10, +8.19]	5/6
Three families	1 / 2	+5.33 [+4.82, +6.62]	[+0.67, +9.15]	3/3

results in Section 4.3 to two-task selection and reports split ranges. Two-task selection gives positive mean gains in every family, with two negative splits in TrueMICL and one in MuirBench. Positive split counts refer to mean gains over target models and tasks.

Selection across benchmark families. We also select on one of the four families and evaluate on the other three, or select on two and evaluate on the other two. These comparisons use *Joint* and *Role* because *Example* is unavailable for MuirBench. A separate three-family analysis excludes MuirBench and selects among all three configurations on one family before evaluating on the other two. Table 50 reports all three settings. Mean gains are positive, but the four-family two-source setting has one negative split: selection on VL-ICL and TrueMICL favors *Role*, whose mean gain on fine-grained classification and MuirBench is -0.10 pp. This analysis includes shared images between families and therefore does not constitute evaluation on an independent set of images.

Table 51 gives the selection frequencies. The accompanying result file records every split, source score, target model/task gain, and source-file hash. These post-hoc analyses quantify transfer within the existing evaluation collection.

Selection on four models. Using the previously reported condition means, selecting on four models improves accuracy on the omitted model in all 20 evaluations, with mean gains of 5.37, 14.53, 4.43, and 2.89 pp for fine-grained classification, VL-ICL, TrueMICL, and MuirBench, respectively. Table 51 gives the selected configurations; the VL-ICL fold omitting GLM-4.6V-Flash selects *Example*, yielding a 9.19 pp gain on that model.

Table 51: **Frequencies of selected groupings.** Counts cover all source splits. Unit identifies the selection/evaluation split unit; source size four gives the supplementary 4/1 model comparison.

Unit	Scope	Source size	*Joint*	*Role*	*Example*
Model	Fine-grained	1	1	0	4
Model	Fine-grained	2	4	0	6
Model	Fine-grained	4	0	0	5
Task	Fine-grained	1	1	0	4
Task	Fine-grained	2	1	0	9
Model	VL-ICL	1	1	2	2
Model	VL-ICL	2	2	6	2
Model	VL-ICL	4	0	4	1
Task	VL-ICL	1	1	1	2
Task	VL-ICL	2	0	3	3
Model	TrueMICL	1	1	3	1
Model	TrueMICL	2	2	7	1
Model	TrueMICL	4	0	5	0
Task	TrueMICL	1	1	4	2
Task	TrueMICL	2	2	13	6
Model	MuirBench	1	1	4	–
Model	MuirBench	2	0	10	–
Model	MuirBench	4	0	5	–
Task	MuirBench	1	3	9	–
Task	MuirBench	2	10	56	–
Family	Four families	1	1	3	–
Family	Four families	2	3	3	–
Family	Three families	1	0	2	1

E COMPLETE INPUT CONTEXTS

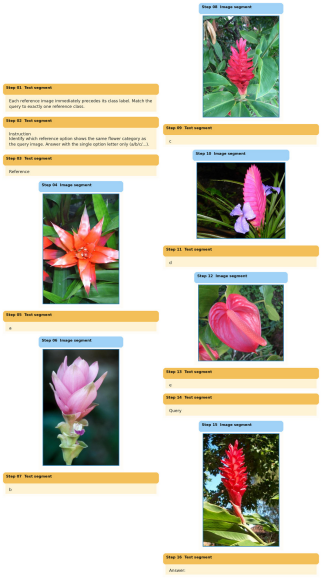
The examples show the complete ordered user content, including external text and every input image. Numbered blocks indicate presentation order; colored borders and headings are annotations and are not model inputs. Images are scaled for display. Each context shows one representative query and shot setting. Main-benchmark examples use Qwen3-VL-32B and seed 0; shot counts are stated below. *Joint* and *Role* use the same grouping definitions throughout the appendix; rendering settings remain specific to each study.

Study	Contexts
Fine-grained classification	10
VL-ICL	8
TrueMICL	14
MuirBench	24
Zero-shot responses	4
Image instructions	10
Counting	2
Caption-only classification	5
Query conflict	2
Concentric compositions	30
Joint mask and drop	10

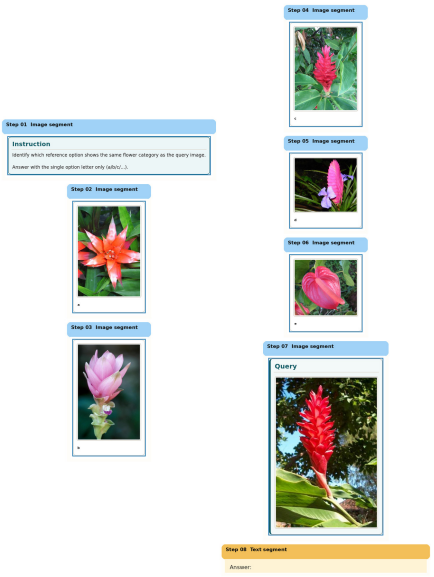
All 28 main-benchmark tasks are represented by Baseline and the adopted VCC configuration: *Example* for fine-grained classification and *Role* for VL-ICL, TrueMICL, and MuirBench. Shot counts do not apply to MuirBench and are omitted. Resolution controls are specified by their matching rules and measured input sizes in Appendix D.1.1; rescaled context images cannot demonstrate these constraints. The readout input is already illustrated in Figure 4. Token removal after encoding leaves the displayed pixels unchanged. The full-resolution context images and text transcripts accompany these reduced-size views.

E.1 FINE-GRAINED CLASSIFICATION

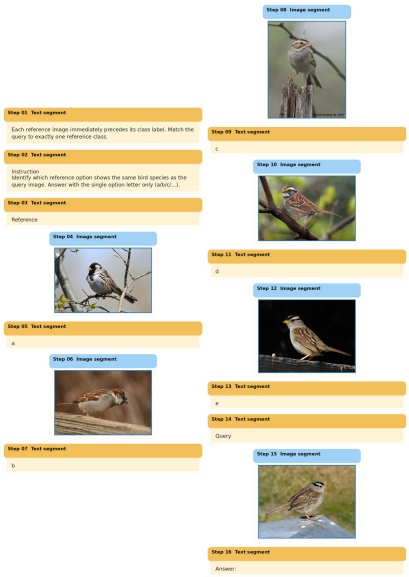
102Flowers: Baseline (1 shot)



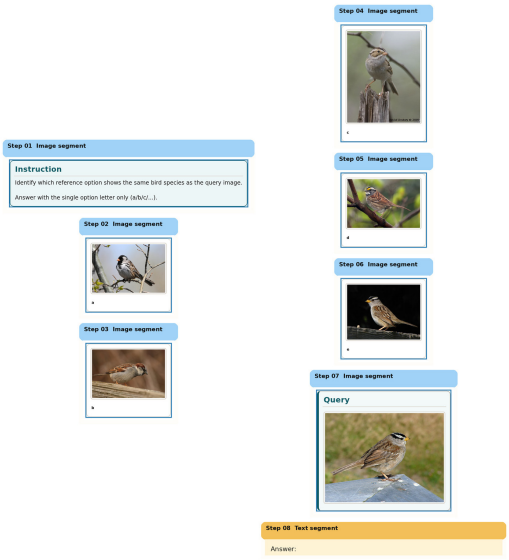
102Flowers: *Example* (1 shot)



CUB-Birds: Baseline (1 shot)

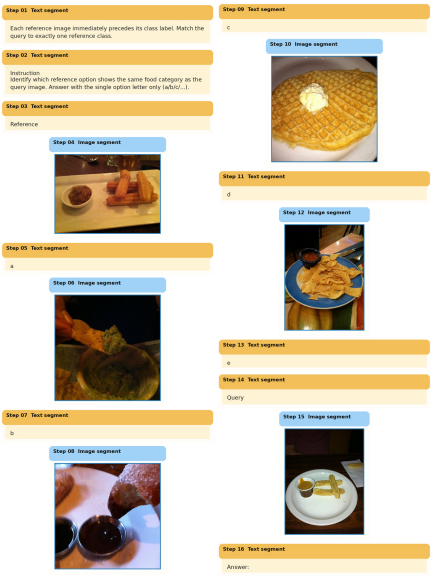


CUB-Birds: *Example* (1 shot)

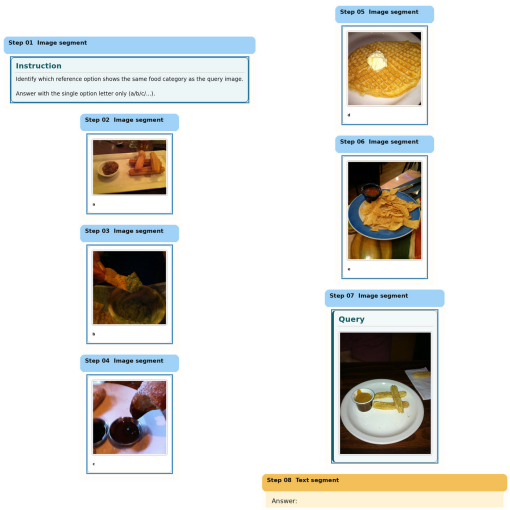


Fine-grained classification (continued)

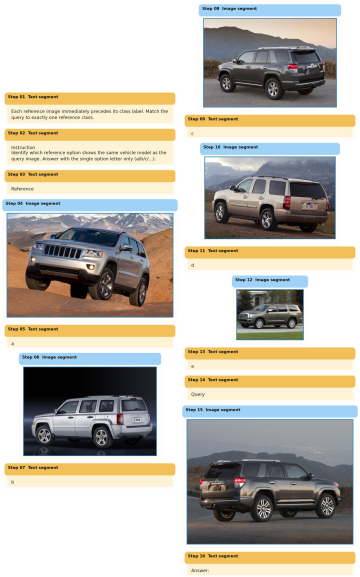
Food-101: Baseline (1 shot)



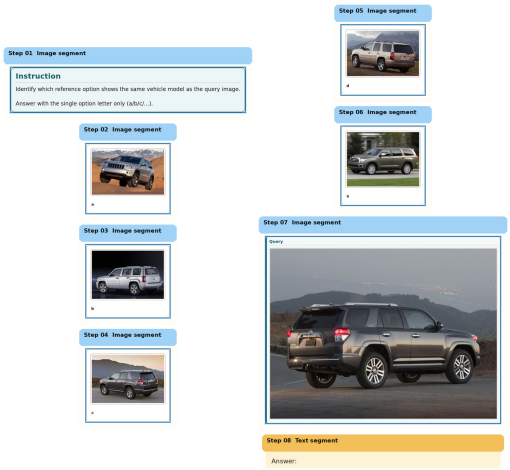
Food-101: *Example* (1 shot)



Stanford Cars: Baseline (1 shot)

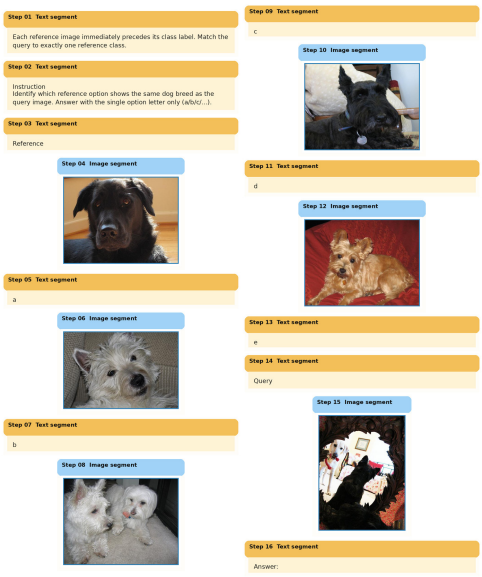


Stanford Cars: *Example* (1 shot)

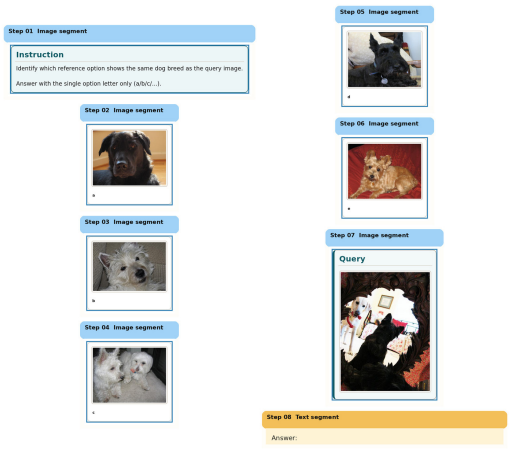


Fine-grained classification (continued)

Stanford Dogs: Baseline (1 shot)

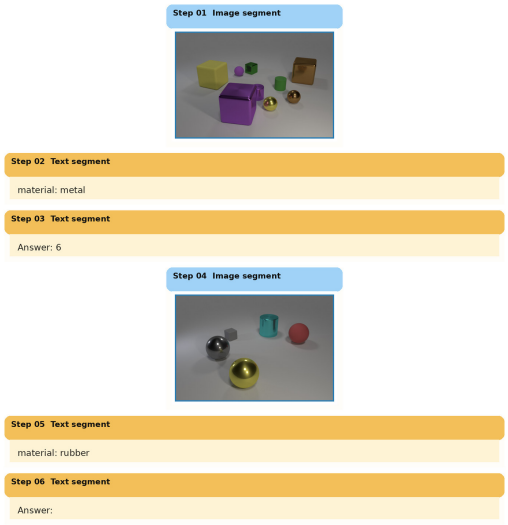


Stanford Dogs: *Example* (1 shot)

