M^{2}CHANGE: CAPTIONING AND PAIRED GROUNDING OF MULTIPLE MINIMAL CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounded image-difference understanding asks models to describe what changes between two images and locate the visual evidence for each change. Most existing benchmarks consider one change at a time, but a pair may contain several changes whose descriptions must be linked to the right objects and regions. We introduce M^{2}Change, a benchmark of natural-image pairs with 2–4 simultaneous changes drawn from object, attribute, counting, and relation categories. For each change, models produce a description, entity labels, and boxes in the original and changed images. The test set contains 1,063 pairs and 2,525 annotated changes for evaluating individual predictions and complete-set recovery. We also propose Paired-view Distillation and Set Optimization (PDSO), a three-stage training framework. Supervised fine-tuning first teaches the structured output. In paired-view on-policy distillation, a frozen teacher scores the same response with change-focused and unrelated control crops. This comparison gives greater distillation weight to tokens whose teacher predictions differ between the two views. Finally, set-level reward optimization encourages complete, correctly grounded predictions. Inference uses only the image pair. PDSO reaches 47.2 MCJS, 5.8 points above the strongest evaluated baseline.

1 INTRODUCTION

Grounded image-difference understanding asks a multimodal large language model (MLLM) to explain what changed between two images and locate the visual evidence for each change. This ability matters when verifying image changes, inspecting products, or monitoring scenes over time (Pan et al., 2026; Liberatori et al., 2026). A pair may contain an object replacement, a subtle attribute change, and a shift in spatial relations. Describing only the most obvious change leaves the comparison incomplete. Recovering several changes also requires linking each description to its supporting regions, so the model must identify the full set and ground each change.

Prior work addresses different parts of this problem. Image difference captioning (IDC) benchmarks emphasize image-pair descriptions, including subtle or context-dependent differences (Jhamtani & Berg-Kirkpatrick, 2018; Liu et al., 2025a; Kim et al., 2026b;a). DiffGround, DiG, and OddGrid-Bench instead focus on locating queried differences, discrepancies, or anomalous elements (Wang et al., 2025; Tao et al., 2026; Weng et al., 2026). Descriptions or locations alone, however, do not establish which regions support each detected change. The closest work, MINCU, links a description and entity labels to paired boxes for one atomic change per image pair (Mu et al., 2026). With several simultaneous changes, the model must recover these links separately for each event while identifying the full set. This creates a further challenge: discovering multiple changes and keeping their descriptions and visual evidence correctly associated.

Evaluating these associations requires more than scoring descriptions and regions separately. Because predicted changes may differ in number and order from the references, evaluation must align them one-to-one before assessing their quality. A caption score can overlook missing changes, while a localization score over pooled boxes can reward a correct region attached to the wrong description. A joint evaluation must therefore account for missing or redundant changes while checking whether each description and its regions refer to the same event.

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.

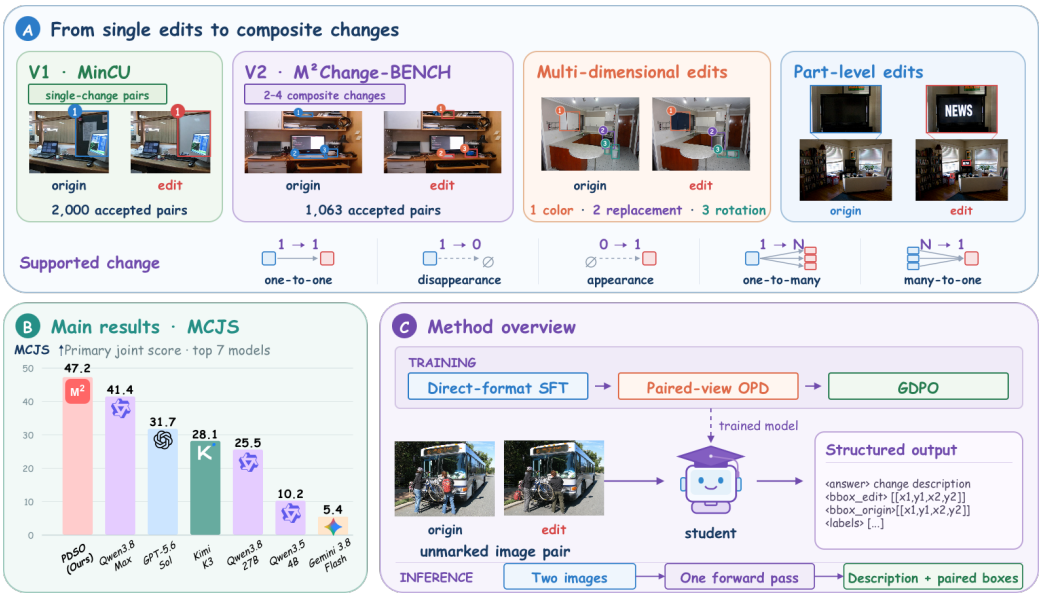


Figure 1: Overview of M^{2}Change. The benchmark evaluates complete understanding of 2–4 heterogeneous changes in natural image pairs. The model must describe every change and associate it with corresponding evidence in the original and changed images.

These associations can involve more than one region per change. As Figure 1 illustrates, the evidence may be confined to an object part, span several regions, or occur on only one image side. Removing an object, for example, leaves evidence in the original image but none in the changed image. Such cases require a change-level representation that links each description to its supporting regions.

To make these associations explicit, we formulate grounded multi-change understanding as variable-size change-set prediction. Given an original and a changed image, the MLLM predicts the changes without being told the number of changes. Each prediction contains a description, entity labels, and side-specific spatial evidence. We introduce M^{2}Change, built from natural photographs edited with 2–4 changes across object, attribute, counting, and relation categories. Its test set contains 1,063 image pairs and 2,525 annotated changes. Our primary metric, MCJS, matches predictions to references and assesses grounded descriptions together with change-set completeness.

Learning grounded change sets requires sensitivity to local evidence and supervision of the complete response. We develop Paired-view Distillation and Set Optimization (PDSO) to address these requirements in three stages. Supervised fine-tuning first establishes the structured output. To focus subsequent learning on evidence for individual changes, paired-view on-policy distillation (PV-OPD) compares a frozen teacher's predictions under change-focused and unrelated control crops (Agarwal et al., 2024). The teacher's response to these views determines the token-level distillation weights. Finally, GDPO combines rewards for coverage, descriptions, labels, and paired grounding under a shared assignment between predicted and reference changes (Liu et al., 2026). The privileged crops guide training, while inference uses only the image pair.

On M^{2}Change, PDSO reaches 47.2 MCJS, 10.8 points above direct-format SFT and 5.8 points above the strongest evaluated baseline.

Our contributions are threefold:

- We formulate grounded image-difference understanding as prediction of a complete set of changes and introduce MCJS to assess both change-level grounding and set coverage.
- We construct M^{2}Change from natural photographs with generated edits, providing 1,063 pairs and 2,525 annotated changes with 2–4 changes per pair.
- We develop PDSO, which uses paired teacher views to guide token-level distillation and a shared change assignment to align set-level rewards.

2 Related Work

Visual difference understanding. Early visual difference benchmarks focused on image-pair descriptions: Spot-the-Diff used real surveillance images, while CLEVR-Change supplied controlled synthetic edits (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019). More recent benchmarks cover diverse scenes and subtle or context-dependent comparisons (Liu et al., 2025a; Kim et al., 2026b;a). These tasks assess comparison and captioning, but do not require a separate description with paired spatial evidence for every change. DiffGround localizes a difference specified in language (Wang et al., 2025), while DiG discovers an unknown number of discrepancies in rendered 3D scenes and returns boxes without per-change descriptions or entity labels (Tao et al., 2026). Spot-the-Shift links descriptions to masks for long-term scene changes (Liberatori et al., 2026). Closest to our output, MINCU associates a caption and entity labels with paired boxes for one controlled change (Mu et al., 2026). M^{2}Change asks for these links across a variable-size set of heterogeneous changes in natural-image pairs, including changes that involve several entities.

