RESYNTH: VERIFIED SYNTHETIC DATA FOR MULTI-IMAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

People often understand the world by comparing different views of it. Vision-language models (VLMs), however, receive these views as separate images and still struggle to relate them. One problem is data. Training on these tasks requires groups of images with a known relationship: the same object from different viewpoints, or the same scene before and after a change. Such examples are hard to collect at scale from existing photo collections. We introduce ReSynth, which creates them entirely from generated images. Localized editing produces before-and-after pairs, viewpoint generation produces multiple views of the same instance, and verified scene contents support comparisons of counts and attributes. Generated images do not always match their prompts, so ReSynth does not assume that a requested relation was successfully rendered. Instead, it verifies facts such as object counts and whether an edit really happened, then builds questions only when the answer is supported by those verified facts. Using GRPO on 13K ReSynth questions, with no real images during training, Qwen2.5-VL-7B improves by **3.8** to **5.1** points on BLINK, MUIRBench, and VLM2-Bench, while single-image performance remains essentially unchanged. In a controlled comparison, using the object counts actually visible in generated images rather than the counts requested in their prompts improves counting accuracy by **9.75** points. ReSynth suggests that generated images can provide useful multi-image supervision. The key is to verify what the generated images actually contain before using them for training.

1 INTRODUCTION

Vision-language models (VLMs) have become increasingly capable at understanding individual images. In the real world, however, visual understanding often requires comparing observations across viewpoints and time. We may recognize the same suitcase from a new angle, spot what changed between two photos of a desk, or match a detail in one image to its counterpart in another. Such judgments require linking evidence across images and remain challenging for VLMs on BLINK (Fu et al., 2024), MUIRBench (Wang et al., 2024), and VLM2-Bench (Zhang et al., 2025a).

Training models for these tasks requires image groups with known relationships, such as multiple views of one instance or a scene before and after a local change. Such groups are difficult to collect at scale from existing photographs. Multi-image instruction datasets commonly assemble or retrieve available images before constructing QA (Jiang et al., 2024; Li et al., 2025; 2024). Other methods organize those images into permutations, contrastive groups, or grounded compositions (Zhang et al., 2025b; Chen et al., 2025; Zheng et al., 2026). These approaches remain limited by the relationships represented in their source images: rearranging images cannot recover a viewpoint or change that was never captured. Figure 1 contrasts this dependence on existing images with ReSynth's generation-based approach.

Existing photographs provide only the relationships that happened to be captured, whereas generative models allow us to construct candidate cross-image comparisons deliberately. We can plan source images with different objects, settings, and physical relations, then use local editing to create before-and-after states or viewpoint generation to show the same subject from another angle. Yet specifying a relation does not guarantee that it appears in the rendered images, and related images do not by themselves provide questions with well-supported answers. This raises two linked

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.

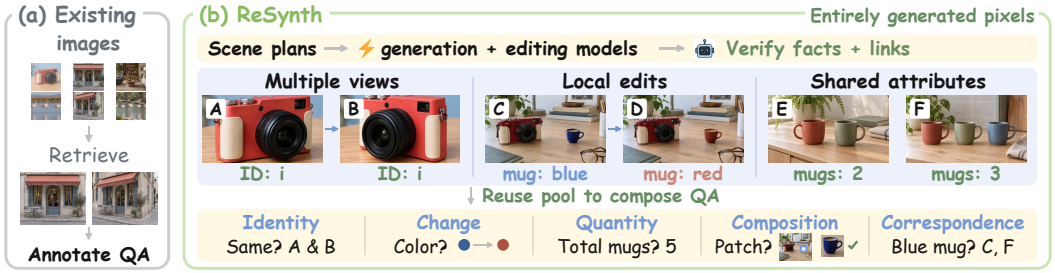

Figure 1: Existing-image curation retrieves and annotates available images. ReSynth plans and generates candidate relations, verifies the resulting visual evidence, and reuses it to construct questions across five multi-image abilities.

questions: **can entirely generated images improve multi-image understanding, and how can relations visible in those images be verified and reused to construct reliable QA supervision?**

Turning generated images into supervision requires broad content coverage and consistency between planned relations, rendered images, and QA answers. This raises three challenges. First, unconstrained caption generation with a large language model (LLM) can concentrate on a small set of recurring objects and scenes, limiting the content coverage of source images. Second, generation intent does not always match the resulting images: source images may fail to exhibit the specified attributes or counts, local edits may alter unrelated content, and new viewpoints may fail to preserve subject identity. Visible image attributes and cross-image associations therefore require verification to provide a reliable basis for subsequent QA construction. Third, reusable evidence does not by itself specify a training task; suitable images must be selected and their facts and associations combined into questions that require cross-image comparison.