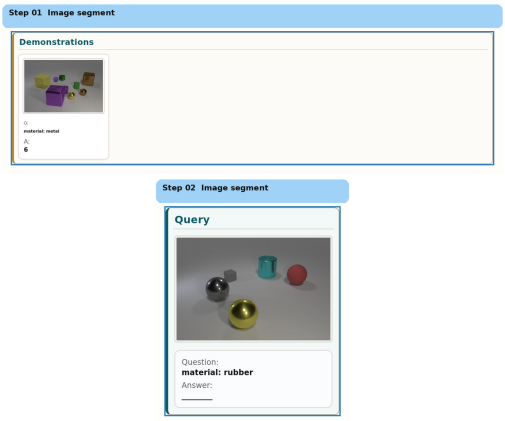


E.2 VL-ICL

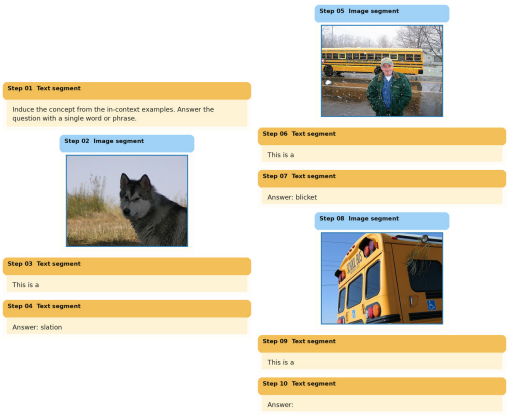
clevr: Baseline (1 shot)



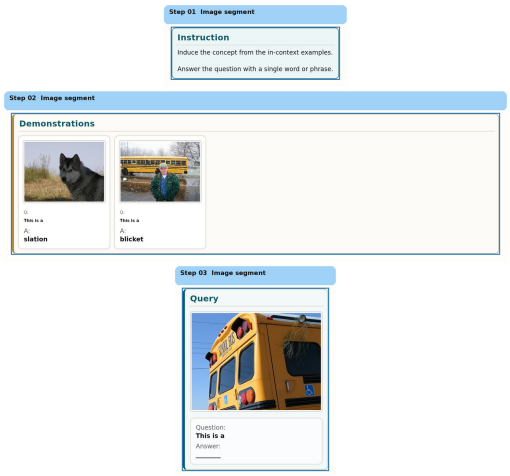
clevr: *Role* (1 shot)



open mi: Baseline (1 shot)

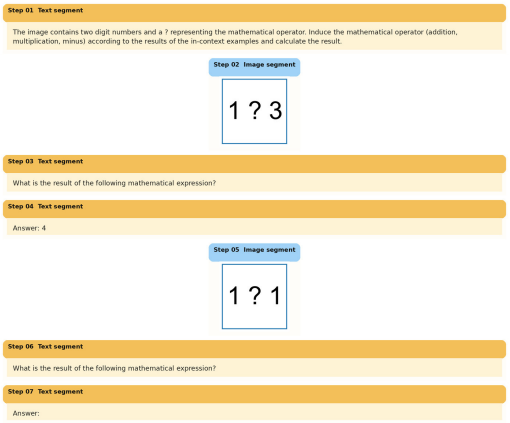


open mi: *Role* (1 shot)

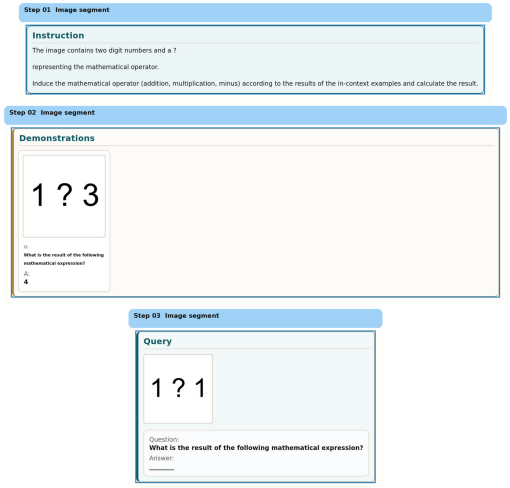


VL-ICL (continued)

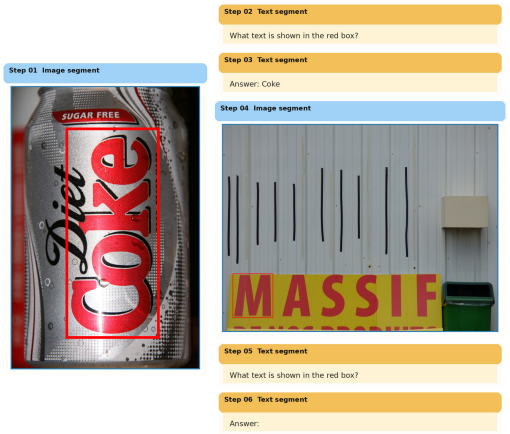
operator induction: Baseline (1 shot)



operator induction: *Role* (1 shot)



textocr: Baseline (1 shot)

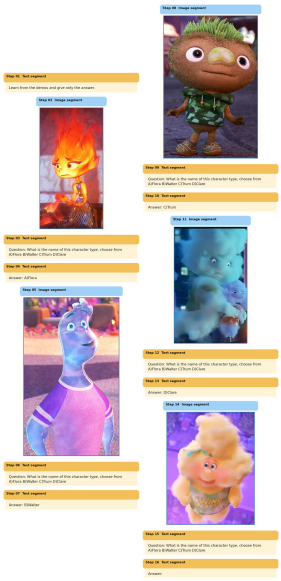


textocr: *Role* (1 shot)

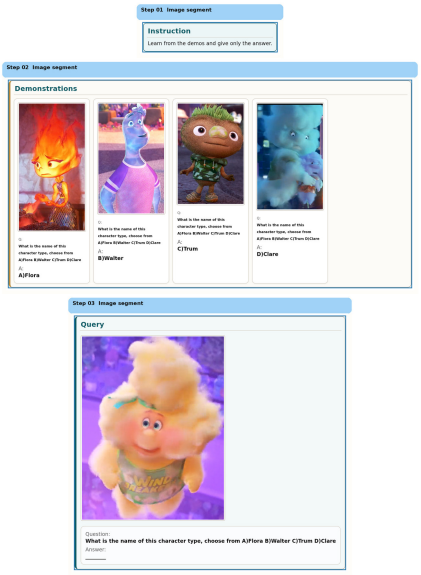


E.3 TRUEMICL

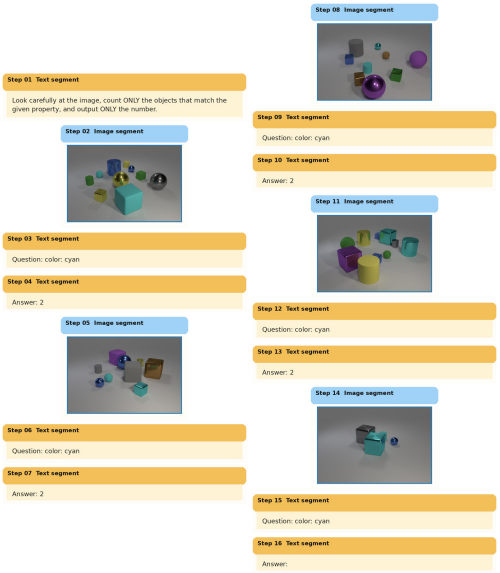
classification: Baseline (4 shot)



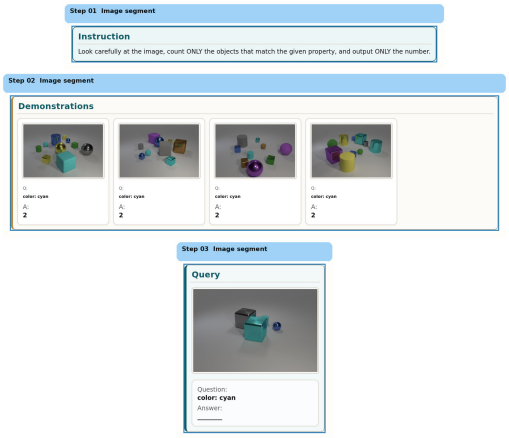
classification: *Role* (4 shot)



clevr: Baseline (4 shot)

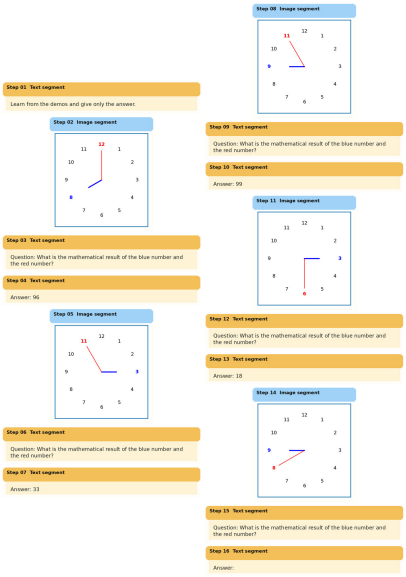


clevr: *Role* (4 shot)

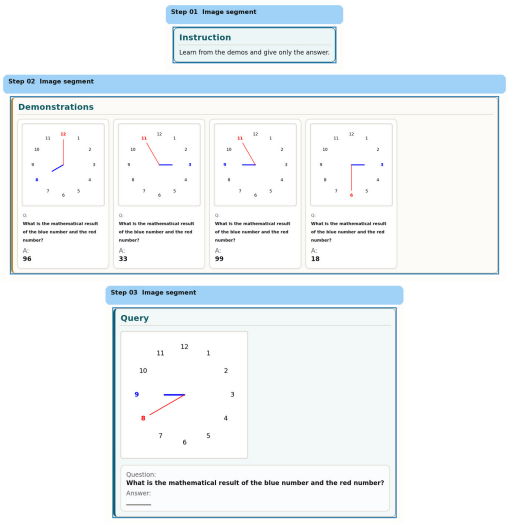


TrueMICL (continued)

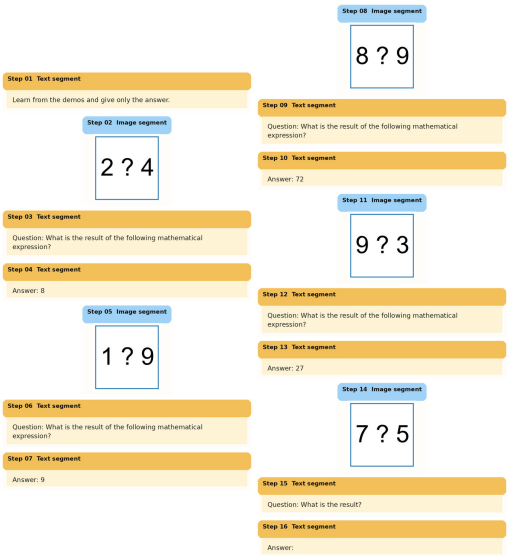
clock: Baseline (4 shot)



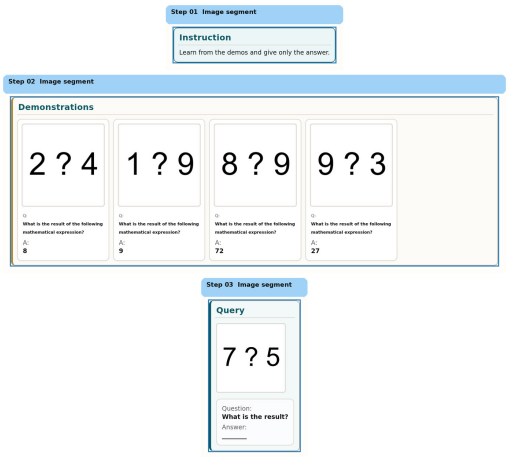
clock: *Role* (4 shot)



operator induction: Baseline (4 shot)

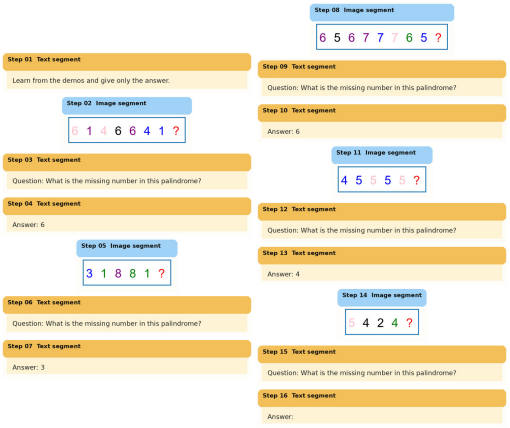


operator induction: *Role* (4 shot)