Cross-image region grounding. Region-aware MLLMs connect language to coordinates or visual regions (Peng et al., 2024; Chen et al., 2023; You et al., 2024). Migician extends grounding to free-form references across multiple images (Li et al., 2025). GeM-VG further covers varied target counts and image relations (Zheng et al., 2026). Panoptic Pairwise Distortion Graph represents matched regions and their distortion type, severity, and relative quality across an image pair (Janjua et al., 2026). Together, these works support multi-image localization and structured regional comparison. Grounded change-set prediction also requires discovering a variable number of semantic edits, describing each one, and binding its side-specific labels and boxes to the right description. A change can involve several regions or have evidence on only one side.

On-policy distillation and reinforcement learning. On-policy distillation scores student-generated responses with a teacher rather than relying only on fixed teacher samples (Agarwal et al., 2024). Vision-OPD uses privileged visual regions, and Masters uses progressive masking to focus multimodal distillation (Yuan et al., 2026; Lee et al., 2026). In PV-OPD, we compare the teacher's predictions under relevant and control insets of the same image pair to weight tokens that respond to local change evidence, while the student sees the full images. GRPO optimizes sampled responses with group-relative rewards (Shao et al., 2024). Visual-RFT applies verifiable rewards to perception tasks (Liu et al., 2025b), while visual grounding methods use spatial or IoU-based feedback (Shen et al., 2025; Bai et al., 2025; Kang et al., 2026; Weng et al., 2026). GDPO normalizes each reward dimension separately (Liu et al., 2026). In PDSO, a shared predicted–reference assignment keeps coverage, description, label, and paired-region rewards aligned.

3 The M^{2}Change Benchmark

3.1 Task and Change Representation

Given an original image and its changed counterpart, an MLLM must identify every change, describe it, and locate its visual evidence across the pair. We denote the ordered images by $X = (I_{\mathrm{A}}, I_{\mathrm{B}})$ and the predicted set of changes by $\widehat{\mathcal{C}}$. Each reference set contains $K \in \{2, 3, 4\}$ changes:

$$\mathcal{C}^* = \{c_i^*\}_{i=1}^{K}, \qquad c_i^* = (a_i, B_i^{\mathrm{A}}, L_i^{\mathrm{A}}, B_i^{\mathrm{B}}, L_i^{\mathrm{B}}), \tag{1}$$

Here a_i describes a change, B_i^{A} and B_i^{B} locate its evidence, and L_i^{A} and L_i^{B} name the entities in the corresponding regions. Boxes use normalized $xyxy$ coordinates in $[0, 1000]$ with a top-left origin. Each pair combines changes from the *object*, *attribute*, *counting*, and *relation* categories, with each category used at most once. The unit of prediction is a change, not a region: one counting change, for example, may involve several objects. When an object appears or disappears, only the image containing it has an annotated region and label. If a change is confined to part of an object, the annotation follows that part. Appendix A details these conventions.

At inference, the MLLM sees the pair without a change count or categories. Predictions may appear in any order, as evaluation treats them as a set. Appendix B.3 gives the response format.

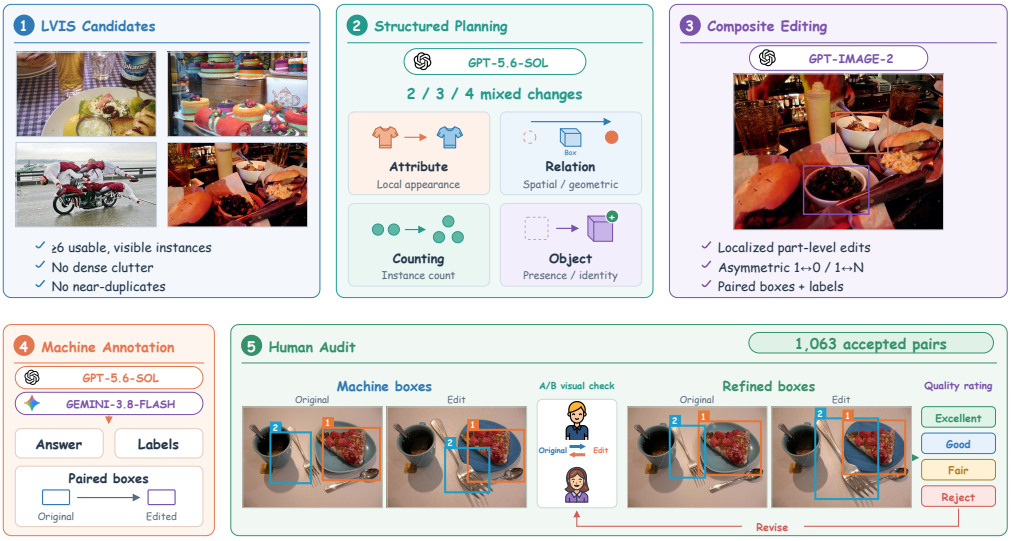


Figure 2: Construction of M^{2}Change from LVIS photographs to paired change annotations.

3.2 Construction and Quality Control

To create pairs with several distinct changes in one scene, we start from LVIS photographs (Gupta et al., 2019) (Figure 2). We select scenes with enough separate targets for multiple changes. For each photograph, a seeded planner chooses 2–4 categories and compatible targets, using each source instance at most once. The changes are then generated together while preserving the viewpoint, illumination, background, and other scene content.

We annotate the changes visible in the generated pair, which may differ from the plan. LVIS regions initialize annotations in the original image, while automatic proposals initialize regions in the changed image. Descriptions and entity labels are drafted from the plan and checked against the pair. After automated checks for valid coordinates and matching regions and labels, reviewers inspect both images, revise the annotations, and reject pairs with unresolved changes. This process yields 1,063 image pairs with 2,525 annotated changes. Appendix A gives the selection thresholds and dataset composition.

3.3 Comparison with Existing Difference Benchmarks

Table 1 compares what related benchmarks ask a model to produce. DiG (Tao et al., 2026) finds an unknown number of discrepancies in rendered 3D scenes, but predicts boxes rather than describing what each discrepancy means. MINCU (Mu et al., 2026) links a description and entity labels to regions in both images, but each pair contains one controlled change. M^{2}Change brings these challenges together in pairs generated from natural photographs, with 2–4 changes from different semantic categories in each pair. Because a change can span several regions and an appearance or disappearance has evidence on only one side, the model must group evidence by change while keeping distinct changes separate.

4 Paired-View Distillation and Set Optimization

4.1 Overview

Grounded image-difference understanding requires a model to identify local evidence in the image pair and recover every change with its description, labels, and regions correctly associated. We address these requirements in three stages (Figure 3). Direct-format SFT establishes the structured change-block output. PV-OPD compares a frozen teacher's predictions under change-focused and control crops to weight token-level distillation into a student that sees the image pair. GDPO then uses a shared predicted–reference assignment to optimize change-set coverage and paired grounding.