To address these challenges, we introduce ReSynth (Relation-based Synthesis), which constructs multi-image supervision from verified visual relations in generated images. Its three-stage framework builds a reusable image pool and composes QA from it (Figure 2). First, ReSynth samples compatible combinations of object categories, scenes, activities, and physical relations before an LLM expands them into captions for source-image generation. This controls source-image content while allowing the same category to appear in different settings for later comparison. Second, ReSynth records what the images actually show instead of treating the plans as labels: it annotates and verifies visible objects, counts, attributes, and relations. Localized editing and viewpoint generation then add candidate change and same-instance associations, which are retained only after visual checks. Finally, ReSynth selects related images or combines independent sources according to these verified facts and associations, then constructs QA from their cross-image relations. Templates and computations determine answers when the records suffice; otherwise, a multimodal large language model (MLLM) proposes QA from the selected images and observations. For model-proposed QA, ReSynth checks the visual evidence in each image and the verified cross-image relations to confirm that the answer is supported and unique.

Using this pipeline, we construct 13K multi-image examples across five core abilities entirely from generated images. Although the ReSynth training set contains no real images, training Qwen2.5-VL-7B (Bai et al., 2025b) on it with GRPO (Shao et al., 2024; Meng et al., 2025) improves the model's average score across BLINK, MUIRBench, and VLM2-Bench by 4.25 points.

Our contributions are:

1. We introduce ReSynth, a relation-based synthesis framework that generates source images, verifies visible facts and cross-image associations, and constructs multi-image QA from supported relations. Its reusable image pool allows the same visual evidence to support different questions.
2. We construct a **13K**-example training corpus entirely from generated images, covering identity, change, quantity, composition, and correspondence. Each example contains two to eight images.
3. We show that reinforcement learning on generated images alone can improve multi-image understanding. On Qwen2.5-VL-7B, ReSynth raises the average score across BLINK, MUIRBench, and VLM2-Bench by **4.25** points. The resulting average even exceeds the reported score obtained with real-world image data.

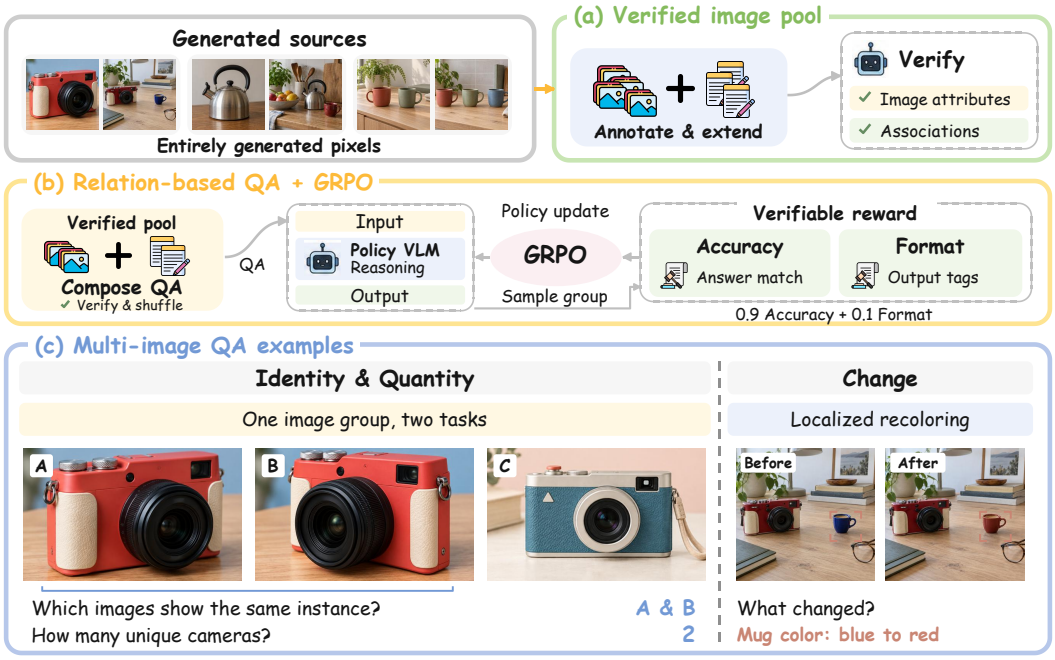

Figure 2: Overview of ReSynth. (a) Verified image facts and associations form a reusable pool. (b) Selected images yield checked QA for GRPO training. (c) Example identity, counting, and change tasks.

2 Related Work

Multi-image evaluation. Multi-image benchmarks test whether models can connect evidence across images. BLINK (Fu et al., 2024) casts 14 perception tasks, including relative depth and visual correspondence, as multiple-choice questions. MUIRBench (Wang et al., 2024) covers ten multi-image relation categories with paired answerable and unanswerable examples. VLM2-Bench (Zhang et al., 2025a) focuses on linking people and objects across views, while MMIU (Meng et al., 2024) spans 52 tasks over seven relation types.

Existing multi-image supervision. Existing methods primarily construct supervision from available images. Mantis (Jiang et al., 2024) aggregates interleaved instructions, SMiR (Li et al., 2025) retrieves correlated images, and LLaVA-NeXT-Interleave (Li et al., 2024) combines image, video, and 3D inputs. PeRL (Zhang et al., 2025b) trains with image permutations, MiCo (Chen et al., 2025) contrasts image groups, and CGC (Zheng et al., 2026) generates grounded examples from detections. Multi-SpatialMLLM (Xu et al., 2025a) uses annotated 3D and 4D scenes, while ReMoT (Wan et al., 2026) derives motion triplets from video. Their available images and annotations constrain which identities, states, and viewpoints can be related.