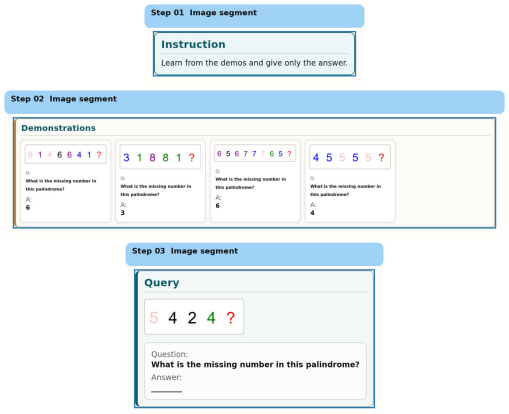


TrueMICL (continued)

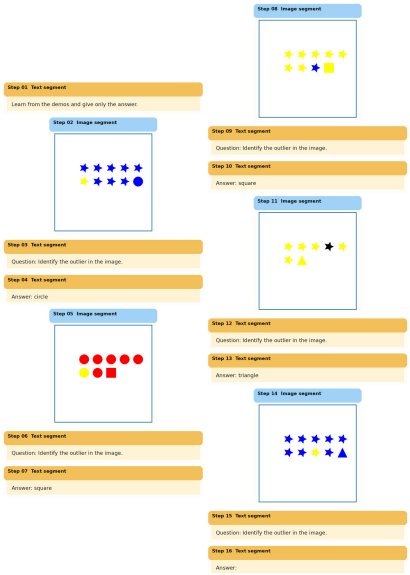
palindrome: Baseline (4 shot)



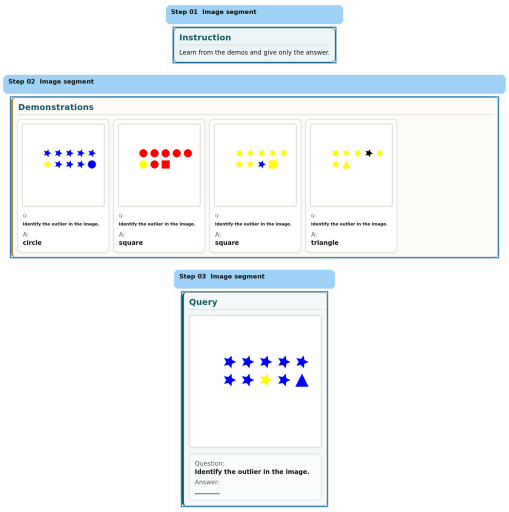
palindrome: *Role* (4 shot)



shapes count: Baseline (4 shot)

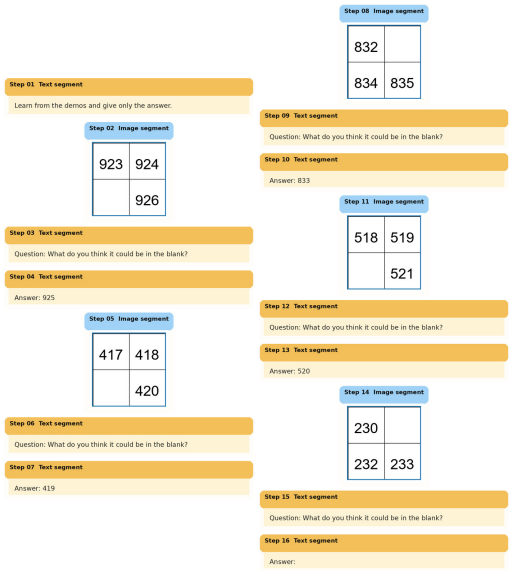


shapes count: *Role* (4 shot)

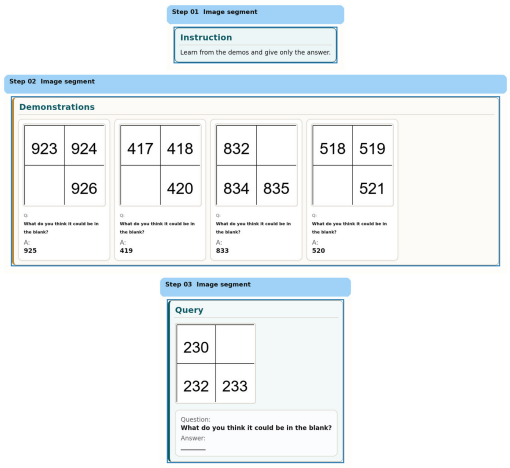


TrueMICL (continued)

sudoku: Baseline (4 shot)

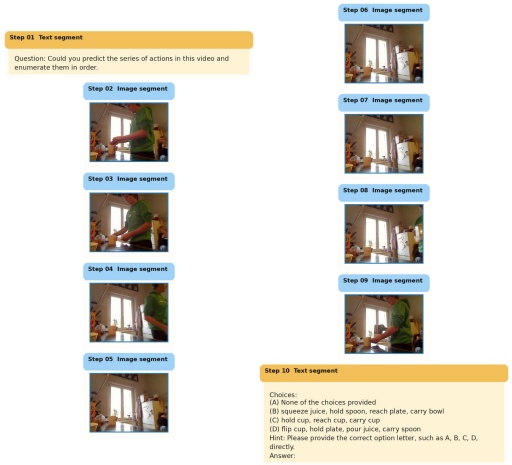


sudoku: *Role* (4 shot)

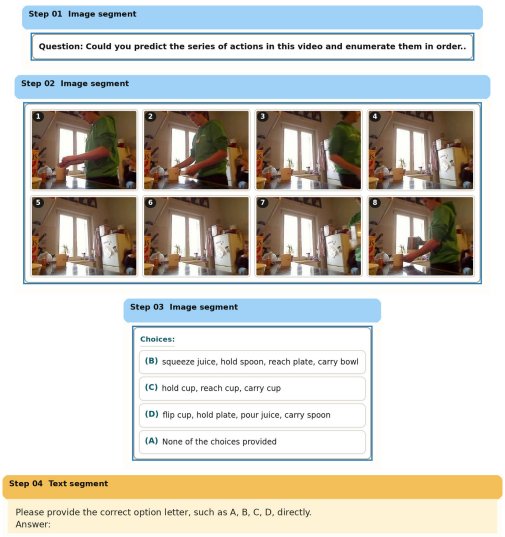


E.4 MUIRBENCH

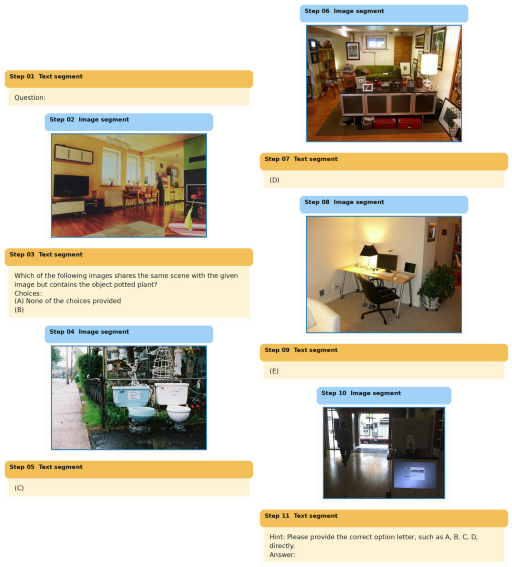
action understanding: Baseline



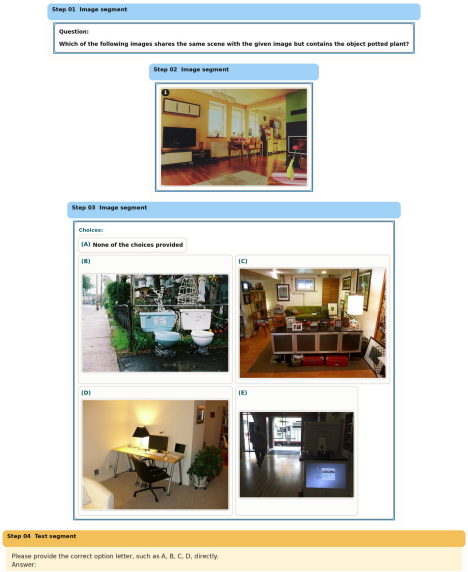
action understanding: *Role*



attribute similarity: Baseline



attribute similarity: *Role*



MuirBench (continued)

cartoon understanding: Baseline

Step 01 Text segment

Question: What are the jokes of these pictures?

Step 02 Image segment



Step 03 Image segment



Step 04 Text segment

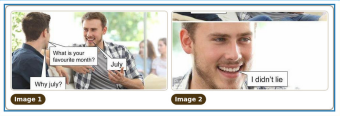
Choices:
(A) None of the choices provided
(B) The man in the plaid shirt heard "why july" as "why you lie".
(C) The man in the plaid shirt lied.
(D) The man in the plaid shirt looks funny.
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

cartoon understanding: *Role*

Step 01 Image segment

Question: What are the jokes of these pictures?

Step 02 Image segment



Step 03 Image segment

Choices:
(A) None of the choices provided
(B) The man in the plaid shirt heard "why july" as "why you lie".
(C) The man in the plaid shirt lied.
(D) The man in the plaid shirt looks funny.

Step 04 Text segment

Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

counting: Baseline

Step 01 Text segment

Question: How many other garments besides a complete mitten pair are shown in each image?

Step 02 Image segment



Step 03 Image segment



Step 04 Text segment

Choices:
(A) Four
(B) Two
(C) Three
(D) None of the choices provided
(E) Zero
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

counting: *Role*

Step 01 Image segment

Question: How many other garments besides a complete mitten pair are shown in each image?

Step 02 Image segment



Step 03 Image segment

Choices:
(A) 4
(B) 2
(C) 3
(D) None of the choices provided
(E) 0

Step 04 Text segment

Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

MuirBench (continued)

diagram understanding: Baseline

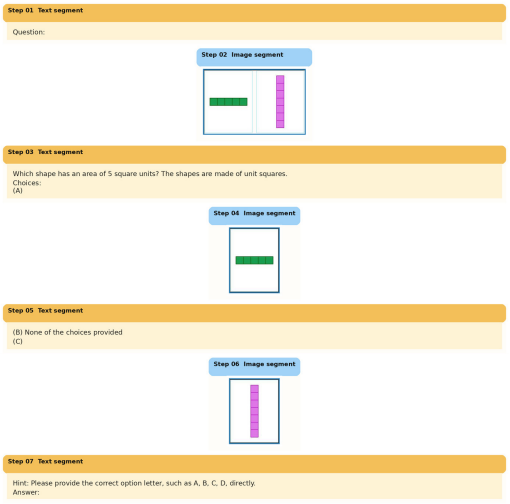
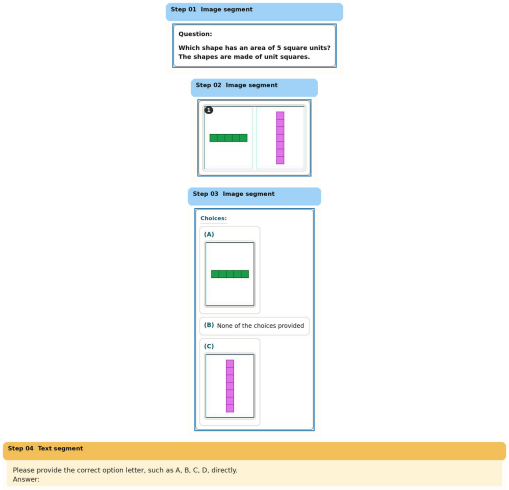
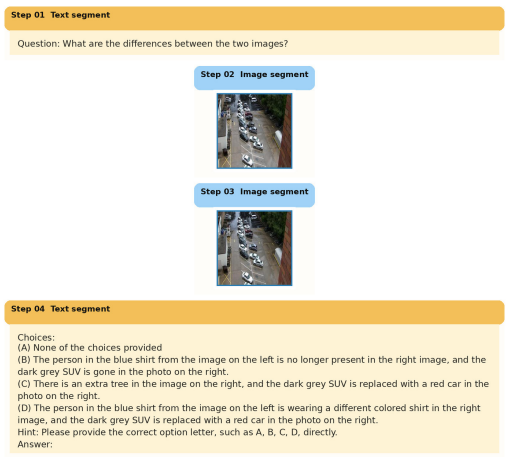


