VISANALOG: A Diagnostic Suite for Visual Concept Transfer on Natural Images

Zhaonan Li[1], Kyle R. Chickering[2,†], Bangzheng Li[3], Jacob Dineen[1], Xiao Ye[1], Zhikun Xu[1], Shijie Lu[1]
Yuxi Huang[1], Ming Shen[1], Bach Nguyen[1], Jaya Adithya Pavuluri[1], Mau Son Nguyen[1]
Sanika Chavan[1], Ngoc Minh Thu Le[1], Muhao Chen[3], Ben Zhou[1]

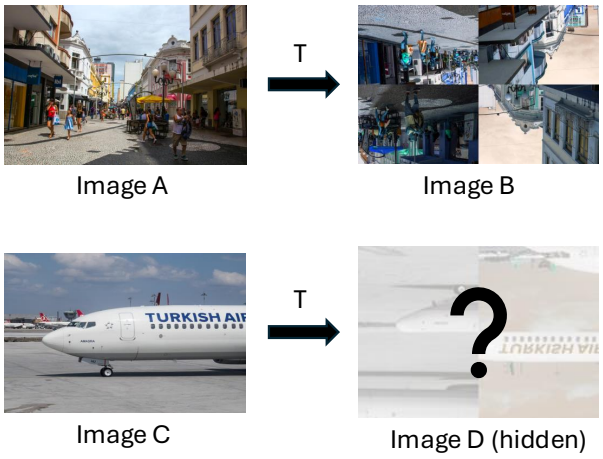
[1]Arizona State University [2]Luma AI [3]UC Davis

Abstract

A useful test of visual concept learning is not just whether a model can recognize a concept in a single image, but whether it can preserve and manipulate concept-level properties under transformation and transfer them to new scenes. We introduce VISANALOG*, a controlled suite for this setting on natural images. Each example instantiates $A : B :: C :?$: images B and a hidden target image D are produced by applying the same deterministic transformation sequence to source images A and C. Given A, B, and C, a model must answer a multiple-choice question about D. The benchmark contains 617 human-validated questions spanning one- to four-step transformations such as zoom, quadrant swap, rotation, flip, and hue rotation. Across strong proprietary and open-source VLMs, end-to-end accuracy is substantially lower than oracle accuracy when D is directly shown, and degrades sharply as transformation depth increases, while human performance remains near the ceiling. A program-conditioned evaluation further separates failures of relation inference from failures of transformation application, showing that inferring the visual relation from $A \rightarrow B$ is the dominant bottleneck, with additional application errors emerging on harder multi-step cases. The dataset is publicly available at* https://huggingface.co/datasets/zli99/VisAnalog.

1. Introduction

A central goal of visual concept learning is to understand whether models can use visual concepts and relations beyond one-shot recognition. If a multimodal model has learned meaningful visual concepts, it should not only recognize them in a single image, but also track how visual properties and relations change under transformation and transfer those changes to new instances. This capability matters for visual reasoning, scene understanding, and controllable generation, where success depends on preserving and manipulating visual information rather than merely describing what is directly visible. We study this capability through a controlled visual analogy task on natural images.



Question: What are the characteristics of the large text visible on the main aircraft's fuselage?

Options:
A: The text is blue-colored, appears upside down and horizontally mirrored, and is in the bottom-right quadrant.
B: The text is gold-colored, appears upside down and horizontally mirrored, and is in the bottom-right quadrant.
...

Figure 1. An example of the visual analogy questions for our benchmark. In these questions, source images A and C undergo the same transformation T, a sequence of image-editing operations, to become images B and D, respectively. During evaluation, the model is provided with images A, B, and C, and is tasked with answering a question about the hidden image D. To solve the question, the model must infer the abstract visual relationships between A and B, then apply them to C to mentally visualize D. The question assesses specific visual details in D that depend on accurately inferring the transformed image. In this example, the question focuses on the color, orientation, and position of the text.

[†]Work done while at UC Davis.