Synthetic data for LLMs and VLMs. Self-Instruct (Wang et al., 2023), STaR (Zelikman et al., 2022), and Magpie (Xu et al., 2025b) generate training data for language models. Visual generation supports synthetic scenes (Z-Image Team, 2025), local edits (Tencent Hunyuan Foundation Model Team, 2026), and new viewpoints (Wu et al., 2025; Odin & fal, 2026). R1-SyntheticVL (Zhang et al., 2026) studies generated images for multimodal learning, while VisionFoundry (Zhou et al., 2026) verifies task-directed visual examples. ReSynth verifies and reuses cross-image associations to construct multi-image QA.

3 Method

ReSynth builds a reusable image pool with verified facts and checked edit and viewpoint associations, then composes multi-image QA from it (Figure 2).

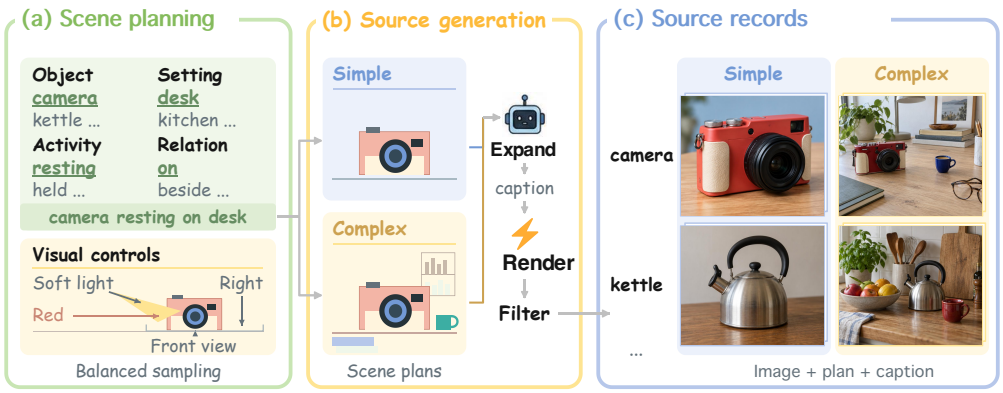

Figure 3: Structured source generation. (a) Compatible elements form scene plans. (b) Simple and complex plans become captions and images. (c) Accepted sources retain their plans and captions.

1. **Structured Source Generation.** Compatible scene elements are combined into plans that specify the intended content and candidate contrasts between sources. A language model expands each plan into a caption, which an image generation model renders as a source image.
2. **Verified Image Facts and Cross-Image Relations.** The content actually visible in each source image is checked and recorded as image facts, rather than inferred from its scene plan. Localized editing and viewpoint generation then establish checked change and same-instance relations between images.
3. **Composing Multi-Image Supervision.** Each task type selects and combines images using their verified facts and checked associations to construct questions and answers.

3.1 Structured Source Generation

Scene planning. To produce sources that can be reused across questions, ReSynth plans their content before generating images. It samples a main object category, a compatible setting and activity, and a visible physical relation from predefined pools; appearance, lighting, viewpoint, and placement provide further variation. Each plan records intended attributes that can be shared or contrasted across source images. For example, two plans can share the bicycle category while placing one bicycle on a street and another beside a workbench in a repair shop. Balanced sampling covers different compatible settings and relations for each category. Figure 3 shows how these plans are expanded into captions, rendered as images, and filtered for the source pool.

Scene richness control. Simple plans keep the main subject, activity, and essential context; complex plans add compatible objects across the scene. A bicycle repair plan, for example, may contain only a bicycle, mechanic, and tools, or also tire racks, other bicycles, and participants. Simple plans provide clear subjects for viewpoint generation, while complex plans provide more objects for annotation and comparison.

Source generation and records. A caption generator expands each scene plan into a caption, which an image generator renders into a source image. Images with missing subjects, severe artifacts, poor framing, or near-duplicate content are filtered out. Each accepted image is retained with its plan and caption, which describe the intended objects, setting, and relations. The plans help retrieve candidate sources with shared or contrasting content; the next stage checks what the images actually show before these differences are used for QA.

3.2 Verified Image Facts and Cross-Image Relations

Scene plans specify intended content, not image labels. ReSynth first checks each source image for visible objects, locations, counts, and attributes, recording them as image-level facts. It then applies localized editing or viewpoint generation to selected sources and checks whether the edits produce the intended changes or the new views preserve instance identity. These checks establish cross-

image relations between each source and its derived images. The pool thus supports two routes to QA: combining independent sources using their verified facts, or selecting images already linked by a checked relation.

Source-image fact verification. A multimodal model annotates object categories, boxes, visible counts, and attributes, with repeated annotations checked by voting. When a source differs from its plan, the record follows the image: a planned count of three becomes two if only two objects are verified. These records guide later image selection and answer construction, while object boxes also locate edit targets and regions referenced by questions.

Locally edited images. Object boxes guide the selection of relatively small secondary targets for local editing. An image editing model applies operations such as `add`, `remove`, `replace`, or `recolor` within the target region. When appropriate, the region is cropped for editing and composited back into the source image to limit the affected area. The source and edited images are compared to confirm the target change and surrounding consistency; accepted pairs retain the target location and observed before-and-after states.