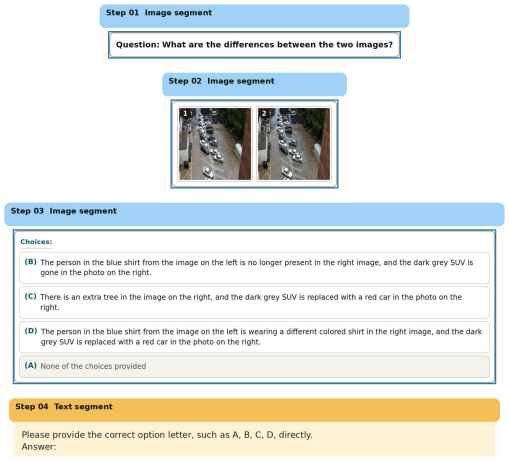
diagram understanding: *Role*



difference spotting: Baseline

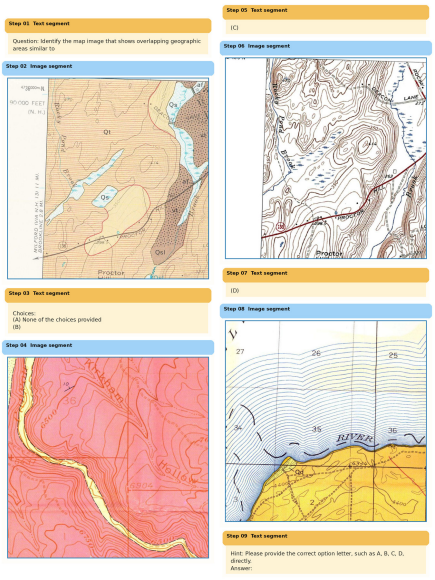


difference spotting: *Role*



MuirBench (continued)

geographic understanding: Baseline



geographic understanding: *Role*

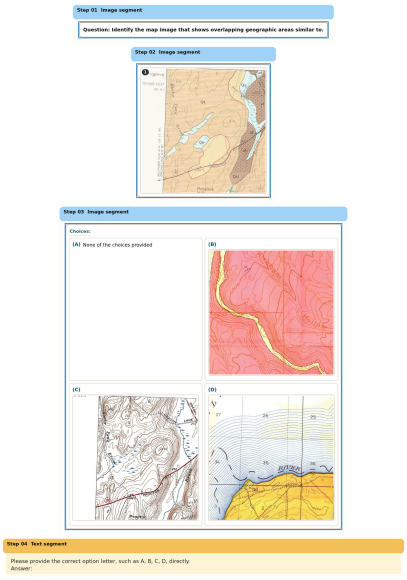


image text matching: Baseline

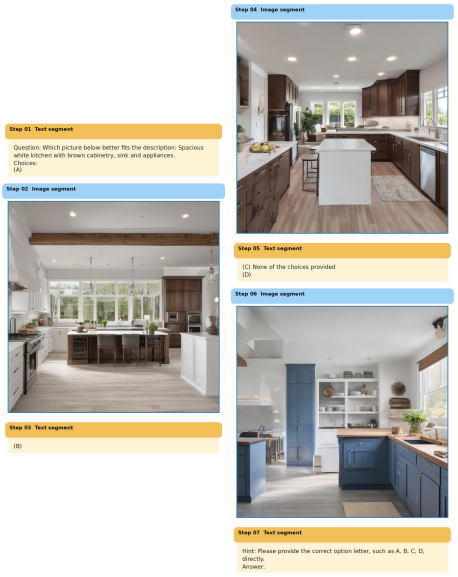
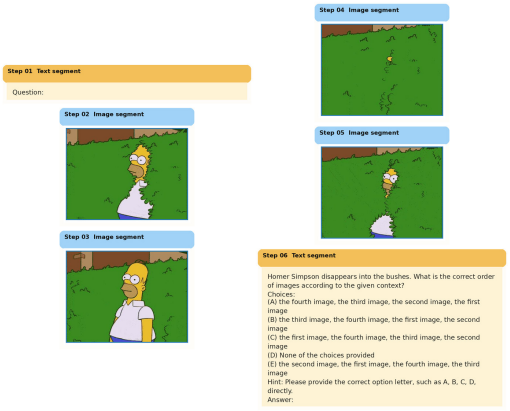


image text matching: *Role*

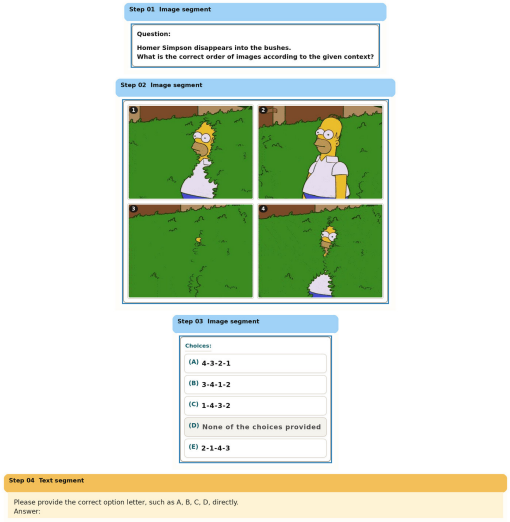


MuirBench (continued)

ordering: Baseline



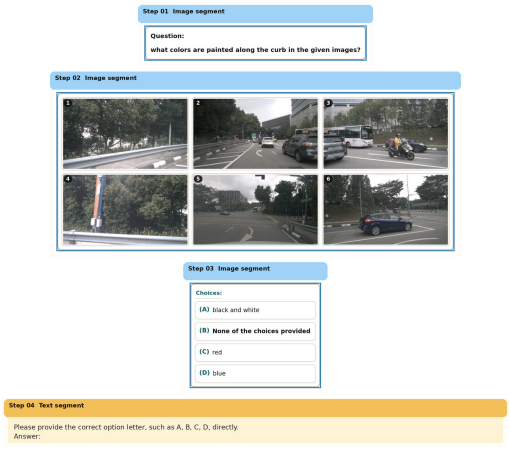
ordering: *Role*



scene understanding: Baseline

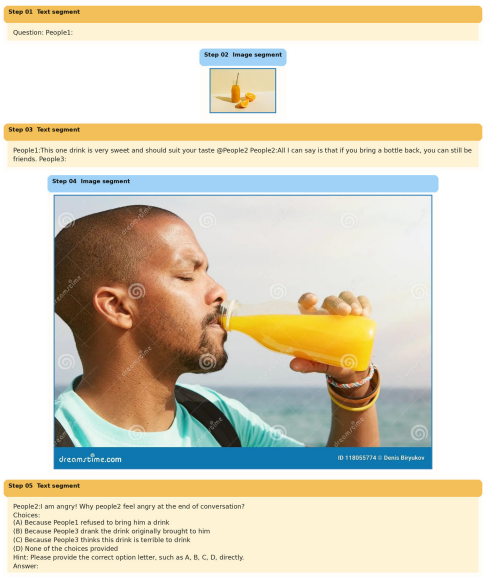


scene understanding: *Role*



MuirBench (continued)

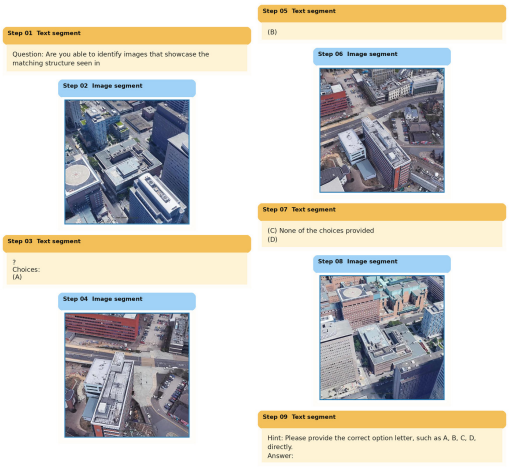
visual grounding: Baseline



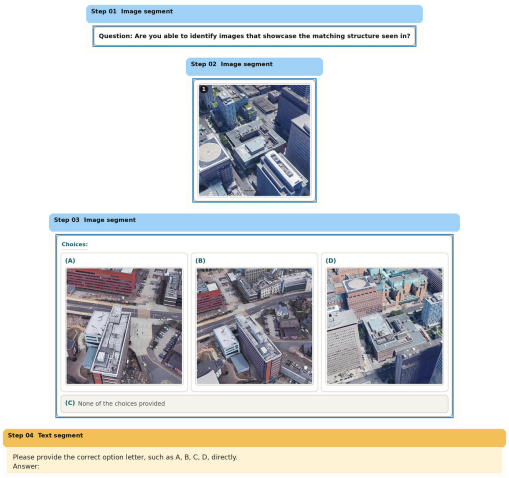
visual grounding: *Role*



visual retrieval: Baseline



visual retrieval: *Role*



E.5 ZERO-SHOT RESPONSES

ImageQuery: Baseline (0 shot)

Step 01 Text segment

Instruction
Match the query pattern to patterns a, b, c, d, and e shown in the reference. The images contain synthetic concentric shapes. Classes are defined by the base shape and the ordered colors of the nested layers. Some tasks may describe the query pattern in text instead of showing a query image.

Step 02 Text segment

Query

Step 03 Image segment

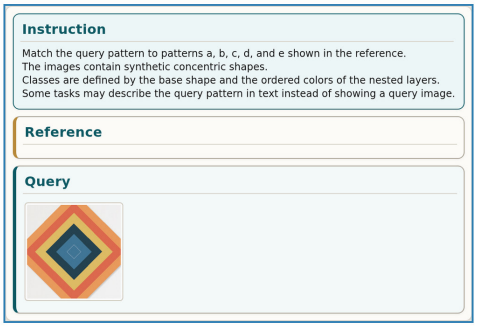


Step 04 Text segment

Answer:

ImageQuery: *Joint* (0 shot)

Step 01 Image segment



Step 02 Text segment

Answer:

TextQuery: Baseline (0 shot)

Step 01 Text segment

Instruction
Match the query pattern to patterns a, b, c, d, e shown in the reference. The images are synthetic concentric shapes. Classes are defined by shape type and the ordered ring colors. Some tasks may describe the query pattern in text instead of showing a query image.

Step 02 Text segment

Query

Step 03 Text segment

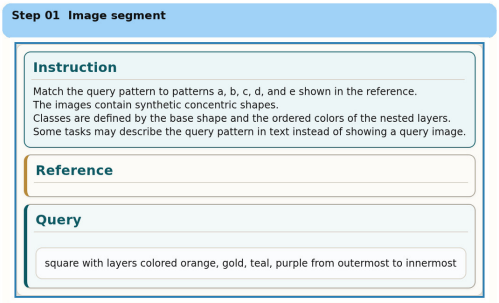
square with rings colored orange, gold, teal, purple from outermost to innermost

Step 04 Text segment

Answer:

TextQuery: *Joint* (0 shot)

Step 01 Image segment

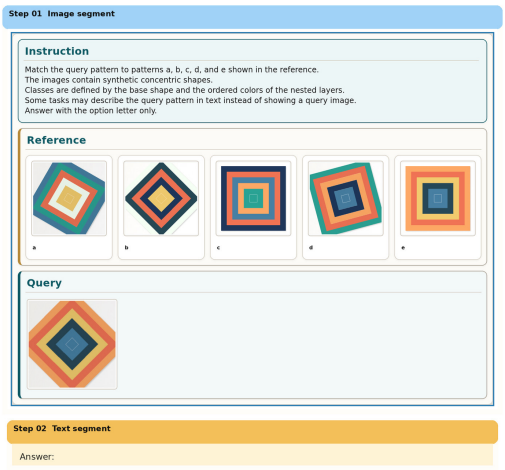


Step 02 Text segment

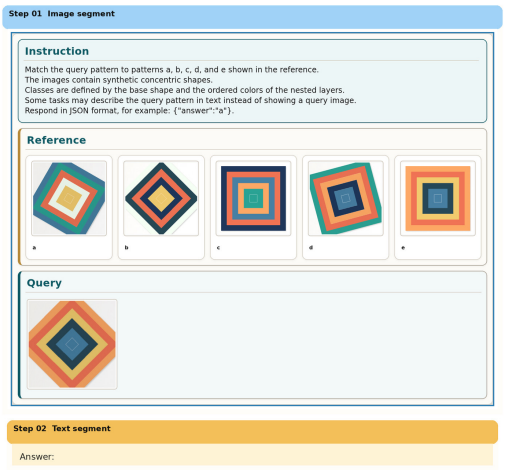
Answer:

E.6 IMAGE INSTRUCTIONS

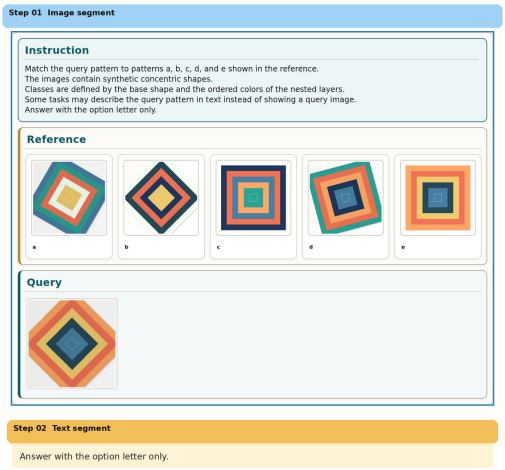
ImageQuery: Image letter (1 shot)



ImageQuery: Image JSON (1 shot)



ImageQuery: Congruent letter (1 shot)



ImageQuery: Image letter and external JSON (1 shot)