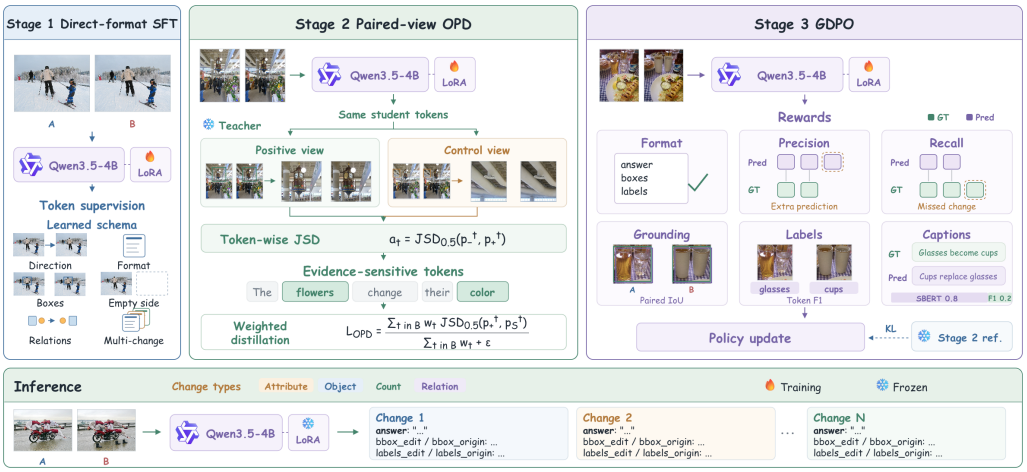


Figure 3: Overview of PDSO. SFT learns structured change blocks. PV-OPD weights token-level distillation using change-focused and control teacher views. GDPO optimizes set coverage and paired grounding through shared matching. Privileged views are used only during training.

Table 1: Comparison of visual-difference tasks. Paired boxes denote side-specific outputs, allowing an empty side for additions or removals. For photo source, ✓ indicates photographic images, ✗ synthetic-only images, and ✓✗ a mixture of both.

Benchmark	Caption	No query	Boxes	Paired boxes	Change set	Photo source
Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018)	✓	✓	✗	✗	✗	✓
CLEVR-Change (Park et al., 2019)	✓	✓	✗	✗	✗	✗
VisMin (Awal et al., 2024)	✓	✗	✗	✗	✗	✓✗
OmniDiff (Liu et al., 2025a)	✓	✓	✗	✗	✗	✓✗
DiffGround (Wang et al., 2025)	✗	✗	✓	✗	✓	✓
DiG (Tao et al., 2026)	✗	✓	✓	✗	✓	✗
MINCU (Mu et al., 2026)	✓	✓	✓	✓	✗	✓
M^{2}Change	✓	✓	✓	✓	✓	✓

4.2 Stage 1: Direct-Format SFT

Given the image pair $X_{\text{raw}} = (I_{\text{A}}, I_{\text{B}})$, let Y^* be the reference change set serialized as numbered blocks with descriptions, labels, and boxes. Absent entities have no labels or boxes. We minimize the autoregressive negative log-likelihood,

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|Y^*|} \log p_\theta \left(y_t^* \mid y_{<t}^*, X_{\text{raw}} \right). \tag{2}$$

Direct-format supervision learns variable-size structured prediction and grounding. Without a change-count prompt, the model infers set cardinality and termination for later stages.

4.3 Stage 2: Paired-View On-Policy Distillation (PV-OPD)

To focus distillation on the evidence supporting an individual change, PV-OPD applies on-policy distillation (Agarwal et al., 2024) to a matched change block. The student samples $\widehat{Y} \sim p_\theta(\cdot \mid X_{\text{raw}})$ from the image pair. We align its change blocks with the references one-to-one using paired box overlap, label similarity, and description similarity, then select one eligible match for token-level distillation. If no match is eligible, the example uses the reference sequence with the SFT objective.

For the selected change, we form two teacher inputs that retain the same full-image pair and differ in their added crops. The positive input X_+ adds crops around the matched change using a shared padded window in normalized coordinates, preserving correspondence across the two images. The

control input X_- instead adds unrelated crops. The frozen teacher scores the same student-generated sequence under both inputs. We use the difference between its token distributions to weight distillation within the selected block.

Let $\mathcal{B}$ be the selected change block's token span. At each $t \in \mathcal{B}$, p_S^t is the student's next-token distribution, while p_+^t and p_-^t are the frozen teacher's distributions under X_+ and X_-. We compute the token-wise view advantage and normalize it within the block:

$$a_t = \mathrm{JSD}_{0.5}(p_-^t, p_+^t), \qquad w_t = \mathrm{clip}\left(\frac{a_t}{\mathrm{mean}_{j\in\mathcal{B}}(a_j) + \epsilon}, 0, 4\right). \tag{3}$$

We use w_t to weight distillation toward the teacher under the positive input:

$$\mathcal{L}_{\mathrm{OPD}} = \frac{\sum_{t\in\mathcal{B}} w_t \,\mathrm{JSD}_{0.5}(p_+^t, p_S^t)}{\sum_{t\in\mathcal{B}} w_t + \epsilon}, \tag{4}$$

Tokens whose teacher predictions differ between the two views receive greater distillation weight.

4.4 Stage 3: GDPO with Set-Level Rewards

PV-OPD focuses token-level distillation on a selected change block, while the final response must cover all changes and keep their descriptions and regions correctly associated. Stage 3 therefore uses GDPO (Liu et al., 2026) to optimize set-level rewards. For each image pair, the policy samples G completions and parses each into change blocks.

We match predicted blocks p to reference blocks g using an affinity that combines paired grounding, labels, and captions. On side s, $b^s(p, g)$ sums the IoUs of one-to-one matched boxes (Kuhn, 1955) and divides by the larger box count. Two empty box lists score one; an empty/nonempty pair scores zero. We combine the two sides as $b(p, g) = \sqrt{b^{\mathrm{A}}(p, g)b^{\mathrm{B}}(p, g)}$. Label agreement $l(p, g)$ is the mean side-specific token-F1, while caption agreement $d(p, g)$ combines Sentence-BERT (Reimers & Gurevych, 2019) and token-F1 with weights 0.8 and 0.2. The assignment affinity is

$$q(p, g) = \begin{cases} b(p, g)\big(0.3l(p, g) + 0.7d(p, g)\big), & \text{if spatial evidence exists,} \\ 0.3l(p, g) + 0.7d(p, g), & \text{otherwise.} \end{cases} \tag{5}$$

Maximizing total affinity gives a shared one-to-one assignment $\mathcal{M}^*$ for set precision, set recall, paired grounding, label, and caption rewards. Format validity is scored separately. This assignment credits a caption and its regions to the same reference change, while precision and recall penalize extra and missing blocks. Appendix B.2 gives the reward definitions.

The six rewards can vary in scale. Following GDPO, we normalize each reward separately across the G sampled completions before combining them into an advantage,

$$\widehat{r}_{i,m} = \frac{r_{i,m} - \mu_m}{\sigma_m + \epsilon}, \qquad A_i = \sum_m \alpha_m \widehat{r}_{i,m}, \tag{6}$$

Here μ_m and σ_m are the group mean and standard deviation for reward m, and α_m sets its weight. We use A_i in a clipped policy-ratio objective (Schulman et al., 2017) with KL regularization toward the frozen Stage 2 policy. Completions that cannot be parsed into change blocks receive zero for all six raw rewards.

5 Experiments

5.1 Experimental Setup

On the 1,063-pair test set of M^{2}Change, we compare PDSO with closed-source general-purpose, open-weight general-purpose, and task-specialized MLLMs. Each model predicts the complete change set from an image pair, including a description, entity labels, and regions for each change. Tables 2 and 3 report joint quality, localization, and language scores.

Training. We train Qwen3.5-4B (Qwen Team, 2026a) with SFT on mixed single- and multi-change examples, followed by PV-OPD and GDPO. Direct-format SFT provides the same-backbone baseline. Appendix B.1 gives the optimization settings.

Table 2: Main results on M^{2}Change. Bold marks the best score within each model group.