Multiview images. Sources generated in the simple mode with a large, complete, and clearly visible main subject are selected for viewpoint generation. An image editing model with viewpoint control generates candidate left-side, right-side, and back views. Visual checks assess subject identity and structural quality; accepted views are linked to the same instance for later QA.

3.3 Composing Multi-Image Supervision

The image pool does not bind each source or derived image to one question. For each task type, ReSynth selects images from the pool, combining independent sources using their verified facts or using images linked by a checked edit or instance association. Scene plans help find candidate independent sources, but their observed content determines which comparisons are valid. The resulting QA covers identity, change, quantity, composition, and correspondence, with the dataset distribution summarized in Table 1. Questions use multiple-choice, true/false, or integer-answer formats.

Image selection by task type. Identity and change tasks use existing associations: views of the same instance support matching and unique-instance counting, while before-and-after edit pairs support change questions. Quantity and correspondence tasks can also combine independent sources according to their verified objects, counts, attributes, and relations. For example, the bicycle plans in Section 3.1 suggest a scene-based comparison. If the images confirm a bicycle beside a workbench in one source and a bicycle on a street in the other, the pair can support a question asking which image shows the bicycle beside the workbench. Likewise, the same sources containing two and three cups can support either a total-count question with answer five or a comparison that selects the second image, without generating new images. For composition tasks, ReSynth masks a source-image region and combines the original patch with distractor patches as answer candidates.

QA construction. When answers follow from verified records, as in count aggregation, templates and computations produce the QA. For visual descriptions or comparisons, a multimodal model uses the selected images and observation records to generate questions, options, and candidate answers. Some tasks combine both routes: the model checks visible states, while templates arrange the options and determine the answer label. The same images can support different questions: an edited pair can be used to select the correct change description or judge whether a specified change occurred. Distractors preserve some similarities but fail to satisfy the target relation, such as a different instance of the same category or a change description with the before and after states reversed.

Relation and answer checks. For model-generated candidate QA, ReSynth checks the question, options, and proposed answer against each image in turn, recording visible evidence that supports or contradicts the answer. It then combines these per-image findings with verified edit or instance associations to check whether the proposed answer holds and whether another option could also be correct. Candidates that fail these checks are discarded. Image and option order are varied, with the corresponding answer mappings updated and checked to reduce positional shortcuts.

Table 1: Distribution of approximately 13K ReSynth examples across five abilities.

Ability	# Ops.	QA operations	Share
Identity	3	Instance grouping; evidence selection; unique counting	19%
Change	3	Change comparison; state binding; verification	32%
Quantity	2	Count aggregation; cross-view instance statistics	22%
Composition	1	Missing-patch recovery	12%
Correspondence	1	Attribute and relation matching	15%
Total	**10**		**100%**

3.4 TRAINING

We train the model with GRPO (Shao et al., 2024), using rule-based correctness and format rewards as in DeepSeek-R1 (DeepSeek-AI et al., 2025). For each question, the model samples a group of responses. Each response places visual reasoning inside `<think>...</think>`, followed by the final answer inside `<answer>...</answer>`. Given a response o and reference answer a, the reward is

$$r(o, a) = 0.9\, r_{\mathrm{acc}}(o, a) + 0.1\, r_{\mathrm{fmt}}(o), \tag{1}$$

where r_{acc} scores the correctness of an option letter, T/F label, or integer, and r_{fmt} checks the required output format. The reasoning content is not scored separately. We use the same objective across all task types.

4 EXPERIMENTS

We test whether entirely generated images and their verified cross-image relations can provide effective training supervision for multi-image understanding. We first evaluate ReSynth across models and tasks, then analyze how data scale, relation verification, and image construction affect training gains. Our experiments address three questions:

1. **Effectiveness and transfer:** Can supervision constructed entirely from generated images improve multi-image understanding across VLMs and benefit different tasks?
2. **Data scale and quality:** How does performance change as the model sees more training examples? How much does checking visible facts before QA construction help at a fixed data size?
3. **Discussion:** Which data-construction choices broaden the visual evidence available for supervision, and what limitations remain in the resulting data?

4.1 EXPERIMENTAL SETUP

Training Data. ReSynth comprises approximately 13K multi-image examples covering five abilities and ten QA operations, as shown in Table 1. Each example contains two to eight generated images, and the corpus includes 30,053 unique images. We construct these data using existing open models to examine whether generated images, organized through relation construction and verification, can provide effective supervision for multi-image understanding. Qwen3.6-27B (Qwen Team, 2026) performs caption expansion and visual annotation, while Z-Image (Z-Image Team, 2025) provides efficient source-image generation. We then use the instruction-guided editing capabilities of HunyuanImage 3.0 (Tencent Hunyuan Foundation Model Team, 2026) to modify selected objects locally, and Qwen-Image-Edit-2511 with the Multiple-Angles LoRA (Wu et al., 2025; Odin & fal, 2026) to generate different views of the same instance. These operations establish change and instance associations for subsequent QA construction.