Each example instantiates an analogy $A : B :: C :?$, where B and a hidden target image D are produced by applying the same deterministic transformation sequence to two source images A and C. The model receives A, B, and C, but not D, and must answer a question about D. To succeed, it must infer the relation expressed by $A \rightarrow B$, transfer that relation to C, and reason about the resulting hidden image. We use *visual concepts* to refer to controllable image properties and *visual relations* to refer to deterministic transformations over those properties. VISANALOG tests whether models can use such concepts and relations in a way that supports manipulation, composition, and transfer, rather than only recognizing them in a single image. Unlike ARC- or RAVEN-style analogy benchmarks built on synthetic patterns, our setting uses natural images, requiring models to identify and track these relations in realistic scenes while preserving precise control over the underlying transformation. Across strong proprietary and open-source VLMs, performance is substantially higher when the true target image D is directly visible than when it must be inferred, and end-to-end accuracy declines sharply as transformation depth increases, even while human performance remains near ceiling.

Contributions. We make three contributions: (1) we introduce VISANALOG, a controlled benchmark for testing concept-level visual transfer under transformation on natural images; (2) we show that strong VLMs struggle on this task, especially under multi-step composition; and (3) we introduce a program-conditioned diagnostic that makes the benchmark more interpretable by separating failures of relation inference from failures of transformation application.

2. Related Work

Vision-centric evaluation of VLMs. Vision benchmarks for multimodal models span several settings. Knowledge-centric VQA benchmarks test whether models can identify scene content and combine it with external or world knowledge [8, 14, 18, 19, 22]. Relational and perception-focused benchmarks instead emphasize properties that must be read from the image itself, including spatial relations, orientation, and perceptual judgments [6, 10, 21]. These works show that modern VLMs often struggle even when the relevant evidence is directly present in the input image. Our benchmark is complementary: rather than asking about visible evidence alone, we ask models to reason about a *hidden transformed target* derived from an observed image pair.

Analogical reasoning and relational transfer. Analogical reasoning has long been studied as the transfer of relational structure rather than surface similarity [7], with links to visuospatial simulation and mental rotation in cognitive science [12, 17, 20]. In language models, analogical ability has been shown to emerge in some settings, but remains brittle when structure must be induced and transferred rather than recalled [9, 23, 27]. Visual analogy benchmarks such as ARC, ConceptARC, and RAVEN similarly probe abstract relational transfer, but typically do so in synthetic or symbolic domains [4, 15, 16, 29]. We view VISANALOG as complementary to these benchmarks: instead of abstract symbolic patterns, we study controlled transformation-based transfer in natural images.

Multimodal models and concept use under transformation. Recent unified multimodal models increasingly support both visual understanding and generation within a single architecture [1, 3, 5, 24, 25]. At the same time, recent evaluations continue to find major weaknesses in abstract, spatial, and transformation-based reasoning [2, 13, 26, 28, 30]. Our work differs from open-ended generation benchmarks and broad capability surveys in two ways. First, we use a deterministic transformation family, which makes the underlying relation explicit and interpretable. Second, we focus on whether visual information is processed in a way that supports compositional transfer and downstream reasoning. In this sense, VISANALOG is best viewed not as a general benchmark of visual reasoning, but as a controlled probe of how multimodal models use visual concepts under relational transformation.

3. VISANALOG: Task and Construction

Task definition. We study a narrow form of visual analogy: *visually grounded relational transfer*. Each example instantiates an analogy of the form $A : B :: C :?$, where B and a hidden target image D are obtained by applying the same transformation sequence T to two source images A and C, respectively. At test time, the model is given A, B, and C, but not D, and must answer a multiple-choice question about D. To succeed, it must infer the relation expressed by $A \rightarrow B$, transfer that relation to C, and reason about the resulting hidden image.

We view this setup as a *controlled concept probe* rather than a comprehensive benchmark of visual reasoning. The goal is to test whether models can preserve and manipulate concept-level visual properties and relations—such as color, orientation, and spatial position—under controlled transformations in natural scenes.

Controlled transformation family. We restrict T to a fixed family of deterministic image edits: (i) centered zoom (an 80% crop resized back to the original resolution), (ii) quadrant swap (exchanging two tiles in a 2×2 grid), (iii) counterclockwise rotation ($90°$, $180°$, or $270°$), (iv) horizontal or vertical flip, and (v) hue rotation ($90°$, $180°$, or $270°$). Any example may contain a subset of these operations, with each operation used at most once. When multiple operations are present, they are applied in a fixed order. This keeps the task narrow but well specified: the underlying transformation is explicit, ambiguity from edit ordering is reduced, and difficulty can be controlled by the number of composed steps.