Model	MCJS↑	mAP↑	AP50↑	mIoU↑	Acc@.50↑	BERTScore-F1↑	Label-F1↑
Closed-source general-purpose MLLMs							
Qwen3.8-Max (Qwen Team, 2026b)	**41.4**	**26.2**	**45.6**	**49.0**	**28.6**	90.7	50.7
Gemini 3.8-Flash (Google, 2026)	5.4	0.5	1.0	12.1	4.3	91.0	49.2
GPT-5.6-Sol (OpenAI, 2026a)	31.7	12.5	28.6	37.3	18.6	**91.4**	**53.0**
GPT-5.6-Terra (OpenAI, 2026b)	26.7	11.5	26.2	33.4	16.1	91.1	49.2
Grok 4.6 (xAI, 2026)	12.8	1.4	6.0	22.1	5.6	90.7	47.1
Open-weight general-purpose MLLMs							
Qwen3.5-4B (Qwen Team, 2026a)	10.2	4.6	10.6	22.5	5.4	90.4	41.8
Qwen3.8-27B (Qwen Team, 2026b)	25.5	**11.1**	**25.2**	33.4	**16.3**	90.5	50.0
Kimi-K3 (Kimi Team, 2026)	**28.1**	10.6	20.5	**33.5**	15.9	**90.8**	**51.4**
DeepSeek-V4.1-Flash (DeepSeek-AI, 2026)	11.6	4.8	8.8	15.6	3.7	89.1	34.6
MiniMax-M3 (MiniMax, 2026)	1.5	0.1	0.4	9.3	1.1	90.1	36.8
GLM-5.3-Flash (GLM-5-Team, 2026)	8.4	1.8	3.9	10.9	5.1	**90.8**	45.6
Task-specialized MLLMs							
EGM-4B (SFT) (Zhan et al., 2026)	28.2	15.3	30.2	31.2	7.6	90.9	51.9
EGM-8B (SFT) (Zhan et al., 2026)	26.3	15.3	30.6	28.5	3.8	90.7	54.7
Migician (SFT) (Li et al., 2025)	25.9	11.7	23.8	28.4	4.2	90.6	51.8
OddGrid-GRPO (SFT) (Weng et al., 2026)	20.4	10.9	21.3	23.1	1.5	90.5	48.3
ReGround (SFT) (Peng et al., 2026)	12.7	1.9	6.4	17.9	1.3	90.8	52.5
UniVG-R1 (SFT) (Bai et al., 2025)	23.9	11.3	22.7	27.6	4.8	90.4	52.1
VLM-R1 (SFT) (Shen et al., 2025)	8.3	1.0	3.6	13.2	0.7	**91.5**	51.4
SG-ISA (SFT) (Mu et al., 2026)	28.3	12.9	24.2	33.8	12.2	89.0	48.4
Qwen3.5-4B (SFT) (Qwen Team, 2026a)	36.4	17.0	31.7	37.3	13.1	89.6	51.6
PDSO (Qwen3.5-4B) (ours)	**47.2**	**25.8**	**48.4**	**49.4**	**27.6**	90.7	**55.1**

Table 3: Additional localization and language metrics on M^{2}Change. Bold marks the best score within each model group. Dashes indicate unavailable scores.

Model	Localization				Language			
	AP75↑	Box-F1@.50↑	Box-F1@.75↑	Acc@.75↑	BLEU-4↑	METEOR↑	CIDEr↑	SPICE↑
Closed-source general-purpose MLLMs								
Qwen3.8-Max	**26.6**	**64.4**	**48.5**	**19.4**	**7.3**	**16.9**	0.36	20.4
Gemini 3.8-Flash	0.4	8.7	4.8	3.2	6.6	16.0	**0.39**	20.8
GPT-5.6-Sol	8.9	49.8	26.1	8.6	6.9	15.1	0.36	**22.0**
Grok 4.6	0.2	22.3	3.3	1.3	6.4	14.8	0.32	19.2
Open-weight general-purpose MLLMs								
Qwen3.5-4B	3.7	28.8	15.5	2.7	4.1	11.9	0.18	–
Kimi-K3	**9.8**	**38.3**	**25.1**	**10.0**	**7.6**	**16.4**	**0.40**	**21.4**
Task-specialized MLLMs								
EGM-4B (SFT)	13.7	41.0	26.6	4.5	3.8	13.6	0.20	17.2
EGM-8B (SFT)	13.6	40.2	24.7	2.6	3.1	13.7	0.09	16.9
SG-ISA	12.3	43.5	29.3	7.7	3.7	13.6	0.18	–
Qwen3.5-4B (SFT)	16.4	47.9	33.1	8.1	4.0	16.1	0.24	18.0
PDSO (ours)	**24.8**	**62.7**	**43.1**	**14.5**	**6.9**	**16.3**	**0.46**	**21.6**

Evaluation metrics. MCJS is our primary metric, combining matched description–label–region quality (Q) and change-set coverage (C): $\mathrm{MCJS} = 100\,\mathrm{HM}(Q, C)$, where HM is the harmonic mean. Grounded-F1 requires semantic matching and grounding on both image sides. To examine individual prediction components, we report localization measures (Lin et al., 2014), caption scores (BLEU-4, METEOR, CIDEr, and SPICE), BERTScore-F1, and Label-F1. AP uses unit box-confidence scores. Appendix C defines the metrics and matching procedures.

5.2 Quantitative Results

PDSO achieves the highest MCJS among the evaluated models at 47.2, improving by 10.8 points over direct-format SFT with the same backbone and exceeding the strongest evaluated baseline by 5.8 points (Table 2). Grounded-F1 also rises from 18.0 to 25.6 (Appendix D), supporting the effectiveness of the full training sequence for grounded change-set prediction.



Figure 4: Predictions on three image pairs. Rows show reference annotations, GPT-5.6-Sol (OpenAI, 2026a), Qwen3.8-Max (Qwen Team, 2026b), and our model. Matching colors link each change's regions in the original and edited images.

The improvement extends to the components of each change prediction. PDSO achieves the highest AP50 (48.4), mIoU (49.4), and Label-F1 (55.1), compared with 31.7, 37.3, and 51.6 for direct-format SFT. These gains combine better localization and entity-label agreement. The joint and component scores also give different rankings: PDSO leads in MCJS even though its mAP of 25.8 is below the highest reported 26.2. Pooled box metrics disregard change membership, whereas MCJS assesses descriptions, labels, and regions within matched blocks. Thus, localization alone does not characterize grounded change-set quality.

Under stricter overlap requirements, PDSO also improves over SFT: AP75 rises from 16.4 to 24.8, and Box-F1@.75 from 33.1 to 43.1 (Table 3). Both scores rank second among the reported models. The gains therefore extend beyond the looser AP50 threshold, while precise box alignment remains a source of error. Acc@.50 also rises from 13.1 to 27.6, measuring recovery of full region lists with equal box counts and sufficient overlap for every matched box on an image side.

Language results further show improvements in the descriptions accompanying these regions. PDSO increases CIDEr from 0.24 to 0.46 and SPICE from 18.0 to 21.6 over SFT, achieving the highest reported CIDEr (Table 3). Yet high caption similarity can coexist with poor grounded change-set quality. VLM-R1 (Shen et al., 2025) obtains 91.5 BERTScore-F1 but only 8.3 MCJS, compared with 90.7 and 47.2 for PDSO. Caption metrics concatenate descriptions within each image pair and do not assess their association with regions. Together, these results show better descriptions and stronger agreement between changes and their supporting regions.

5.3 Qualitative Results

We use Figure 4 to examine change coverage and paired localization in three scenes. In the shop, Qwen3.8-Max (Qwen Team, 2026b) omits the blue-marked sign change, while PDSO includes the corresponding sign regions in both images. In the room, our model keeps three changes separate, with region groups for the bed, small red-marked objects, and green-marked object replacement. These examples illustrate recovery of multiple changes within the same scene.