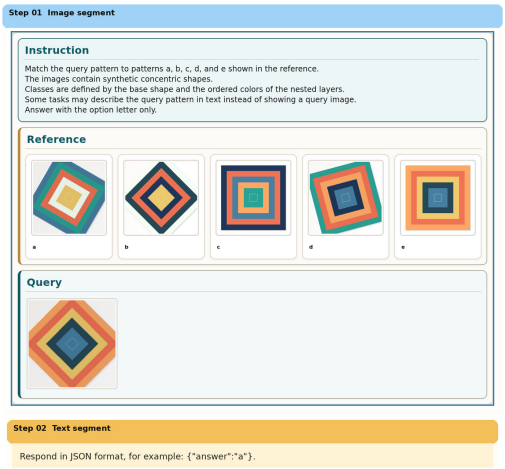
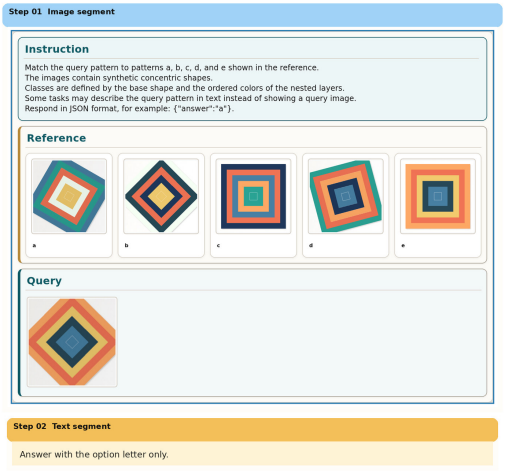
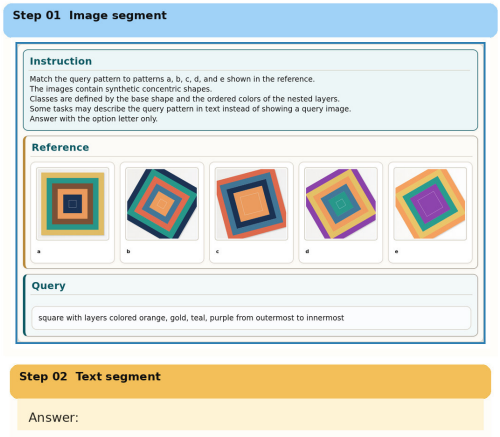


Image instructions (continued)

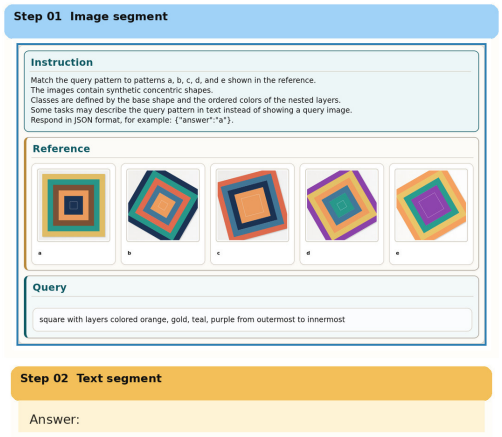
ImageQuery: Image JSON and external letter (1 shot)



TextQuery: Image letter (1 shot)



TextQuery: Image JSON (1 shot)



TextQuery: Congruent letter (1 shot)

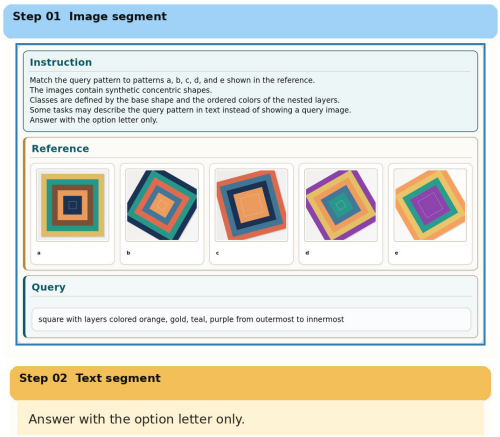
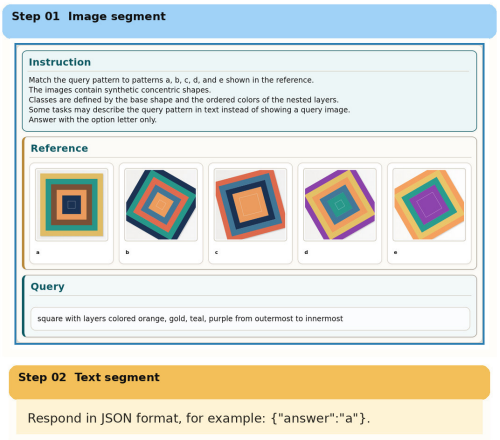
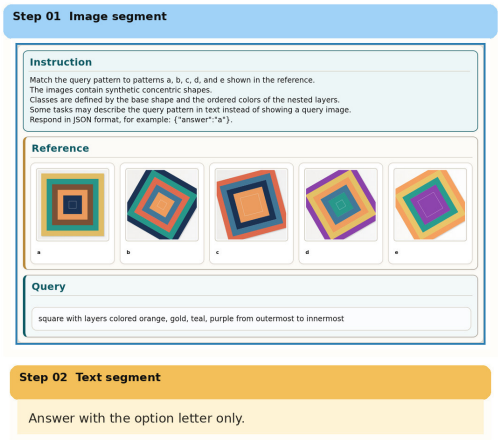


Image instructions (continued)

TextQuery: Image letter and external JSON (1 shot)

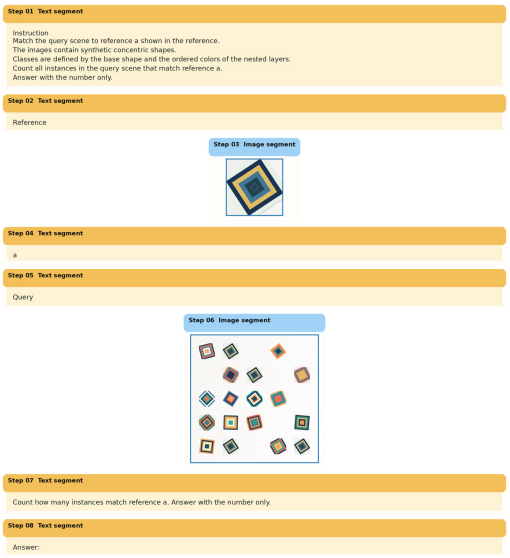


TextQuery: Image JSON and external letter (1 shot)



E.7 COUNTING

Baseline (1 shot)



Joint (1 shot)



E.8 CAPTION-ONLY CLASSIFICATION

102Flowers (1 shot)

Step 01 Text segment

Choose which reference option describes the same fine-grained category as the query.
Use only the supplied visual descriptions. Answer with one option letter only.
References:
a: The central object is a star-shaped bloom with pointed, overlapping petals radiating outward. Each petal is elongated, tapering to a sharp tip, with a gradient from deep red at the base to bright orange and pale pink at the edges. The surface appears smooth and slightly glossy, with subtle veining. At the center, small, clustered structures emerge, resembling tiny buds or stamens, surrounded by tightly packed, reddish bracts. The bloom is set against broad, green, lance-shaped leaves.
b: A delicate, conical inflorescence rises vertically, composed of overlapping, pointed pink bracts that taper to soft tips. The bracts exhibit a gradient from pale pink at the edges to deeper pink toward the center, with subtle greenish-yellow bases. A small, white and purple flower emerges from the lower bracts, nestled within green, cup-shaped structures. The surface appears smooth and slightly waxy, with fine veining visible on the bracts.
c: A vibrant, deep red inflorescence rises prominently, composed of overlapping, pointed bracts arranged in a dense, conical spiral. Each bract has a smooth, waxy texture with subtle veining and slightly curled tips. The structure tapers toward the top, with a pale green base transitioning into the red. Surrounding broad, glossy green leaves with prominent veins frame the bloom, enhancing its striking contrast.
d: The central object features a tall, pointed, pink structure with overlapping, ridged segments resembling a spiky cone. Its surface is smooth and glossy, with a gradient from deep pink at the base to lighter pink at the tip. Surrounding it are two delicate, purple-blue petals with ruffled edges, speckled with glistening water droplets. The petals appear soft and slightly translucent, contrasting with the firm, upright pink form. Green, slender leaves radiate outward, adding a vibrant backdrop.
e: A vibrant, heart-shaped structure in glossy, deep pink-red dominates the image. Its surface is smooth and waxy, with subtle, flowing ridges radiating from the base. Emerging from the center is a thick, pale pink, finger-like appendage with a textured, bumpy surface. The edges of the heart-shaped form are slightly curled, and the background features broad, green leaves with visible veins.
Query:
A vibrant, upright inflorescence displays a dense, conical cluster of overlapping, pointed bracts in deep crimson to magenta hues. Each bract is smooth, slightly glossy, and tapers to a sharp tip, arranged in a tight spiral. The base transitions into a pale pinkish stem, emerging from broad, green, veined leaves. Sunlight highlights the bracts' waxy texture and subtle gradations in color.
Answer:

CUB (1 shot)

Step 01 Text segment

Choose which reference option describes the same fine-grained category as the query.
Use only the supplied visual descriptions. Answer with one option letter only.
References:
a: A small bird perches on a slender, grayish branch. Its head is black with a sharp, conical orange beak. The face has a white patch behind the eye, blending into a grayish-white throat and chest. The back and wings are mottled brown and black with fine streaks. The belly is pale, and the tail is long and dark, slightly fanned. Feathers appear soft and slightly ruffled.
b: A small bird perches on a weathered wooden surface, its body compact and rounded. The head features a warm reddish-brown crown, transitioning to gray on the cheeks and throat, with a dark eye and short, pointed black beak. The back and wings display intricate patterns of brown, black, and buff, with fine streaks. The underparts are soft gray, and the legs are slender and orange. The tail is short and slightly fanned.
c: A small bird perches on a weathered wooden post, its body compact and upright. It has a pale grayish-white breast, buffy flanks with fine dark streaks, and a brownish back with subtle barring. The head features a distinct black-and-white striped crown, a pale supercilium, and a dark eye-line. Its beak is short, conical, and pale pinkish-brown. The legs are slender and pinkish, gripping the rough, splintered wood. The tail is short and slightly rounded, with pale edges.
d: A small bird perches on a dark, textured branch. Its head features a bold black-and-white striped crown with a bright yellow patch above the eye. The back and wings display intricate brown and black streaks over a warm buff base. The underparts are pale grayish-white. Fine, pointed feathers form a short tail, and delicate pink legs grip the branch. The eye is dark and alert, surrounded by a pale gray face.
e: A small bird perches on a weathered wooden surface, its body compact and slightly angled. It has a grayish-brown back with fine, dark streaks and warm, buffy underparts. The wings show rich brown feathers with darker barring. A striking black crown stripe contrasts with a white supercilium and pale gray face. Its short, conical beak is yellowish-orange, and its legs are pinkish. The tail is long and pointed, blending into the dark background.
Query:
A small bird perches on a weathered, gray surface. Its body is compact, with mottled brown and gray feathers on the back and wings, showing fine, dark barring. The underparts are pale grayish-white. A bold black stripe runs through the eye, bordered above by white, creating a distinct facial pattern. The beak is short, conical, and yellowish-orange. Legs are slender and pinkish-brown. Tail is long and slightly rounded, blending into the brown plumage.
Answer:

Food-101 (1 shot)

Step 01 Text segment

Choose which reference option describes the same fine-grained category as the query.
Use only the supplied visual descriptions. Answer with one option letter only.
References:
a: The main object consists of several golden-brown, elongated, cylindrical fried items arranged in a small pile on a rectangular white plate. They have a slightly rough, textured surface with visible oil sheen, suggesting crispiness. Their ends are rounded, and they vary slightly in thickness. A small white bowl beside them contains a chunky, dark brown sauce with visible bits, possibly fruit or spice.
b: A tortilla chip, irregularly triangular with a golden-yellow hue and speckled surface, is held at an angle. Its edge is dipped into a thick, chunky green dip, revealing a coarse texture with visible flecks of red and dark green. The dip fills a shallow, rounded bowl with a glossy, slightly uneven rim, showing smears and residue along the inner walls.
c: The main object is a golden-brown, crescent-shaped pastry with a flaky, porous interior. Its surface is dusted with fine white powder, likely sugar. One end is dipped into a dark, glossy liquid in a small metallic cup, creating a shiny, wet coating. The pastry's curved edge is slightly irregular, with a soft, airy texture visible where broken. The overall shape is elongated and gently arched.
d: A golden-brown, circular waffle with a grid of raised square pockets covers most of the white plate. Its surface is slightly uneven, with a crisp, textured exterior and soft, pale edges. A dollop of creamy, off-white butter sits centered, slightly melting into the waffle's crevices, creating a glossy sheen. The waffle's edges are slightly curled and irregular.
e: A blue-rimmed white plate holds irregularly shaped, golden-yellow tortilla chips with a slightly rough, speckled texture. A small black ramekin with red salsa sits beside them. The plate rests on a colorful, glossy surface with green, yellow, and red stripes. A knife and fork lie on a dark napkin in the background.
Query:
The main object is a white, circular ceramic plate with a slightly raised rim. On it rest four golden-brown, elongated, slightly curved sticks arranged in a cross shape. Their surface is textured with fine ridges and speckled with seasoning. A small, round, metallic cup with a white rim holds a smooth, orange-brown sauce. The plate sits on a dark wooden surface.
Answer:

Cars (1 shot)