Source images and instance generation. We construct examples from natural images sampled from SA-1B [11], favoring images with sufficient resolution and scene complexity. For each instance, we sample two source images A and C, generate a random transformation sequence T, and produce $B = T(A)$ and $D = T(C)$ using PIL. The use of natural images places these controlled transformations in visually rich scenes, while the synthetic edit family keeps the task programmatically precise.

Question generation. For each pair (C, D), we generate one multiple-choice question about D. Questions are written to target visual consequences of the transformation sequence. The question, ground-truth answer, and distractor options are proposed by Gemini 2.5 Pro, conditioned on C, D, and the ground-truth transformation sequence. Distractors are designed to reflect plausible partial failures, such as omitting one step or misapplying the transformation order. This design ensures success depends on correctly transferring the visual relation, rather than on language priors alone.

Quality control. We apply a two-stage filtering process. First, annotators verify that the transformation relating A and B is uniquely identifiable, removing ambiguous cases such as near-symmetries or repeated quadrants. Second, annotators verify that the question has a unique correct answer given accurate reasoning about D. We retain only items accepted by two annotators on both criteria. To reduce residual annotation errors in answerability, we then apply a lightweight oracle-solvability check with two open-source VLMs, Qwen2.5-VL-72B and Llama-3.2-90B-Vision, using the question together with oracle D, and keep only items solved by at least one model. We use this step only to remove clearly problematic items, not to argue that the benchmark is easy or saturated. The final benchmark contains 617 questions spanning one- to four-step sequences.

4. Experiments

Evaluation settings. We report three settings. **End-to-end analogy** is the main task: the model receives A, B, and C, and answers a multiple-choice question about hidden D. **Oracle VQA** provides the ground-truth target image D directly and measures whether the question is answerable given correct visual evidence. **Wrong-image control** replaces D with C and tests whether success can be explained by generic priors or superficial cues instead of reasoning about the transformed target.

Model selection. We report a representative set of strong proprietary and open-source VLMs, spanning frontier closed models (GPT-o3, Gemini-2.5-Flash, Gemini-2.5-Pro) and strong open models (Qwen2.5-VL-72B, InternVL3.5-38B-Instruct, Qwen3-VL-8B-Instruct).

Human baseline. We treat the human result as a sanity-check estimate rather than a high-precision benchmark. Following the benchmark protocol, we randomly sample 40 questions (10 per step count), ask two annotators to solve them independently with access to A, B, C, and the transformation family, and adjudicate disagreements to form consensus labels.

Model	Step1	Step2	Step3	Step4
Human	100.0	90.0	90.0	90.0
GPT-o3	77.4	66.0	49.0	30.7
Gemini-2.5-Flash	67.3	59.7	51.6	24.0
Gemini-2.5-Pro	71.2	49.7	48.4	29.3
Qwen2.5-VL-72B	51.8	39.6	43.3	52.0
InternVL3.5-38B-Instruct	51.3	32.1	21.7	26.7
Qwen3-VL-8B-Instruct	58.4	34.0	29.3	30.7

Table 1. End-to-end analogy accuracy (%) on VISANALOG.

4.1. Main benchmark results

Strong models still struggle with transformation-based visual analogy. Table 1 reports representative end-to-end results. Humans remain near ceiling across all step counts, while strong closed and open models are substantially worse and, for most models, degrade as the number of composed transformations increases. This trend is especially clear for GPT-o3 and the Gemini models: for example, GPT-o3 drops from 77.4% on 1-step questions to 30.7% on 4-step questions, while humans remain at 90–100% across all four bins. Overall, the table shows that transferring a visually specified relation and then reasoning about the hidden result is much harder than ordinary single-image question answering.

The benchmark requires transferring transformed visual concepts, not guessing from the source image. Figure 2 provides the key control. Questions are explicitly written to target properties and relations that are *true in* D, must follow from the transformation sequence, and are *not answerable from* C *alone*; distractors are designed to reflect plausible misapplications such as omitted steps or wrong order. Consistent with that design, models perform below chance in the wrong-image control, where they are asked the question using the source image C instead of the hidden target D. In contrast, when the ground-truth target image D is directly provided, oracle accuracy remains high across models, often above 80%. Together, these controls show that the task is not solved by generic priors or by reading off concepts already visible in the source image. Instead, success requires *transferring* concept-level visual information through the inferred transformation and then reasoning over the resulting hidden target.