Models and Training. We train Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) and Qwen3-VL-8B-Instruct (Bai et al., 2025a) to evaluate ReSynth on two generations of Qwen-VL models. Both models undergo full-parameter fine-tuning with the same GRPO training settings and a learning rate of 10^{-6}. Both the global rollout and update batch sizes are 256, and each prompt receives eight

Table 2: Results on BLINK (Fu et al., 2024), MUIRBench (Wang et al., 2024), and VLM2-Bench (Zhang et al., 2025a). Avg. (Average) is the mean of BLINK Overall, MUIRBench, and VLM2. Scores are accuracies (%).

Model / Method	#Train	BLINK			MUIR	VLM2	Avg.
		Single	Multi	Overall			
Qwen2.5-VL-7B and reference models							
GPT-4o† (OpenAI, 2024)	-	-	-	68.0	68.0	59.6	65.2
Qwen2.5-VL-7B (Bai et al., 2025b)	-	55.2	54.0	54.5	58.4	60.9	57.9
MM-Eureka† (Meng et al., 2025)	15.6K	-	-	54.4	60.6	59.0	58.0
PeRL† (Zhang et al., 2025b)	58K	-	-	58.5	-	-	-
MiCo† (Chen et al., 2025)	87K	-	-	57.2	60.5	63.2	60.3
CGC-7B† (Zheng et al., 2026)	24K	-	-	56.2	60.9	63.7	60.3
ReSynth	13K	56.9	59.5	58.4	63.5	64.6	62.2
Δ over Base	-	+1.6	+5.5	+3.9	+5.1	+3.8	+4.3

Table 3: Ability-level and single-image benchmark results for Qwen2.5-VL-7B. Ability scores are unweighted subtask averages; Δ denotes ReSynth minus Base. Scores are percentages.

	Multi-image abilities					Single-image benchmarks			
Model	**Identity**	**Change**	**Quantity**	**Compos.**	**Corresp.**	**MMStar**	**MMMU**	**MathVista**	**Hallusion**
Base	57.2	59.1	59.0	55.9	40.6	65.7	55.0	67.0	68.7
ReSynth	60.6	64.1	63.5	56.2	50.3	64.9	54.6	67.9	69.9
Δ	+3.4	+4.9	+4.4	+0.3	+9.7	−0.8	−0.4	+0.9	+1.2

sampled responses at temperature 1.0 and top-p 1.0, without KL regularization. The Qwen2.5-VL-7B training configuration and response format are provided in Appendix C.

Evaluation Protocol. We use VLMEvalKit (Duan et al., 2024) to evaluate on BLINK (Fu et al., 2024), MUIRBench (Wang et al., 2024), and the eight image-based tasks of VLM2-Bench (Zhang et al., 2025a). For BLINK, we report overall, single-image, and multi-image results on 1,901, 793, and 1,108 examples, respectively. We report overall accuracy for MUIRBench and the macro-average over the eight VLM2-Bench tasks using their official scoring rules. Results for prior methods are taken from their papers and marked accordingly in the table.

4.2 Main Results

Overall Performance. As shown in Table 2, training Qwen2.5-VL-7B with ReSynth improves performance on BLINK, MUIRBench, and VLM2-Bench. Its average across the three benchmarks also exceeds the reported scores of MiCo and CGC, which construct supervision from existing visual data, showing that generated images can provide effective supervision for multi-image understanding. On BLINK, gains appear in both the multi-image and single-image portions, indicating that the benefit extends to single-image understanding.

Ability and Task Analysis. We group evaluation subtasks into identity, change, quantity, composition, and correspondence to examine the 13K corpus by ability. Table 3 shows gains across all five abilities, with the largest gain on correspondence. The same table reports MMStar (Chen et al., 2024), MMMU (Yue et al., 2023), MathVista (Lu et al., 2024), and HallusionBench (Guan et al., 2024) to assess single-image performance. ReSynth training has little effect on single-image performance across these four benchmarks.

Qualitative Comparison. Figure 4 compares responses from the base and ReSynth-trained models under the same inference settings. On two MUIRBench examples, the trained model correctly counts sculpture pairs and matches images to the specified sky colors, correcting errors made by the base model.