Step 01 Text segment

Choose which reference option describes the same fine-grained category as the query.
Use only the supplied visual descriptions. Answer with one option letter only.
References:
a: A silver-gray SUV with a boxy, rugged design is positioned on a reddish-brown rocky incline. It features a seven-slot chrome grille, round headlights with clear lenses, and fog lights below. The body has black cladding along the lower edges and wheel arches. Chrome door handles and side mirrors contrast with the metallic paint. The wheels are multi-spoke alloy with black tires. Snow-capped mountains appear in the background.
b: A silver SUV with a boxy, angular design is shown in a three-quarter rear view. It features a glossy metallic finish, black trim along the lower body, and five-spoke alloy wheels. The rear includes vertical taillights, a spare tire mounted under the tailgate, and a small antenna. The windows are tinted, and the side mirrors are body-colored. The vehicle sits on a reflective dark surface with a gradient blue-black background.
c: A dark gray SUV is shown from a rear three-quarter view, parked on asphalt with mountains in the background. It features a boxy, robust shape with a slightly sloped rear window, chrome roof rails, and silver door handles. The vehicle has five-spoke alloy wheels, tinted windows, and a chrome rear bumper with a tow hitch. The tail lights are red with clear lenses, and a chrome badge is centered on the tailgate. A California license plate is visible.
d: A beige SUV with a smooth, metallic finish is shown from a rear three-quarter view. It features a boxy shape with rounded edges, tinted rear windows, and a roof rack. Chrome door handles and side moldings contrast with the body. The rear includes a horizontal taillight bar, a centered emblem, and a license plate. Silver multi-spoke wheels and black tires are visible. The vehicle rests on a dirt path with mountains in the background.
e: A large, olive-gray SUV is shown in motion, with a glossy, metallic finish. It features a prominent front grille with horizontal slats, clear headlamps, and silver multi-spoke alloy wheels. The body has smooth, rounded contours with black-tinted windows, roof rails, and chrome door handles. The side mirrors are body-colored, and the rear includes a visible red taillight. The vehicle's surface reflects light, indicating a polished texture.
Query:
A dark metallic gray SUV is shown from a rear three-quarter view, parked on asphalt. It features a boxy, robust shape with a raised roofline and silver roof rails. The rear has a prominent tailgate with a chrome emblem and red taillights. Chrome door handles and side steps contrast with the body. Multi-spoke silver alloy wheels with black tires are visible. The side window trim is black, and a "LIMITED" badge is on the rear door. The background shows hazy hills and distant city lights.
Answer:

Caption-only classification (continued)

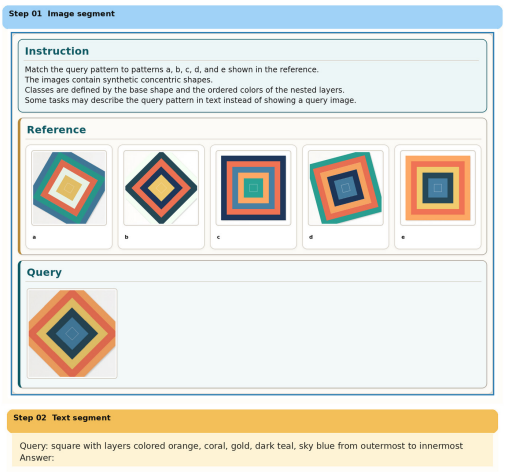
Dogs (1 shot)

Step 01 Text segment

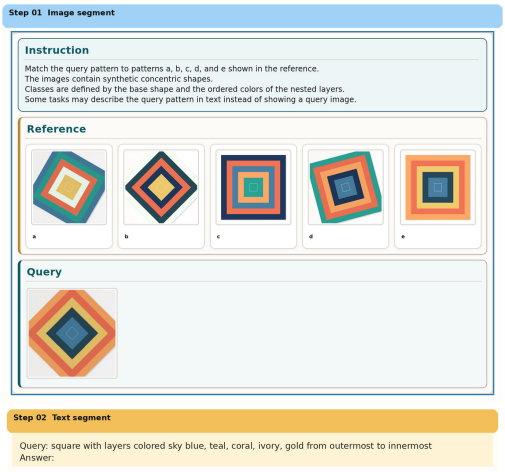
Choose which reference option describes the same fine-grained category as the query.
Use only the supplied visual descriptions. Answer with one option letter only.
References:
a: A close-up of a dark, glossy-coated animal with a broad, rounded head and soft, floppy ears that hang down on either side. Its eyes are large, round, and warm amber-brown, reflecting light with a subtle shine. The nose is wide, black, and moist-looking, with visible nostrils. Fine, dark whiskers extend from the muzzle. The fur appears short, smooth, and uniformly black, with slight sheen under warm lighting. The background is a blurred wooden floor and beige wall.
b: A fluffy, off-white creature with a dense, slightly tousled coat lies on a textured beige surface. Its face is round with dark, expressive eyes and a small, black, moist nose. Ears are pointed, upright, and covered in soft fur, revealing pink inner linings. The muzzle is short, with a hint of darker fur around the mouth. The fur appears soft and slightly wavy, with subtle shading from light to cream.
c: Two fluffy white animals lie side by side on a beige carpet. Their fur is long, wavy, and dense, with slight off-white or cream shading around the face and paws. One has upright, pointed ears with pink inner linings; the other has floppy, rounded ears. Both have dark, round eyes and black noses. One holds a small pink and blue object in its mouth. Paws are broad with dark pads.
d: A black, shaggy-coated animal with upright, pointed ears and a dense, wiry fur texture. Its face is framed by long, dark fur, obscuring the lower jaw and eyes, which appear dark and deep-set. A plaid collar with a round, metallic tag hangs around its neck. The body is covered in thick, curly black fur, resting on a patterned surface with a soft, cream-colored fabric in the background.
e: A small, fluffy creature with golden-tan fur lies on a red, patterned fabric. Its fur is soft and slightly wavy, with longer tufts around the face and ears. The ears are upright, triangular, and fringed with hair. Dark, round eyes and a small black nose are prominent. A small, golden tag hangs from its collar. Its front paws are extended forward, and its body curves gently to the side.
Query:
The main object is a black, shaggy-coated animal with a dense, wiry texture, seen from behind and slightly to the side. Its silhouette is compact and rounded, with a thick, fluffy tail curving upward. The ears are small, pointed, and erect, blending into the dark fur. The head is slightly tilted upward, and the back is arched, suggesting alertness. The fur appears uniformly black with subtle variations in shadow and light, giving it a textured, layered look.
Answer:

E.9 QUERY CONFLICT

Congruent (1 shot)

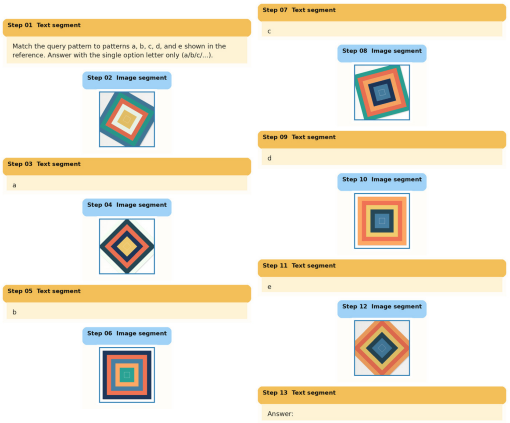


Conflict (1 shot)

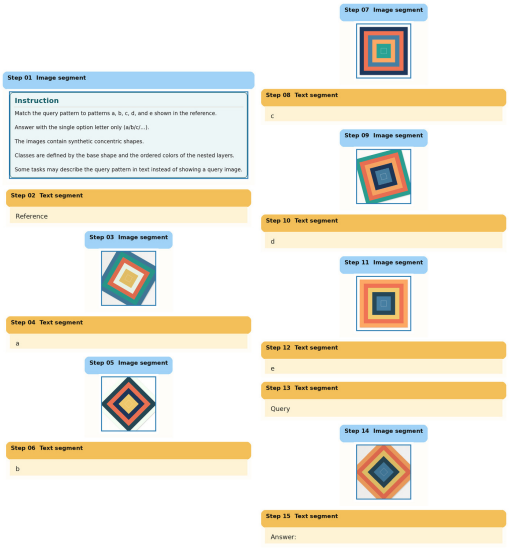


E.10 CONCENTRIC COMPOSITIONS

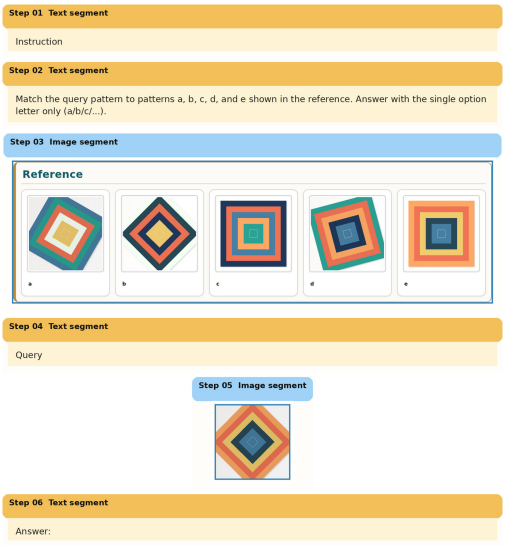
ImageQuery: empty (1 shot)



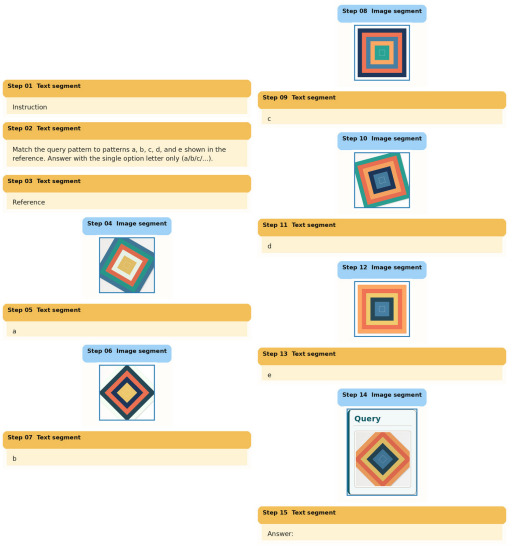
ImageQuery: I (1 shot)



ImageQuery: R (1 shot)

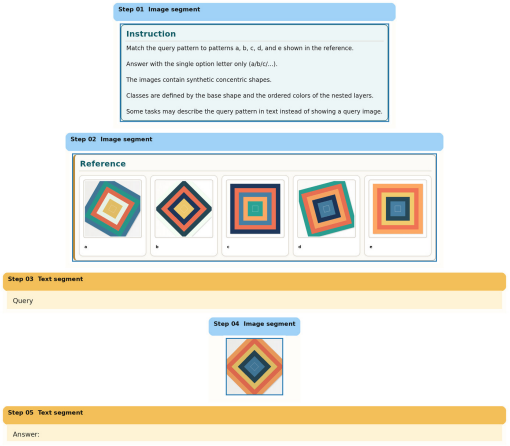


ImageQuery: Q (1 shot)

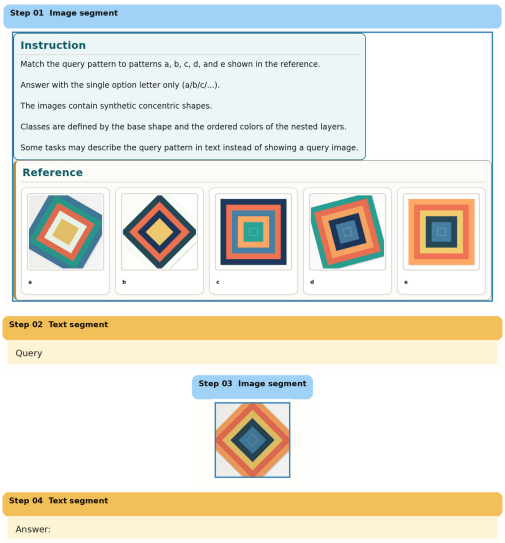


Concentric compositions (continued)

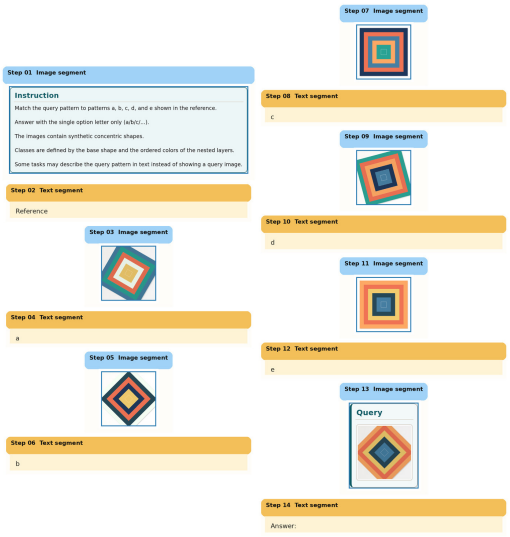
ImageQuery: I|R (1 shot)



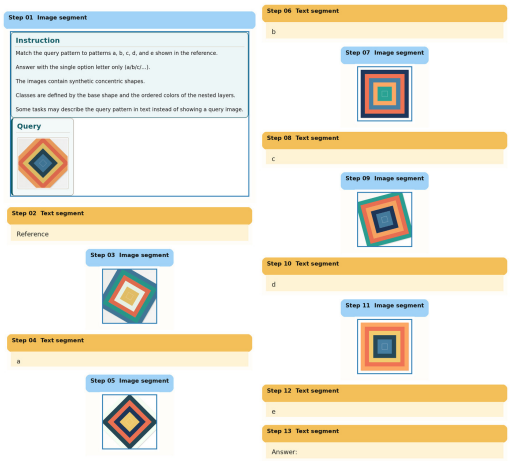
ImageQuery: IR (1 shot)