The main difficulty is the analogy gap, and it widens with composition. Taken together, Table 1 and Figure 2 isolate the core difficulty of VISANALOG. The questions are answerable when correct visual evidence is available, but not from the untransformed source image, which means

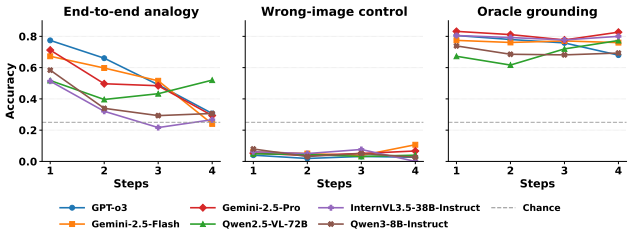


Figure 2. Comparison of end-to-end analogy solving, wrong-image control, and oracle VQA. Models perform much better when the true target image D is visible than when they must infer it, while the wrong-image control stays near or below chance. Dashed lines indicate 25% random-guessing accuracy.

Model	Metric	S1	S2	S3	S4
InternVL3.5-38B	$\mathrm{Acc}(C,T)$	70.8	70.4	62.4	53.3
	Δ_{inf}	19.5	38.4	40.8	26.7
	Δ_{app}	9.7	8.8	15.3	26.7
Qwen3-VL-8B	$\mathrm{Acc}(C,T)$	72.1	67.9	61.8	52.0
	Δ_{inf}	13.7	34.0	32.5	21.3
	Δ_{app}	1.8	0.6	6.4	17.3

Table 2. Program-conditioned evaluation. $\mathrm{Acc}(C,T)$ gives accuracy when the model is provided image C and the exact transformation program T. Δ_{inf} measures the gain from removing relation inference, and Δ_{app} measures the residual gap to oracle grounding. Results are reported separately for each step count.

the remaining gap is specifically about analogical transfer of visual relations. Moreover, the separation between oracle grounding and end-to-end analogy grows as the transformation chain becomes longer, indicating that errors accumulate when models must preserve and manipulate transformed visual concepts across multiple steps. The central empirical finding of our benchmark is a growing composition gap: current VLMs remain much stronger at answering from direct visual evidence than at inferring, composing, and applying multi-step transformations.

4.2. Diagnosing inference and application errors

Motivation. Figure 2 measures end-to-end performance on the analogy reasoning task, requiring the model to both infer the composition of visual concepts and relations from A to B and apply it to C. To better understand this capability, we isolate two failure modes, separating errors in inferring the analogy from errors in applying it. We add a program-conditioned setting in which the model receives C together with the exact ground-truth transformation program T, and must answer the same question about hidden D. This removes relation inference while preserving the need to apply a multi-step transformation and reason over the resulting target.

Metrics. We report program-conditioned accuracy $\mathrm{Acc}(C,T)$ and two signed gaps:

$$\Delta_{\mathrm{inf}} = \mathrm{Acc}(C,T) - \mathrm{Acc}(A,B,C),$$
$$\Delta_{\mathrm{app}} = \mathrm{Acc}(D) - \mathrm{Acc}(C,T).$$

where $\mathrm{Acc}(A,B,C)$ is end-to-end analogy accuracy and $\mathrm{Acc}(D)$ is oracle VQA accuracy. A large Δ_{inf} indicates that inferring the relation from $A \to B$ is a major bottleneck; a positive Δ_{app} indicates residual difficulty in applying the known transformation and using the transformed visual concepts for question answering.