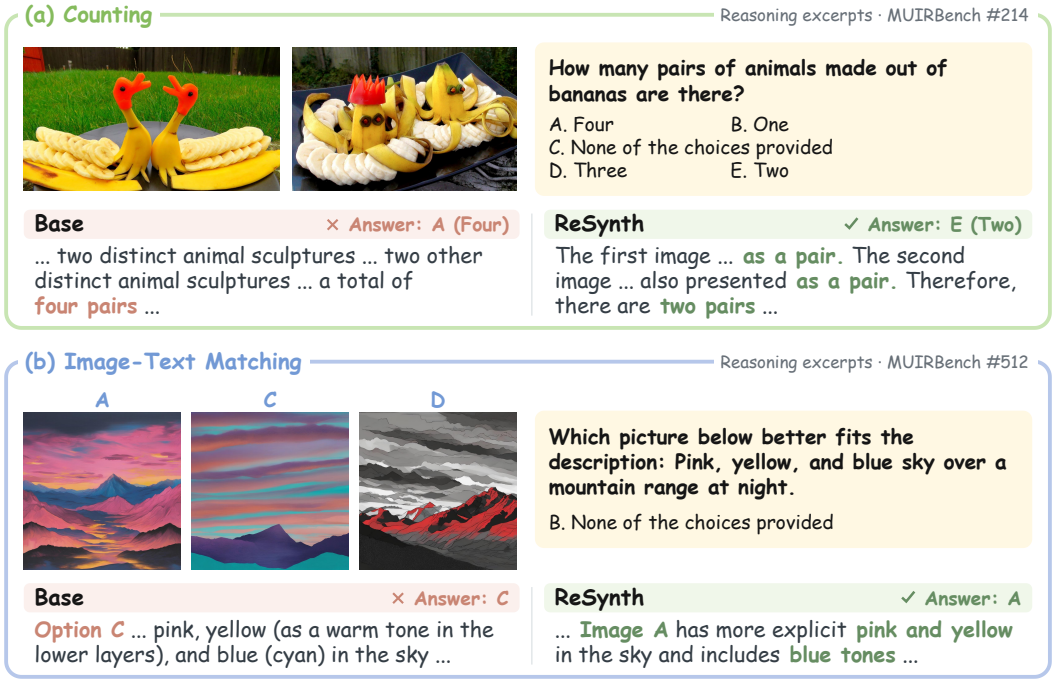

Figure 4: Qualitative comparisons on MUIRBench with matched prompts and inference settings for base and ReSynth-trained Qwen2.5-VL-7B. ReSynth (a) counts pairs across images and (b) matches the specified sky colors. Images and questions are unmodified; both panels show reasoning excerpts, with omissions marked by ellipses.

4.3 Data Scale and Visual Verification

Effect of Data Scale. Figure 5(a) tracks performance as the number of ReSynth training examples increases while retaining the same task types. Performance improves at smaller scales but levels off and fluctuates as the corpus grows further. This motivates adding harder examples and more diverse visual relations, rather than only increasing the number of examples in the existing task types.

Effect of Visual Verification and Filtering. Before constructing QA, ReSynth corrects annotations against the rendered images and filters examples whose visual facts cannot be verified. To test this step, we construct two sets of 1,000 questions about object counts across two or three images. The unverified set uses counts specified in the generation plans, while the verified set corrects answers using visible counts and replaces images whose counts cannot be established reliably. Both sets use the same question forms, option layouts, and training settings. Training on the verified set raises object-counting accuracy from 55.20 to 64.95 (+9.75) compared with the unverified set. This comparison shows that correcting and filtering examples according to the rendered images improves cross-image counting supervision.

4.4 Discussion

Subject Diversity in Captions. To test whether prompting alone broadens the range of visual subjects, we use Qwen3.6-27B to generate 500 image descriptions for each of three prompts under identical sampling settings:

(1) Free caption: choose any subject and scene for a realistic photograph.
(2) Distinctive subject: choose a visually distinctive subject from any domain and describe its appearance and setting.
(3) Plan before captioning: consider subjects from several unrelated domains, then choose a less common, visually concrete subject to describe.

For each prompt, we measure the share of descriptions belonging to its five most frequent subject categories; a higher share means greater category concentration. The free-caption prompt yields

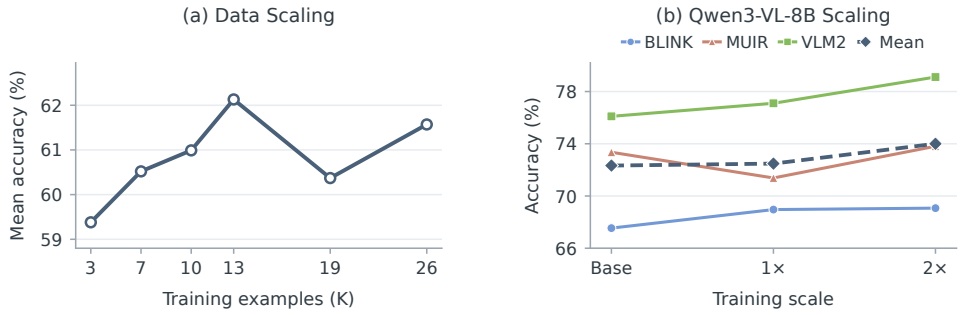

Figure 5: Effect of training data scale on two model backbones. (a) Mean accuracy of Qwen2.5-VL-7B over BLINK, MUIRBench, and VLM2-Bench as training examples increase. (b) Scores of Qwen3-VL-8B on the same three benchmarks and their mean at Base, 1x, and 2x training scales.

29.4%, while asking for a distinctive subject raises this share to 37.8%, with "tree" alone accounting for 16.4%. Planning before captioning lowers the top-five share to 11.2%, but the descriptions still converge in appearance: 48% depict rusted or corroded objects, and 32% use close-up framing. Thus, prompting can reduce repetition of subject categories while leaving a recurring visual style. ReSynth instead samples object categories and compatible scenes before caption expansion to guide coverage of both subjects and settings (Section 3.1).