ImageQuery: I|Q (1 shot)

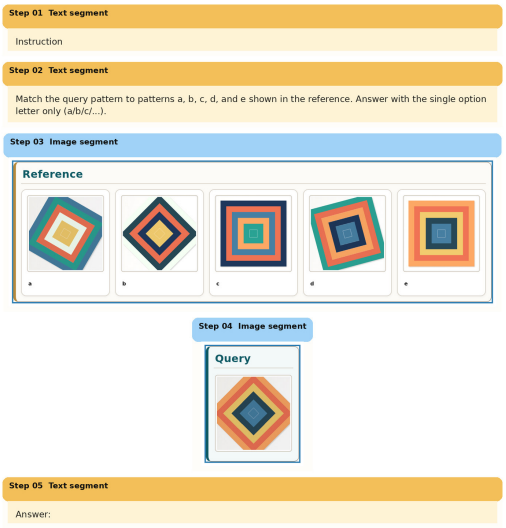


ImageQuery: IQ (1 shot)

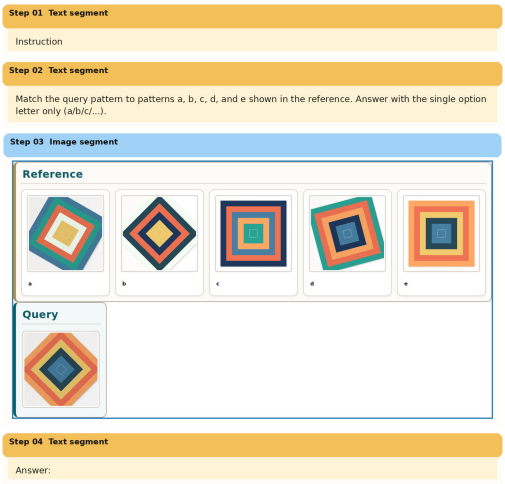


Concentric compositions (continued)

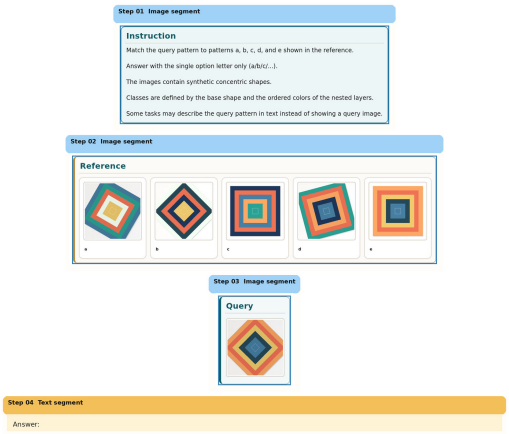
ImageQuery: R|Q (1 shot)



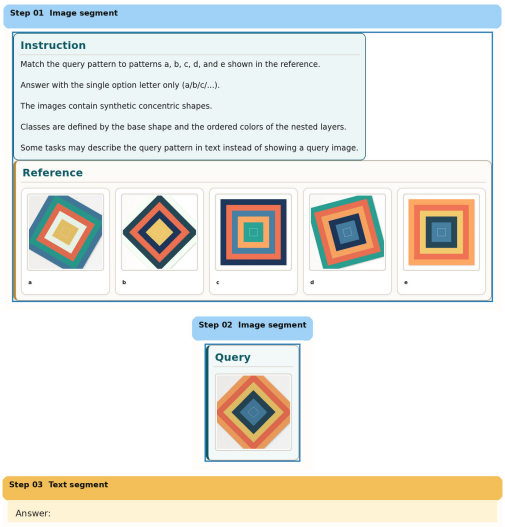
ImageQuery: RQ (1 shot)



ImageQuery: I|R|Q (1 shot)

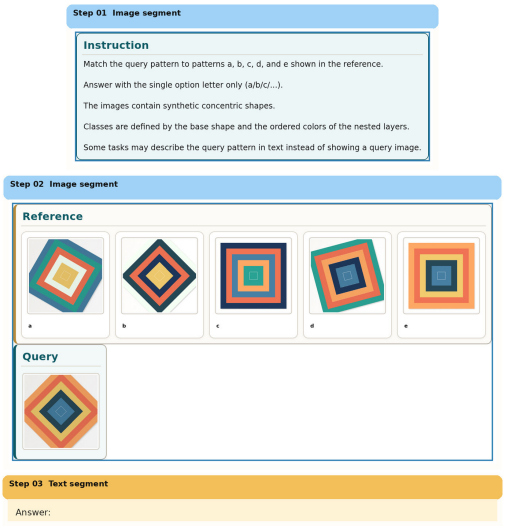


ImageQuery: IR|Q (1 shot)

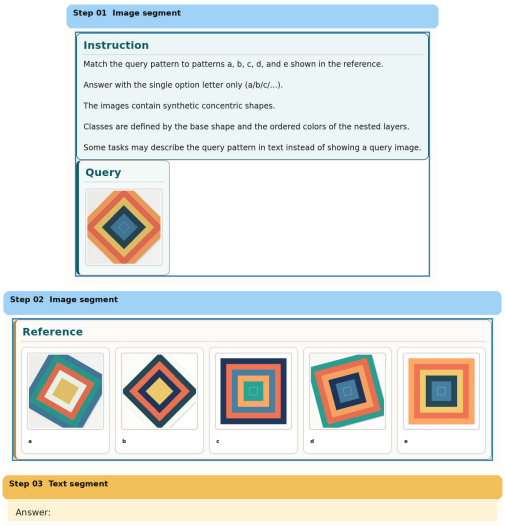


Concentric compositions (continued)

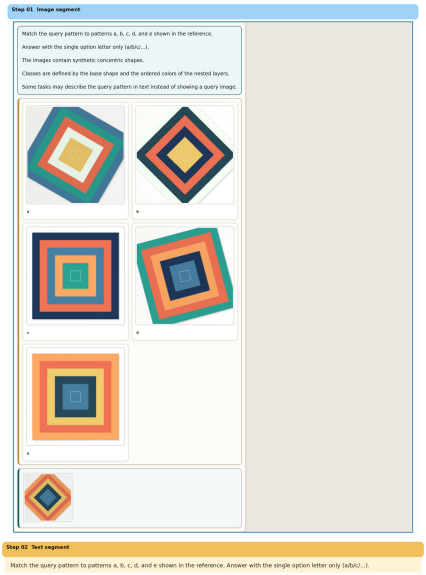
ImageQuery: I|RQ (1 shot)



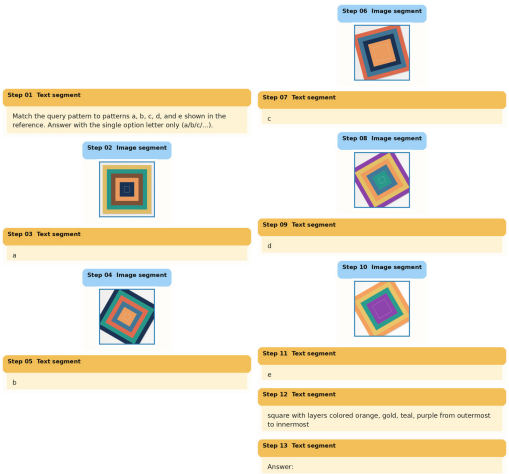
ImageQuery: IQ|R (1 shot)



ImageQuery: IRQ (1 shot)

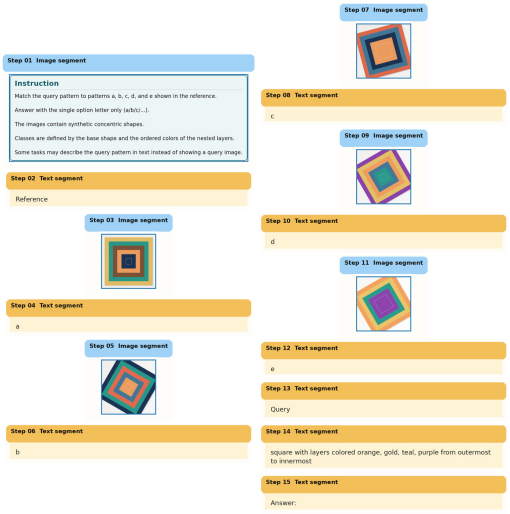


TextQuery: empty (1 shot)

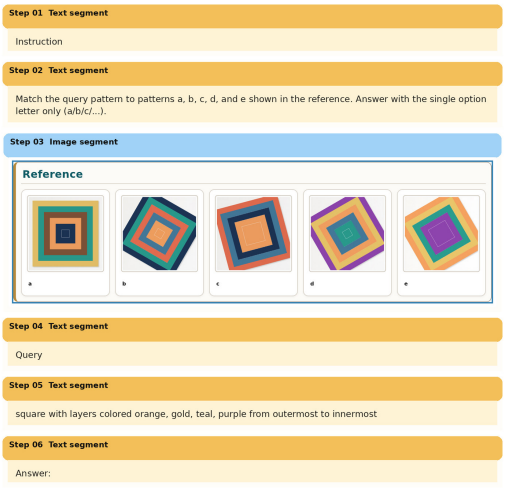


Concentric compositions (continued)

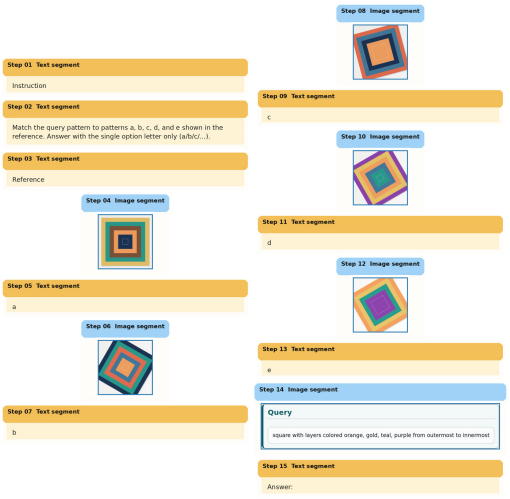
TextQuery: I (1 shot)



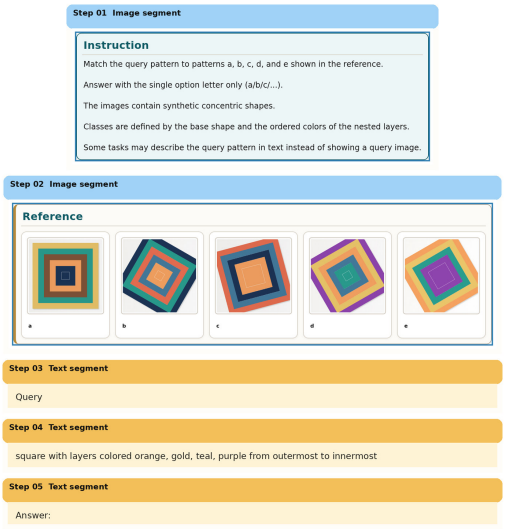
TextQuery: R (1 shot)



TextQuery: Q (1 shot)

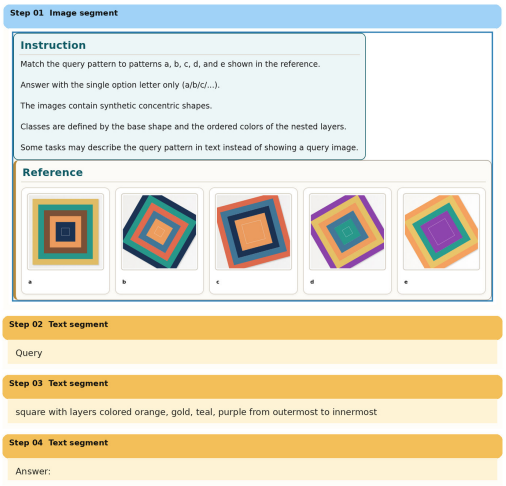


TextQuery: I|R (1 shot)

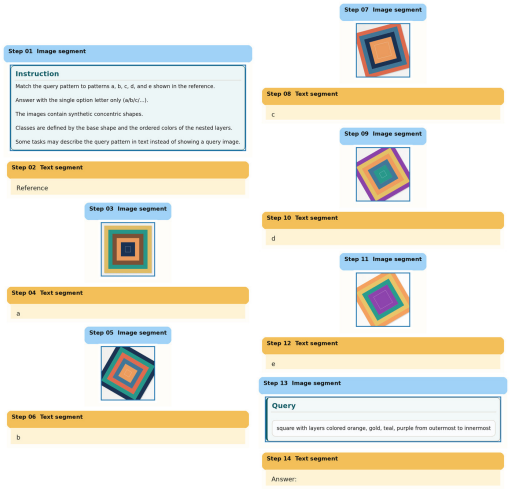


Concentric compositions (continued)

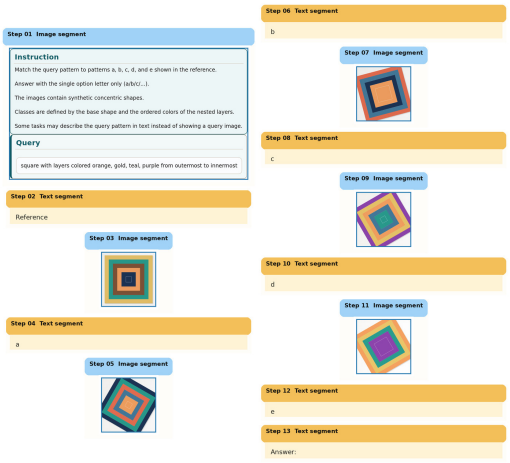
TextQuery: IR (1 shot)