Findings. Across both models, the inference gap is large at every step and is generally larger beyond Step 1, indicating that relation inference is a major source of error in VISANALOG. This suggests that a substantial portion of the end-to-end failure arises before the model ever reasons about the hidden target: it often fails to recover how visual properties and relations change from A to B. At the same time, the application gap is also consistently positive and becomes more pronounced as the composition grows longer. For InternVL3.5-38B-Instruct, Δ_{app} rises from 9.7 points at Step 1 to 26.7 points at Step 4, indicating substantial residual difficulty even when the correct transformation is provided. Qwen3-VL-8B-Instruct shows a smaller residual gap, but it too increases with depth, from near zero on Steps 1–2 to 17.3 points on Step 4. Taken together, these trends show that both error sources are meaningful, and that both become more visible as the sequence of visual concept changes grows. More broadly, the negative correlation with step number suggests that current models do not yet exhibit robust, compositional mastery of visual concept recognition and transfer.

5. Conclusion

We introduced VISANALOG, a diagnostic benchmark for evaluating whether VLMs can recover a visual relation from one pair of natural images and use it to reason about an unseen counterpart. Each example instantiates an analogy $A : B :: C :?$, where the model must infer the edit sequence underlying $A \to B$, compose the corresponding operations on C, and answer a question about the latent target. Controlled evaluations confirm that the benchmark measures visually grounded relation identification and execution rather than single-image recognition or shortcut reasoning from the source image. Our main finding is a widening composition gap: as transformation depth increases, end-to-end accuracy falls substantially, revealing a weakness in current VLMs' ability to maintain and manipulate visual concepts across multi-step edits. Program-conditioned experiments show that this gap reflects both relation-inference failures and residual errors in applying known transformations, both worsening with composition length.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[2] Giacomo Camposampiero, Michael Hersche, Roger Wattenhofer, Abu Sebastian, and Abbas Rahimi. Can large reasoning models do analogical reasoning under perceptual uncertainty? *arXiv preprint arXiv:2503.11207*, 2025. 2

[3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2

[4] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 2

[5] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2

[6] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 2

[7] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. 2

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[9] Xiaoyang Hu, Shane Storks, Richard L. Lewis, and Joyce Chai. In-context analogical reasoning with pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. 2

[10] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[12] Daniel C. Krawczyk. The cognition and neuroscience of relational reasoning. *Frontiers in Human Neuroscience*, 6:64, 2012. 2

[13] Yongyuan Liang, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiuhai Chen, Jiatao Gu, Yue Wang, and Furong Huang. Rover: Benchmarking reciprocal cross-modal reasoning for omnimodal generation. *arXiv preprint arXiv:2511.01163*, 2025. 2

[14] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[15] Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*, 2023. 2

[16] Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*, 2023. 2

[17] Lindsey Engle Richland and Robert G Morrison. Is analogical reasoning just another measure of executive functioning? *Frontiers in Human Neuroscience*, 4:180, 2010. 2

[18] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 2

[19] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8876–8884, 2019. 2

[20] Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 2

[21] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025. 2

[22] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 2

[23] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Proceedings of the National Academy of Sciences*, 120(33):e2300487120, 2023. 2

[24] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[25] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[26] Junyan Ye, Dongzhi Jiang, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, Hongsheng Li, Conghui He, and Weijia Li. Blink-twice: You see, but do you observe? a reasoning benchmark on visual perception. *arXiv preprint arXiv:2510.09361*, 2025. 2

[27] Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. AnaloBench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13060–13082,

Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[28] Nilay Yilmaz, Maitreya Patel, Yiran Lawrence Luo, Tejas Gokhale, Chitta Baral, Suren Jayasuriya, and Yezhou Yang. Voila: Evaluation of MLLMs for perceptual understanding and analogical reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[29] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. 2

[30] Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning? *arXiv preprint arXiv:2403.04732*, 2024. 2