Matching Data Difficulty to Model Capability. The current ReSynth corpus improves Qwen2.5-VL-7B, but the same data yield a much smaller gain for Qwen3-VL-8B, which already has stronger cross-image reasoning ability. When the number of examples is doubled, Qwen3-VL-8B achieves a modest gain (Figure 5(b)). This pattern indicates that many examples in the current corpus are relatively easy for the stronger model. Future data scaling should cover a wider range of visual relations and difficulty levels to support more consistent gains across backbones.

Limits of Viewpoint Generation. Multiview images are an important source of identity supervision in ReSynth: views of the same instance support cross-image matching and distinct-instance counting. Current viewpoint generation is less reliable in cluttered scenes, where it may change identifying details of the subject or rearrange surrounding objects, invalidating the intended same-instance association. We therefore select simple scenes with a large, complete, clearly visible subject and retain only generated views that pass identity and quality checks. This keeps the associations reliable but limits the complexity of scenes represented in the multiview data. Stronger image generation models or world models that preserve object identity and scene layout across viewpoints could extend this supervision to more complex scenes.

5 Conclusion

ReSynth demonstrates that entirely generated images can support multi-image understanding. Its relation-based pipeline generates source images, verifies their visible content, and checks the changes and same-instance links created through editing and viewpoint generation. It retains these facts and links in a reusable pool, allowing the same images to support different QA tasks across five abilities. GRPO training on the resulting corpus improves Qwen2.5-VL-7B on BLINK, MUIR-Bench, and VLM2-Bench, with an average above the results reported by MiCo and CGC. This shows that generated images can provide useful training supervision without real images in the ReSynth training set.

The analyses also clarify where data construction matters. The counting comparison illustrates the value of labeling what the image shows rather than what its generation prompt requested. For stronger models, simply increasing the number of existing examples brings limited gains, while reliable viewpoint generation remains difficult in cluttered scenes. Future work will construct harder and more diverse cross-image relations and improve identity and scene consistency across generated viewpoints.

AI USE STATEMENT

Generative AI is part of our data-construction pipeline. We use language and multimodal models to expand scene plans into captions, annotate visible content, and help generate and check candidate QA. Image generation and editing models create source images, local changes, and alternative views. Generated outputs undergo the visual verification and QA checks described in this paper, and training answers are determined from verified image facts and retained construction records. We also use language models to improve manuscript clarity and assist with code development. The authors reviewed AI-assisted text and code and take responsibility for the reported results and final content of this work.

ETHICS STATEMENT

This work does not involve human-subject studies or the direct collection of personal photographs. The training corpus consists of generated images, while existing public benchmarks are used only for evaluation. We will document the generation and editing provenance of the released corpus and its intended research use.

REPRODUCIBILITY STATEMENT

The Method and Experimental Setup sections describe the data construction, training, and evaluation procedures. We plan to release the corresponding code to support reproducibility.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xi Chen, Mingkang Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. MiCo: Multi-image contrast for reinforcement visual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Zerun Ma, et al. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. *arXiv preprint arXiv:2407.11691*, 2024. URL https://arxiv.org/abs/2407.11691.

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 (ECCV)*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.

Andrew Li, Rahul Thapa, Rahul Chalamala, Qingyang Wu, Kezhen Chen, and James Zou. SMIR: Efficient synthetic data pipeline to improve multi-image reasoning. *arXiv preprint arXiv:2501.03675*, 2025.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, et al. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*, 2024. URL https://arxiv.org/abs/2407.07895.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, 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. *arXiv preprint arXiv:2408.02718*, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Lovis Odin and fal. Qwen-Image-Edit-2511-Multiple-Angles-LoRA. Hugging Face model repository, 2026. URL https://huggingface.co/fal/Qwen-Image-Edit-2511-Multiple-Angles-LoRA.

OpenAI. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Tencent Hunyuan Foundation Model Team. HunyuanImage 3.0 Technical Report, 2026. URL https://arxiv.org/abs/2509.23951.

Cong Wan, Zeyu Guo, Jiangyang Li, SongLin Dong, Yifan Bai, et al. ReMoT: Reinforcement Learning with Motion Contrast Triplets. *arXiv preprint arXiv:2603.00461*, 2026. URL https://arxiv.org/abs/2603.00461.

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. *arXiv preprint arXiv:2406.09411*, 2024.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13484–13508, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, et al. Multi-SpatialMLLM: Multi-Frame Spatial Understanding with Multi-Modal Large Language Models. *arXiv preprint arXiv:2505.17015*, 2025a. URL https://arxiv.org/abs/2505.17015.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *International Conference on Learning Representations (ICLR)*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:2311.16502*, 2023.

Z-Image Team. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15476–15488, 2022.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. VLM2-Bench: A closer look at how well VLMs implicitly link explicit matching visual cues. *arXiv preprint arXiv:2502.12084*, 2025a.

Jingyi Zhang, Tianyi Lin, Huanjin Yao, Xiang Lan, Shunyu Liu, and Jiaxing Huang. R1-SyntheticVL: Is synthetic data from generative models ready for multimodal large language model? *arXiv preprint arXiv:2602.03300*, 2026.