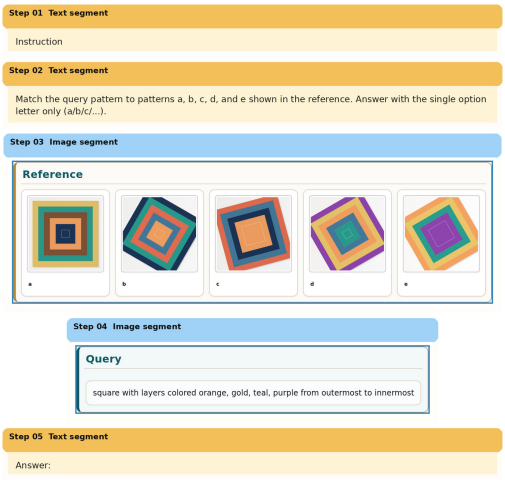
TextQuery: I|Q (1 shot)



TextQuery: IQ (1 shot)

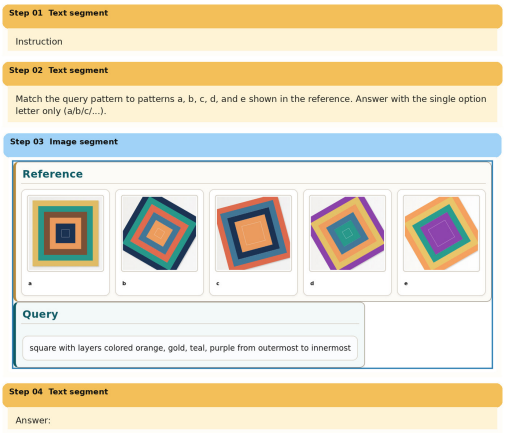


TextQuery: R|Q (1 shot)

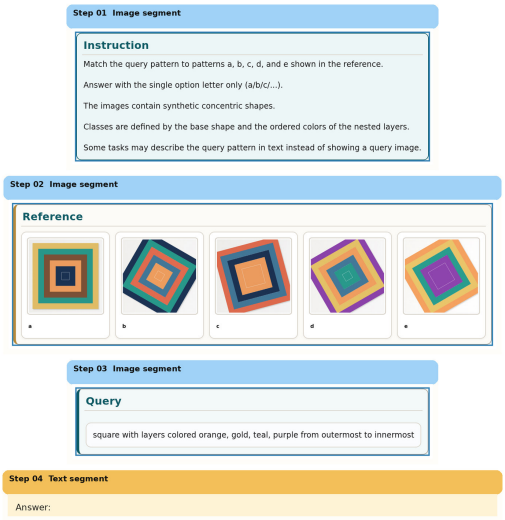


Concentric compositions (continued)

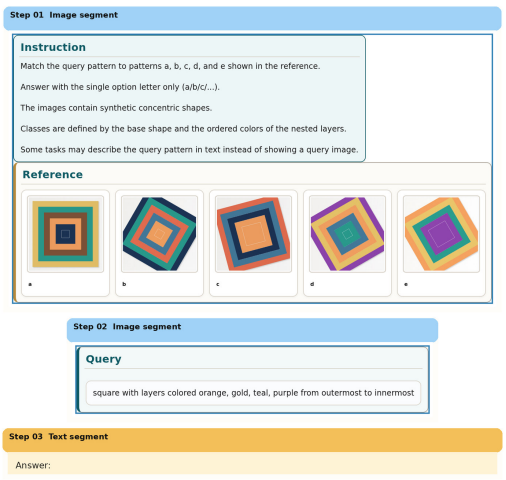
TextQuery: RQ (1 shot)



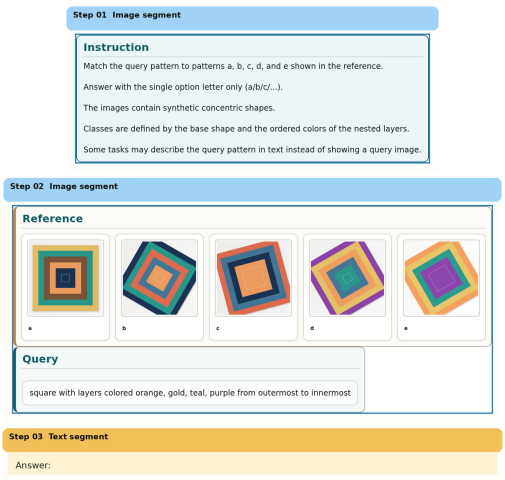
TextQuery: I|R|Q (1 shot)



TextQuery: IR|Q (1 shot)

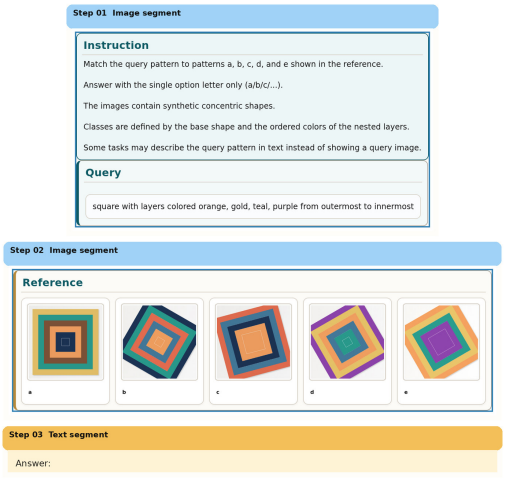


TextQuery: I|RQ (1 shot)

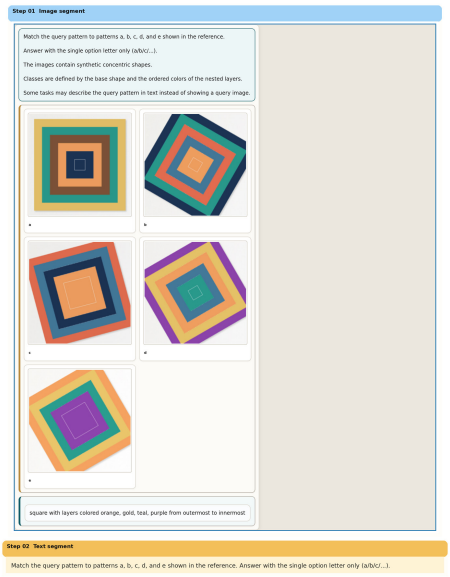


Concentric compositions (continued)

TextQuery: IQ|R (1 shot)

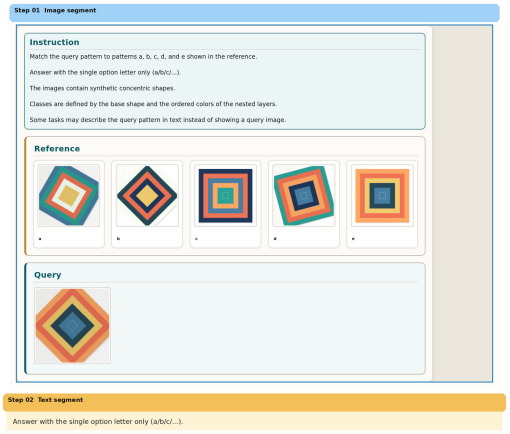


TextQuery: IRQ (1 shot)



E.11 *Joint* MASK AND DROP

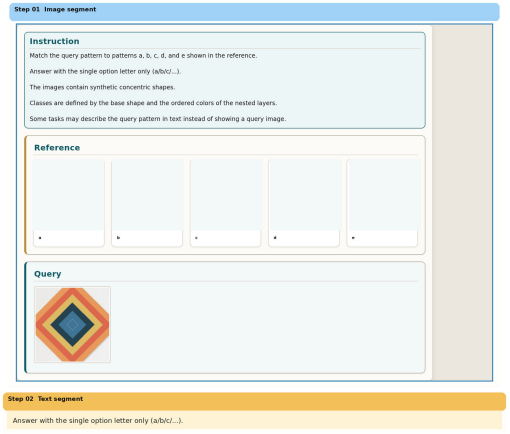
ImageQuery: full (1 shot)



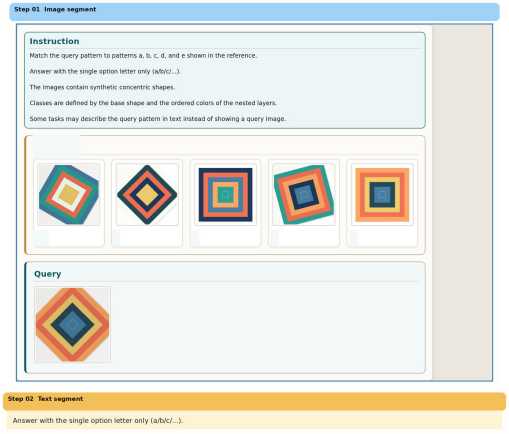
ImageQuery: mask target instruction text (1 shot)



ImageQuery: mask target reference image (1 shot)

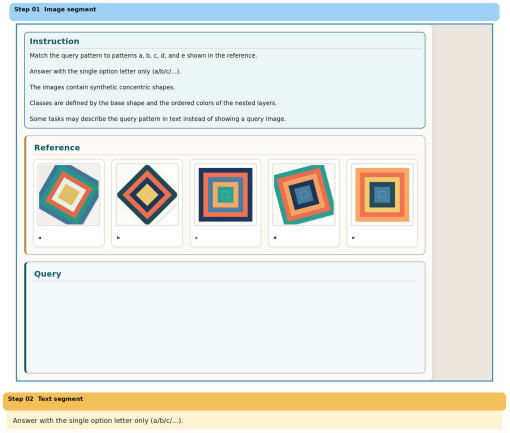


ImageQuery: mask target class label text (1 shot)

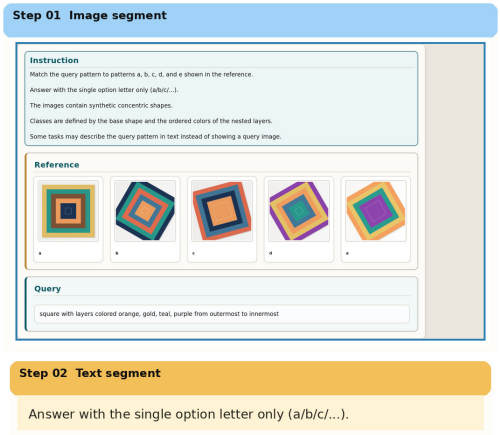


Joint mask and drop (continued)

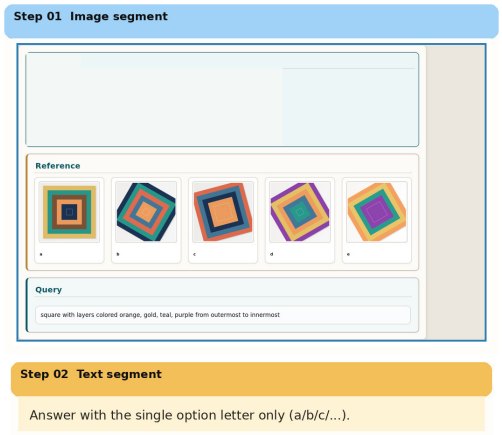
ImageQuery: mask target query image (1 shot)



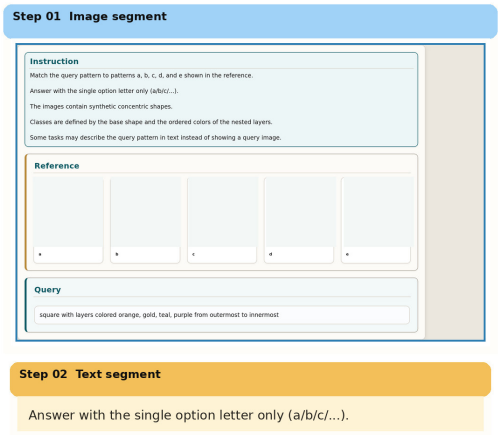
TextQuery: full (1 shot)



TextQuery: mask target instruction text (1 shot)

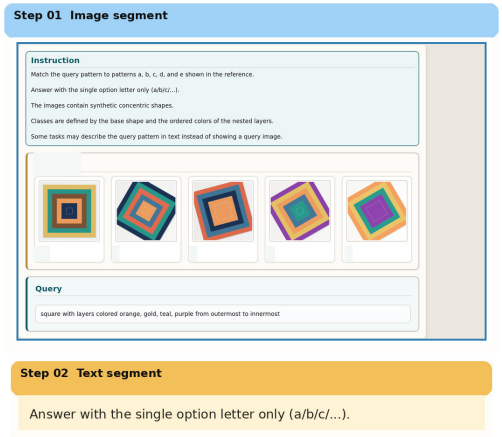


TextQuery: mask target reference image (1 shot)



Joint mask and drop (continued)

TextQuery: mask target class label text (1 shot)



TextQuery: mask target query text (1 shot)