A. Model Prompts

Question Generation

```
========================
1) Context and Inputs
========================
You are a question generator that writes exactly one
     diagnostic multiple-choice question (MCQ) about
     the target image D.

You are given:
- C: the source image.
- D: the ground-truth target image produced by
     applying a sequence of transformations to C.
- Sigma = [tau_1, tau_2, ..., tau_k]: an ordered list
      of transformations mapping C -> D.

Transformation types are for your internal reasoning
     only; do NOT mention them in the question or
     options:
- Operational: pixel-space edits, e.g., rotation,
     center crop, flip, hue shift, quadrant swaps.

Evaluation usage: Later, a solver model will not see
     D. It will receive A, B, and C, infer A -> B,
     apply the inferred transformation(s) to C to
     imagine D, and then answer your question about D
     . Your question must probe visual details of D
     that are consequences of the transformation(s),
     thereby testing whether the solver can simulate
     transformations mentally.

========================
2) Objective
========================
1. Understand the visual consequences that
     distinguish D from C via Sigma.
2. Write one self-contained MCQ about D that:
   - Makes sense on its own.
   - Never mentions C, D, transformation, analogy, or
      step names.
   - Requires correct simulation of the full sequence
      of transformations to answer.
   - Is not answerable from C alone, generic priors,
     or an incorrect visualization of the target
     image.
3. Provide four options, labeled A, B, C, and D, with
      exactly one correct answer.
4. Make each distractor a plausible outcome of a
     specific mis-simulation: omitted step, wrong
     order, or wrong interpretation.
5. Provide an explanation proving why the correct
     option is uniquely true in D and diagnosing each
      distractor.

================================
3) Hard Leak-Prevention Rules
================================
Never reveal, hint at, or imply any of the following
     in the question or options:
- The existence of transformations, Sigma, step types
     , or operation names.
- Any verbs or phrases that imply change or causality
     , e.g., becomes, turned, after, before, now,
     transformed, once rotated, when aged, if winter
     arrives, gets flipped, or shifted.
- Any meta-language about the protocol, e.g., A/B/C/D
     , analogy, simulation, or apply the
     transformation.
- Direct references to D, such as in the final image.

Focus on neutral, stative facts about what is true in
      D: objects, attributes, spatial relations, and
     states. Do not ask what changed; only ask what
     is.

================================
4) Robustness and Diagnostic Power
================================
- Not answerable from C alone or generic priors. The
     correct answer must hinge on the effects of the
     transformation sequence that produces D.
- Plausible failure modes. Write distractors that
     reflect realistic mis-visualizations, e.g.,
     skipped steps, wrong order, or wrong magnitude,
     so the item challenges an unfaithful visualizer.
- Salient, stable consequences. Target robust,
     clearly visible outcomes that persist in a
     correct rendering of D; avoid tiny details or
     subtle, hairline differences.
- Parallel and balanced options. Keep the answer
     choices similar in length and style.

============================
5) Output Format
============================
Print exactly this JSON object, with no extra text
     and no code fences:

{
  rationale: <brief reasoning: which consequences of
     C -> D are targeted; why solving requires the
     entire sequence; how each distractor challenges
     an incorrect or inaccurate visualizer>,
  question: <one single-sentence MCQ about D; no
     mention of C, D, transformations, or analogy>,
  options: [<A>, <B>, <C>, <D>],
  explanation: <why the chosen option is uniquely
     correct for D only when all steps are applied>,
  answer: <A|B|C|D>
}
```

Analogy Reasoning with Search Space Information

```
You are a Visual Analogical Reasoner.

Task
- Solve the analogy A : B :: C : D.
- You are given three images: A, B, and C.
- Infer the minimal transformation T that maps A -> B
     , then mentally apply T to C to imagine D.
- Then answer a fine-grained multiple-choice question
      about the imagined D.

Input
- image_A, image_B, image_C
- question: a fine-grained query about the imagined D
- choices: four options labeled A, B, C, and D
```

```
Output
- Let's think step by step, and then output only one
     capital letter in a LaTeX box, e.g., \boxed{A}.

Each transformed image is produced by up to 4 non-
     repeating operations applied in a fixed order: a
      centered zoom, a swap of two tiles in a 2x2
     quadrant grid, a counter-clockwise rotation of
     90, 180, or 270 degrees, a horizontal or
     vertical flip, and a hue rotation of 90, 180, or
      270 degrees. Any subset of these operations may
      be used, but whenever an operation is included,
      it appears in this order.

Question:
[Question]

Options:
[Options]

Image A:
[Image A]

Image B:
[Image B]

Image C:
[Image C]
```

Standalone VQA Evaluation

```
You are a Visual Question Answering model.

Task
- Answer a multiple-choice question about the image.

Input
- image
- question: a query about the image
- choices: four options labeled A, B, C, and D

Output
- Let's think step by step, and then output only one
     capital letter in a LaTeX box, e.g., \boxed{A}.

Question:
[Question]

Options:
[Options]

[Input Image]
```