Recovering a change also requires locating its corresponding evidence across the pair. At the railway platform, Qwen3.8-Max places its green box on a small train detail in the original image. PDSO instead marks the train region in both views, more closely following the reference. This comparison illustrates how paired localization complements change coverage: a detected change must be linked to the appropriate regions on each image side.

Table 4: Training stages and paired-view distillation ablations on M^{2}Change. Bold marks the best score in each column.

(a) Training stages.

Stage	MCJS↑	Set-F1↑	Gr-F1↑	mIoU↑
SFT	36.4	79.5	18.0	37.3
SFT+PV-OPD	33.2	79.7	16.8	35.3
PDSO (ours)	**47.2**	**85.6**	**25.6**	**49.4**

(b) Paired-view distillation variants.

Stage 2 variant	MCJS↑	Set-F1↑	Gr-F1↑	mIoU↑
w/o control view	31.0	**79.9**	14.1	34.5
w/o view weighting	29.1	79.8	14.3	33.2
PV-OPD	**32.1**	79.8	**15.8**	**34.8**

Table 5: Reward ablations on M^{2}Change. R_{set} denotes set precision and recall, and R_{pg} paired grounding. Bold marks the best score in each column.

Reward variant	MCJS↑	Set-F1↑	Gr-F1↑	mIoU↑
Scalar reward	45.5	83.6	24.0	47.9
GDPO w/o R_{set}	44.9	83.3	23.1	45.9
GDPO w/o R_{pg}	45.4	84.1	24.2	47.6
GDPO w/o shared $\mathcal{M}^*$	27.1	46.6	13.9	20.2
PDSO (ours)	**47.2**	**85.6**	**25.6**	**49.4**

5.4 Ablation Studies

Training stages. Table 4a shows how performance changes as training moves from output formatting to local evidence and then to set-level optimization. PV-OPD uses token-level distillation to expose the student to change-focused evidence, but its intermediate checkpoint scores 33.2 MCJS and 16.8 Grounded-F1, below the SFT scores of 36.4 and 18.0. Local evidence supervision alone therefore does not improve recovery of the complete grounded set. After GDPO optimizes coverage and grounding together, PDSO reaches 47.2 MCJS and 25.6 Grounded-F1. The measured gain thus appears only after set-level optimization.

Contribution of paired-view OPD. The control view helps distinguish tokens supported by a change-relevant inset from those the teacher would produce using unrelated local content. Adding it raises MCJS from 31.0 to 32.1 and Grounded-F1 from 14.1 to 15.8, while Set-F1 stays near 79.8 (Table 4b). The larger change in Grounded-F1 suggests that this contrast chiefly improves how descriptions are tied to regions, rather than how many changes are predicted.

Reward design. Removing the shared assignment causes the largest drop in Table 5: MCJS falls to 27.1, versus 44.9–45.5 for the other reward variants, while Set-F1 and Grounded-F1 fall to 46.6 and 13.9. Different rewards can then credit a predicted description, label, and region against different reference changes. This feedback may favor individually plausible outputs over coherent change blocks, consistent with the loss in both set recovery and grounding. Using $\mathcal{M}^*$ aligns these signals, and the full method reaches 47.2 MCJS.

6 Conclusion

We introduced M^{2}Change for describing and grounding multiple image changes. Its annotations link descriptions and entity labels to supporting regions for evaluating complete change sets. We also proposed PDSO, which combines structured fine-tuning, paired-view distillation, and set-level optimization to describe and locate these changes together. PDSO improves MCJS by 10.8 points over direct-format SFT, supporting its effectiveness for grounded change-set prediction. Future work can extend this setting to naturally occurring changes and less constrained image pairs.

AI USE STATEMENT

We use generative AI tools to assist with benchmark construction, including data generation, editing, and annotation. We also use generative AI tools to polish the writing and improve the clarity of the manuscript.

ETHICS STATEMENT

The benchmark derives source images and annotations from LVIS and will preserve the applicable source licenses and attribution. We do not use the benchmark for identity inference or sensitive-attribute prediction. Before release, we will audit the data for personally identifying or harmful content, publish provenance and checksums, and document the filtering and removal process.

REPRODUCIBILITY STATEMENT

The appendix provides additional details of the paper, including benchmark construction and annotation, model configurations, reward definitions, and evaluation procedures. Upon acceptance, we will release the dataset and code to support reproducibility and future research.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398, 2016.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107795–107829, 2024.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. UniVG-R1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025. URL https://arxiv.org/abs/2505.14231.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. URL https://arxiv.org/abs/2306.15195.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash.

GLM-5-Team. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. URL https://arxiv.org/abs/2602.15763.

Google. Gemini 3.8 Flash, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash. Gemini API documentation. Accessed: 2026-09-18.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Muhammad Kamran Janjua, Abdul Wahab, and Bahador Rashidi. Panoptic pairwise distortion graph. In *International Conference on Learning Representations*, 2026.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4024–4034, 2018.

Weitai Kang, Jason Kuen, Mengwei Ren, Zijun Wei, Yan Yan, and Kangning Liu. VGent: Visual grounding via modular design for disentangling reasoning and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41160–41170, 2026.

Jae-Woo Kim, Hyeongbeom Kim, and Ue-Hwan Kim. C3-Bench: A context-aware change captioning benchmark. In *European Conference on Computer Vision*, pp. 248–269, 2026a.

Minkyu Kim, Sangheon Lee, and Dongmin Park. VLM-SubtleBench: How far are VLMs from human-level subtle comparative reasoning? In *International Conference on Learning Representations*, 2026b.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Byung-Kwan Lee, Yu-Chiang Frank Wang, and Ryo Hachiuma. Masking teacher and reinforcing student for distilling vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10126–10141, 2026.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, and Maosong Sun. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9845–9867, 2025.

Benedetta Liberatori, Nermin Samet, Paolo Rota, Matthieu Cord, Elisa Ricci, Andrei Bursuc, and Monika Wysoczańska. Spot-the-shift: Evaluating grounded image difference captioning of long-term changes. *arXiv preprint arXiv:2609.10356*, 2026. URL https://arxiv.org/abs/2609.10356.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026. URL https://arxiv.org/abs/2601.05242.

Yuan Liu, Saihui Hou, Saijie Hou, Jiabao Du, Shibei Meng, and Yongzhen Huang. OmniDiff: A comprehensive benchmark for fine-grained image difference captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21440–21449, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality. MiniMax Research, June 2026. URL https://www.minimax.io/blog/minimax-m3.

Chaoqian Mu, Wenhao Wu, Zichen Liang, Jiaxu Li, Lijun Wang, Yifan Wang, and Huchuan Lu. MinCU: A fine-grained benchmark for grounded minimal-change understanding in image pairs, 2026. URL https://arxiv.org/abs/2609.23336.

OpenAI. GPT-5.6 Sol model, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Model documentation; accessed July 29, 2026.

OpenAI. GPT-5.6 Terra model, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Model documentation; accessed July 29, 2026.

OpenAI. Gpt-image-2 model card, 2026. URL https://platform.openai.com/docs/models/gpt-image-2.

Kaihang Pan, Weile Chen, Haiyi Qiu, Qifan Yu, Wendong Bu, Zehan Wang, Yun Zhu, Juncheng Li, and Siliang Tang. WiseEdit: Benchmarking cognition- and creativity-informed image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37268–37278, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4624–4633, 2019.

Lei Peng, Shuai Lv, and Wei Hu. ReGround: Restoring visual grounding in multi-step reasoning through self-diagnosis and visual re-examination, 2026. URL https://arxiv.org/abs/2608.04385.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 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)*, pp. 3982–3992, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. URL https://arxiv.org/abs/2504.07615.