Yizhen Zhang, Yang Ding, Shuoshuo Zhang, Xinchen Zhang, Haoling Li, Zhong-zhi Li, Peijie Wang, Jie Wu, Lei Ji, Yelong Shen, Yujiu Yang, and Yeyun Gong. PeRL: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Lihao Zheng, Zhenwei Shao, Yu Zhou, Yan Yang, Xintian Shen, Jiawei Chen, Hao Ma, and Tao Wei. CGC: Compositional grounded contrast for fine-grained multi-image understanding. *arXiv preprint arXiv:2604.22498*, 2026.

Guanyu Zhou, Yida Yin, Wenhao Chai, Shengbang Tong, Xingyu Fu, and Zhuang Liu. VisionFoundry: Teaching VLMs visual perception with synthetic images. *arXiv preprint arXiv:2604.09531*, 2026.

A Data Construction Details

Source images are built in two steps: the system chooses what an image should contain, then expands that plan into a caption for image generation. The plan records intended content, while later visual checks determine what the generated image actually shows. This section gives the sampling rules and prompts for source-image generation and local editing.

A.1 Sampling Scene Plans

Compatible scene configurations. The planner samples object categories uniformly from a predefined COCO category pool. For each category, scene templates specify compatible locations, activities, and relations to nearby objects. The planner samples a valid combination of category, scene, location, activity, and relation while avoiding repeats within each sampling cycle.

Balancing scene coverage. To improve content coverage, the planner favors compatible combinations that have appeared less often in previous scene plans. Appearance, lighting, and placement are sampled separately for each image.

A.2 Image Generation and Local Editing

Input to the image generation model. The generated `caption` serves as the text input to the image generation model. It specifies the planned subject, scene, activity, relation, appearance, lighting, and placement; image dimensions and a random seed are taken from the corresponding source record. The following caption is an example for a bicycle at a town intersection:

Image caption: bicycle

A stationary bicycle rests slightly left of center at a town intersection, captured from an opposite three-quarter view at subject height in a medium-to-wide frame. The bike, featuring clean well-maintained surfaces and healthy tires, sits on the road surface just before a junction marking, waiting at the edge of the route. Cool diffuse illumination bathes the scene, highlighting the quiet surrounding context of the asphalt and pavement without harsh shadows. The composition keeps the bicycle prominent and clearly visible, with only necessary support elements and a few compatible props present to maintain a simple, readable silhouette against the ordinary street background.



(a) Generated source image



(b) Before editing



(c) After editing

Figure 6: Examples of source-image generation and local editing. Panel (a) is generated from the bicycle caption above. Panels (b) and (c) show a separate source image before and after replacing a baseball cap with a motorcycle helmet.

Local editing. We first locate a visible target using its bounding box and crop a region that includes the target and nearby context. The crop is provided to the edit model with an instruction to remove, replace, or recolor the target. The edited crop is blended into the original image to create a before-and-after pair. Candidates are skipped if the target cannot be located or the intended edit cannot be specified clearly. For the baseball-cap replacement in Figure 6, Qwen3.6-27B compares the full image pair and crops around the target. It checks whether the baseball cap was replaced by a motorcycle helmet, the surrounding scene was preserved, and the result has no obvious artifacts.

B QA CONSTRUCTION

How records determine answers. The following examples illustrate the construction rules rather than quote benchmark questions. For identity, a reference and several candidate views support retrieval when the verified instance association identifies a unique match. For change, a source and its edited output support a question about the target's state transition, with distractors reversing the transition or naming an unchanged object. For quantity, two images verified to contain two and three cups support a total-count answer of five; a distinct-instance question would additionally require evidence about whether any cups are shared. For composition, the original patch removed from a source image determines the correct candidate, and the other patch provides a distractor. For correspondence, visible attributes and relations identify which image satisfies the stated condition, such as a dog lying on an unpaved surface.

A concrete QA example. Using Figure 6(b,c) as Images 1 and 2, we ask: "What changed above the television between the two images?" The options are (A) a baseball cap was replaced by a motorcycle helmet, (B) a baseball cap was replaced by a toy car, (C) a beanie was replaced by a motorcycle helmet, and (D) a beanie was replaced by a toy car. The correct answer is (A), determined by the visible object in each image.

Text-only shortcut screening. For multiple-choice questions, we hide the images and ask a verifier to answer using only the question and options, repeating the check with a different option order. Questions that remain consistently answerable without visual evidence are excluded to reduce wording and option-pattern shortcuts.

C TRAINING DETAILS

We train Qwen2.5-VL-7B with GRPO using the settings in Table 4.

Table 4: GRPO training configuration for Qwen2.5-VL-7B.

Parameter	Value
Optimization	
Objective	GRPO
Trainable parameters	Full model
Learning rate	10^{-6}
Global update batch size	256
KL regularization	None
Rollout sampling	
Global rollout batch size	256
Responses per prompt	8
Temperature	1.0
Top-p	1.0
Reward	
Accuracy weight	0.9
Format weight	0.1

For multiple-choice training examples, we use the following response instruction:

Training response instruction

Explain the visual evidence inside `<think></think>`, then put exactly one option letter inside `<answer></answer>`.

The correctness reward compares the extracted final answer with the reference option, truth value, or integer, depending on the task. The format reward checks the response structure, while the reasoning text receives no separate reward.