Zhou Tao, Shida Wang, YongXiang Hua, Haoyu Cao, and Linli Xu. DiG: Differential grounding for enhancing fine-grained perception in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1695–1705, 2026.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Wenxuan Wang, Zijia Zhao, Yisi Zhang, Yepeng Tang, Erdong Hu, Xinlong Wang, and Jing Liu. Image difference grounding with natural language. *arXiv preprint arXiv:2504.01952*, 2025. URL https://arxiv.org/abs/2504.01952.

Tengjin Weng, Wenhao Jiang, Jingyi Wang, Ming Li, Lin Ma, and Zhong Ming. OddGridBench: Exposing the lack of fine-grained visual discrepancy sensitivity in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1575–1584, 2026.

xAI. Model card: Grok 4.6, 2026. URL https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf. Official model card.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026. URL https://arxiv.org/abs/2605.18740.

Guanqi Zhan, Changye Li, Zhijian Liu, Yao Lu, Yi Wu, Song Han, and Ligeng Zhu. EGM: Efficient visual grounding language models. In *European Conference on Computer Vision*, pp. 300–317, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Shurong Zheng, Yousong Zhu, Hongyin Zhao, Fan Yang, Yufei Zhan, Ming Tang, and Jinqiao Wang. GeM-VG: Towards generalized multi-image visual grounding with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28857–28865, 2026.

A Benchmark Construction and Annotation

Starting from LVIS photographs (Gupta et al., 2019), the pipeline has five stages: source selection, change planning, composite image editing, automatic annotation, and human review. The following details specify how these stages produce pairs with several distinct, verifiable changes.

Change taxonomy. Object changes cover category or identity replacement, insertion, and removal. Attribute changes cover color, material, texture, shape, ordinary or functional state, and text, number, or symbol edits. Counting changes include same-class increases and decreases, duplication, and split or merge operations. Relation changes include translation, scale, rotation, relative position, containment, contact or attachment, ordering, and occlusion or visibility. Fine-grained operations are sampled according to their feasibility in the source scene, while every planned change retains its assigned coarse category.

Candidate filtering. To support several independent changes in one scene, we require a source image with a short side of at least 400 pixels and at least six usable instances. A usable instance has an area ratio in $[0.002, 0.20]$, a box short side of at least 18 pixels, normalized width and height of at least 0.035, and limited truncation. Each source must offer at least five independent targets, three categories, a repeated category for counting, and three eligible relation pairs. Independent targets have pairwise IoU at most 0.35. Of 17,514 matched LVIS images, 5,663 pass these filters, from which 4,500 are sampled. The capacity tiers contain 2,250, 1,350, and 900 candidates able to support up to two, three, and four changes, respectively.

Change planning and image editing. For capacity tiers two, three, and four, the probabilities of planning two, three, or four distinct change categories are $(1, 0, 0)$, $(0.35, 0.65, 0)$, and $(0.15, 0.30, 0.55)$. A seeded planner then selects compatible LVIS targets, fine-grained operations, count metadata, and relation references without reusing a source annotation ID. Categories containing at least six small instances with a median area ratio at most 0.02 are excluded as editable targets and relation anchors. The resulting plan also identifies content to preserve during a single composite edit. GPT-5.6-Sol (OpenAI, 2026a) generates the plan, GPT-Image-2 (OpenAI, 2026) produces the changed image, and Gemini 3.8-Flash (Google, 2026) proposes regions on the changed side. Descriptions are initialized from validated plans and refined against the generated pair. Structural validation checks coordinate ranges and box–label alignment before human review. The two images are brought to common dimensions for inspection, and the final annotations describe the realized changes rather than the requested plan.

Relation visibility. Relation edits are planned to be discernible from the image pair. Translation is approximately one quarter of the relevant image axis or crosses a visible reference boundary. Rotation is at least $45°$; scale changes enlarge an entity by at least $1.5\times$ or reduce its linear size to at most two thirds. Other relation types specify distinct before-and-after states with visible reference objects.

Human review. Reviewers check whether each planned transformation occurred and whether the pair contains additional semantic changes. They then revise descriptions, boxes, and labels against both images, verifying that each block refers to one realized change. Inspecting the pair in both directions is particularly useful for insertions and removals, where one evidence list is empty. Pairs with unresolved discrepancies are rejected.

Test-set composition. The accepted test set contains 1,063 pairs: 738 with two changes, 251 with three, and 74 with four. Its 2,525 annotated changes comprise 706 attribute, 693 object, 605 counting, and 521 relation changes. Human review assigns 407 pairs an Excellent rating, 546 Good, and 110 Fair.

Evidence granularity and counting. Boxes follow the edited component when a change is confined to a part; the original whole-object LVIS box is refined when necessary. Unchanged relation anchors may be named in a description but are not included in the changed-region box lists. For example, increasing the number of cups from three to five produces one counting block with an empty original-side list and two boxes for the added cups. A decrease follows the reverse convention, whereas replacement and split or merge operations include participating regions on both sides. Consequently, the change-set cardinality and the number of evidence boxes serve different purposes; recovering full before-and-after counts requires the count metadata rather than box-list lengths alone.

B Model and Output Specifications

B.1 Training Configuration

Training starts from Qwen3.5-4B (Qwen Team, 2026a). We update it with LoRA (Hu et al., 2022) of rank 16, scaling factor 32, and dropout 0.05, while keeping the vision tower and multimodal aligner frozen. Across the three stages, we use BF16, AdamW (Loshchilov & Hutter, 2019), gradient clipping at 1, and 25 warmup steps followed by a constant learning rate. The rate is 10^{-5} for SFT and PV-OPD and 5×10^{-6} for GDPO.

In the three-stage sequence, PV-OPD initializes from SFT and GDPO initializes from the resulting PV-OPD checkpoint. Both on-policy stages sample with temperature 1 and a maximum response length of 2,048 tokens. GDPO draws four responses per prompt and uses a KL coefficient of 0.02. Gradient accumulation is 2 for SFT and PV-OPD and 4 for GDPO.

Ablation settings. In Table 4a, each stage continues from the preceding checkpoint. The PV-OPD reference in Table 4b is a separate Stage 2 run. The variant without view-advantage weighting uses a shorter budget of 2,000 updates. In Table 5, the GDPO reward-variant adapters are merged

with SFT+PV-OPD for evaluation, while the PDSO row reports the sequential model. The variant without R_{set} uses 1,750 updates rather than 3,300.

B.2 Set-Level GDPO Rewards

A single Hungarian assignment (Kuhn, 1955) $\mathcal{M}^*$ maximizes total affinity between predicted and reference change blocks. Let P and T be the numbers of predicted and reference blocks, and let $U = \sum_{(p,g)\in\mathcal{M}^*} q(p,g)$ be the assigned affinity. The six reward dimensions are

$$R_{\text{fmt}} = \mathbf{1}[\text{strictly parseable}], \quad R_{\text{prec}} = \frac{U}{\max(P,1)}, \quad R_{\text{rec}} = \frac{U}{T},$$
$$R_{\text{ground}} = \frac{\sum_{(p,g)\in\mathcal{M}^*} b(p,g)}{\max(T_{\text{sp}}, P_{\text{sp}}, 1)}, \quad R_{\text{label}} = \frac{\sum_{(p,g)\in\mathcal{M}^*} l(p,g)}{\max(P,T)}, \quad R_{\text{desc}} = \frac{\sum_{(p,g)\in\mathcal{M}^*} d(p,g)}{\max(P,T)}. \tag{7}$$

Here T_{sp} and P_{sp} count reference and predicted blocks with at least one supervised box side. In the reward ablations, R_{set} refers jointly to precision and recall, while R_{pg} refers to paired grounding. Because all reward terms use the same assignment, omissions reduce recall, redundant predictions reduce precision, and a correct block count alone cannot compensate for incorrect change–evidence associations.

B.3 Structured Output Schema

Each predicted change is serialized as a separate block:

```
<change_i>
<answer>One concise description.</answer>
<bbox_edit>[[x1,y1,x2,y2], ...]</bbox_edit>
<bbox_origin>[[x1,y1,x2,y2], ...]</bbox_origin>
<labels_edit>["..."]</labels_edit>
<labels_origin>["..."]</labels_origin>
</change_i>
```

The `origin` and `edit` fields refer to I_{A} and I_{B}, respectively. Boxes use normalized $xyxy$ coordinates in $[0, 1000]$, with one label per box and empty lists for an absent entity. Field order within a block is fixed, but the order of change blocks is immaterial. The parser rejects missing or repeated fields, invalid coordinates, free text outside the blocks, and box–label cardinality mismatches.

C Evaluation Protocol

Reported metrics use percentages, except CIDEr on its standard scale. In tables, Gr-F1 denotes Grounded-F1.

Localization and semantics. For component localization, boxes are pooled across change blocks within each image while retaining their image side. We report mAP, AP50, AP75, mIoU, strict box accuracy, and class-aware Label-Box AP. At IoU 0.50, strict accuracy requires equal predicted and reference box counts and every Hungarian-matched box to reach the threshold; scores are averaged across the two image sides. Label-Box AP additionally requires Label-F1≥ 0.50 for each matched label pair. Language measures comprise BLEU-4, METEOR, CIDEr, SPICE, BERTScore-F1, and Label-F1 (Papineni et al., 2002; Banerjee & Lavie, 2005; Vedantam et al., 2015; Anderson et al., 2016; Zhang et al., 2020). These scores describe individual aspects of a prediction without requiring the caption and boxes to refer to the same change.

Joint change-set score. To assess those links, we align predicted and reference change blocks one-to-one. For a predicted block p_i and reference block g_j, let s_{ij}^{desc} denote BERTScore-F1, s_{ij}^{lab} the mean side-specific Label-F1, and b_{ij}^{A} and b_{ij}^{B} the corresponding Box-F1@0.50 scores. The assignment $\mathcal{M}$ maximizes $A_{ij} = \frac{1}{2}(s_{ij}^{\text{desc}} + s_{ij}^{\text{lab}})$. Each assigned pair then receives

$$b_{ij} = \text{HM}(b_{ij}^{\text{A}}, b_{ij}^{\text{B}}), \qquad q_{ij} = \text{HM}(s_{ij}^{\text{desc}}, s_{ij}^{\text{lab}}, b_{ij}), \tag{8}$$

Table 6: AP and overlap metrics on M^{2}Change.

Model	mAP↑	AP50↑	AP75↑	AP90↑	AP95↑	mIoU↑	Label-Box AP50↑
Closed-source general-purpose MLLMs							
Qwen3.8-Max	**26.2**	**45.6**	**26.6**	**5.3**	**0.7**	**49.0**	**13.8**
Gemini 3.8-Flash	0.5	1.0	0.4	0.2	0.2	12.1	0.4
GPT-5.6-Sol	12.5	28.6	8.9	1.8	0.2	37.3	9.7
GPT-5.6-Terra	11.5	26.2	9.0	1.1	0.2	33.4	7.7
Grok 4.6	1.4	6.0	0.2	0.0	0.0	22.1	1.5
Open-weight general-purpose MLLMs							
Qwen3.5-4B	4.6	10.6	3.7	0.4	0.1	22.5	2.6
Qwen3.8-27B	**11.1**	**25.2**	7.8	1.3	**0.5**	33.4	**7.8**
Kimi-K3	10.6	20.5	**9.8**	**2.3**	0.2	**33.5**	7.0
DeepSeek-V4.1-Flash	4.8	8.8	4.7	0.6	0.1	15.6	2.2
MiniMax-M3	0.1	0.4	0.0	0.0	0.0	9.3	0.1
GLM-5.3-Flash	1.8	3.9	1.3	0.1	0.1	10.9	1.1
Task-specialized MLLMs							
EGM-4B (SFT)	15.3	30.2	13.7	1.9	0.1	31.2	10.3
EGM-8B (SFT)	15.3	30.6	13.6	2.1	0.3	28.5	11.0
Migician (SFT)	11.7	23.8	9.8	1.6	0.0	28.4	8.8
OddGrid-GRPO (SFT)	10.9	21.3	10.0	1.4	0.2	23.1	7.4
ReGround (SFT)	1.9	6.4	1.2	0.0	0.0	17.9	2.2
UniVG-R1 (SFT)	11.3	22.7	9.7	1.7	0.1	27.6	8.8
VLM-R1 (SFT)	1.0	3.6	0.6	0.0	0.0	13.2	1.6
SG-ISA (Qwen3.5-4B)	12.9	24.2	12.3	2.4	0.2	33.8	9.6
Qwen3.5-4B (SFT)	17.0	31.7	16.4	3.0	**0.4**	37.3	12.8
PDSO	**25.8**	**48.4**	**24.8**	**3.6**	0.3	**49.4**	**18.9**

where HM is the harmonic mean and is zero if any input is zero. On a given image side, two empty box sets score one, while an empty/nonempty mismatch scores zero. Let $M = |\mathcal{M}|$, and let N_G and N_P be the numbers of reference and predicted changes. We compute

$$Q = \frac{\sum_{(i,j)\in\mathcal{M}} q_{ij}}{\max(N_G, 1)}, \qquad C = \frac{M}{\max(N_G, N_P, 1)}, \qquad \mathrm{MCJS} = 100\,\mathrm{HM}(Q, C). \tag{9}$$

The reported MCJS is averaged over image pairs. Q measures the quality of matched descriptions, labels, and regions; C discounts omitted or redundant change blocks. One-to-one alignment makes the score invariant to output order.

Thresholded joint diagnostic. Grounded-F1 uses a separate assignment that maximizes Sentence-BERT description similarity (Reimers & Gurevych, 2019). An assigned pair is a true positive when similarity reaches 0.65 and Box-F1@0.50 reaches 0.50 on each image side. Label similarity does not enter this assignment. Unmatched or rejected predictions and references contribute false positives and false negatives. In contrast, Acc@.50 requires equal box counts and every matched box to reach IoU 0.50 on the evaluated side. AP uses unit scores because predictions provide no calibrated confidence. Caption metrics concatenate the descriptions within each image pair, whereas MCJS retains individual change membership.

D Additional Results

Detailed Model Results Tables 6 and 7 give the detailed localization results. Table 8 reports language and Grounded-F1 scores. Bold marks the best score within each model group. Dashes indicate unavailable scores.

Table 7: Region matching and strict accuracy on M^{2}Change.

Model	Box-F1@.50↑	Box-F1@.75↑	Box-F1@.90↑	Acc@.50↑	Acc@.75↑	Acc@.90↑
Closed-source general-purpose MLLMs						
Qwen3.8-Max	**64.4**	**48.5**	**19.8**	**28.6**	**19.4**	**6.8**
Gemini 3.8-Flash	8.7	4.8	1.6	4.3	3.2	1.6
GPT-5.6-Sol	49.8	26.1	8.1	18.6	8.6	3.2
GPT-5.6-Terra	46.1	25.8	7.1	16.1	9.1	2.9
Grok 4.6	22.3	3.3	0.2	5.6	1.3	0.8
Open-weight general-purpose MLLMs						
Qwen3.5-4B	28.8	15.5	5.4	5.4	2.7	0.7
Qwen3.8-27B	**46.7**	**25.8**	8.4	**16.3**	8.8	3.6
Kimi-K3	38.3	25.1	**9.8**	15.9	**10.0**	**4.2**
DeepSeek-V4.1-Flash	21.0	14.7	5.0	3.7	2.8	1.4
MiniMax-M3	4.6	0.3	0.0	1.1	0.8	0.8
GLM-5.3-Flash	13.8	5.1	0.7	5.1	2.1	0.9
Task-specialized MLLMs						
EGM-4B (SFT)	41.0	26.6	8.4	7.6	4.5	0.8
EGM-8B (SFT)	40.2	24.7	6.9	3.8	2.6	1.0
Migician (SFT)	38.0	23.2	6.4	4.2	2.0	0.4
OddGrid-GRPO (SFT)	31.7	20.2	6.6	1.5	0.8	0.3
ReGround (SFT)	18.0	3.4	0.1	1.3	0.1	0.1
UniVG-R1 (SFT)	36.3	22.4	6.1	4.8	2.6	0.4
VLM-R1 (SFT)	12.7	2.0	0.2	0.7	0.1	0.0
SG-ISA (Qwen3.5-4B)	43.5	29.3	9.9	12.2	7.7	2.3
Qwen3.5-4B (SFT)	47.9	33.1	11.7	13.1	8.1	1.9
PDSO	**62.7**	**43.1**	**14.0**	**27.6**	**14.5**	**3.1**

Table 8: Language and joint metrics on M^{2}Change. SPICE predictions are truncated at 128 whitespace-delimited tokens; the two parser timeouts for the historical SG-ISA SFT row are scored as zero.

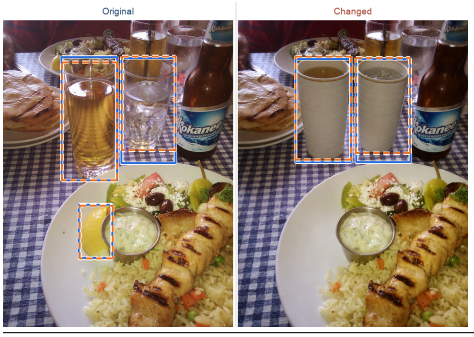
Model	BLEU-4↑	METEOR↑	CIDEr↑	SPICE↑	BERTScore-F1↑	Label-F1↑	Gr-F1↑
Closed-source general-purpose MLLMs							
Qwen3.8-Max	**7.3**	**16.9**	0.36	20.4	90.7	50.7	**20.8**
Gemini 3.8-Flash	6.6	16.0	**0.39**	20.8	91.0	49.2	2.6
GPT-5.6-Sol	6.9	15.1	0.36	**22.0**	**91.4**	**53.0**	18.1
GPT-5.6-Terra	5.5	13.4	0.27	19.2	91.1	49.2	14.7
Grok 4.6	6.4	14.8	0.32	19.2	90.7	47.1	5.5
Open-weight general-purpose MLLMs							
Qwen3.5-4B	4.1	11.9	0.18	–	90.4	41.8	7.7
Qwen3.8-27B	5.3	14.2	0.25	18.5	90.5	50.0	**13.8**
Kimi-K3	**7.6**	**16.4**	**0.40**	**21.4**	**90.8**	**51.4**	11.3
DeepSeek-V4.1-Flash	2.9	10.6	0.11	11.4	89.1	34.6	2.1
MiniMax-M3	4.6	14.2	0.20	15.9	90.1	36.8	1.0
GLM-5.3-Flash	2.4	7.8	0.11	6.5	**90.8**	45.6	4.9
Task-specialized MLLMs							
EGM-4B (SFT)	3.8	13.6	0.20	17.2	90.9	51.9	13.9
EGM-8B (SFT)	3.1	13.7	0.09	16.9	90.7	54.7	13.5
Migician (SFT)	3.5	13.4	0.19	15.7	90.6	51.8	12.7
OddGrid-GRPO (SFT)	2.7	13.7	0.08	15.3	90.5	48.3	9.0
ReGround (SFT)	2.9	14.4	0.18	15.4	90.8	52.5	5.8
UniVG-R1 (SFT)	3.4	13.2	0.16	15.2	90.4	52.1	11.2
VLM-R1 (SFT)	3.4	13.1	0.17	15.1	**91.5**	51.4	3.3
SG-ISA (Qwen3.5-4B)	3.7	13.6	0.18	–	89.0	48.4	14.4
Qwen3.5-4B (SFT)	4.0	16.1	0.24	18.0	89.6	51.6	18.0
PDSO	**6.9**	**16.3**	**0.46**	**21.6**	90.7	**55.1**	**25.6**

E FAILURE CASES

We checked the PDSO outputs against the reference annotations and both images. Figure 5 presents three verified errors: splitting one change into two predictions, omitting changes, and assigning the wrong object identity across the image pair.

Boxes: blue solid, reference; orange dashed, PDSO. In (b), boxes mark only the omitted reference regions.

(a) Change splitting



Ground truth

1. The large lemon wedge and small lemon garnish present in Image A are absent in Image B.

2. The beer tumbler and water tumbler change from transparent glass in Image A to opaque ceramic in Image B.

PDSO

1. The transparent glass changed to a white ceramic cup.

2. The lemon wedge was removed from the plate.

3. The transparent glass changed to a white ceramic cup.

(b) Change omissions



Ground truth

1. The lower book is removed.

2. The serving plate changes from ceramic to wood.

3. Banana pieces increase from five to nine.

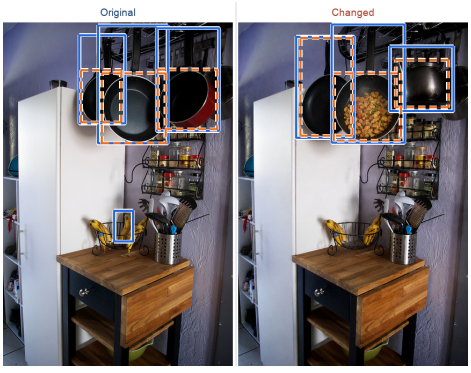
4. The soda can moves from the left foreground to the center table, swapping places with the blackboard.

PDSO

1. The soda can moves from the table edge to the background.

2. The number of banana slices increases.

(c) Cross-image object association



Ground truth

1. The frying pan is larger in Image B than in Image A.

2. The frying pan changes from empty in Image A to containing cooked food in Image B.

3. The pot in Image A is replaced by a kettle in Image B.

4. Image B has one fewer banana than Image A.

PDSO

1. The red frying pan changed to a black teapot.

2. The two frying pans changed to a black frying pan and a wok containing food.

Figure 5: Three verified PDSO failures on M^{2}Change. (a) One material change is split into two predictions. (b) PDSO misses the removed book and plate material change. (c) PDSO misidentifies a pot as a frying pan and a kettle as a teapot, merges two pan changes, and misses the banana removal. Boxes highlight the regions relevant to each error.

F LIMITATIONS AND DISCUSSION

M^{2}Change applies generated changes to real photographs, so subtle editing artifacts or unintended changes may remain after review. The two image sides also have different annotation starting points: original-image regions inherit LVIS annotations, while changed-image regions begin with automatic proposals. Manual correction reduces, but cannot eliminate, this source of asymmetry.

The benchmark measures visible, localized changes with controlled compositional complexity. It covers 2–4 changes per pair, at most one from each of four coarse categories, and does not include unchanged pairs or changes that require external knowledge or long temporal context. These categories do not exhaust the range of possible visual changes. Some combinations are harder to generate reliably than others, producing an uneven distribution of change counts and fine-grained types. Extending the benchmark to repeated same-category changes, larger sets, and no-change abstention would test different aspects of change-set recovery. Finally, PV-OPD relies on a frozen task-trained teacher and can inherit its errors; the effect of teacher quality warrants separate study.