ChainDiff: Controllable Fine-Grained Image-Pair Synthesis and Frequency-Domain Difference Attention for Image Difference Captioning

Anonymous authors
Paper under double-blind review

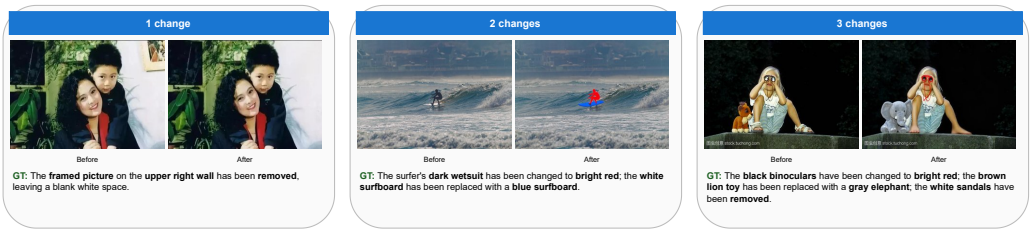

Figure 1: Representative ChainDiff pairs at three change counts (1, 2, 3). Each panel shows the before / after images and the automatically generated reference caption, with attribute keywords in bold. Changes remain at the object sub-part level (e.g. single accessory, small object addition / removal, color / material swap).

Abstract

Fine-grained Image Difference Captioning (IDC) requires precisely describing sub-part-level differences between two highly similar images. Existing work is constrained by two bottlenecks: fine-grained image pairs are scarce and their differences are uncontrollable, and mainstream Vision-Language Models (VLMs) rely on implicit alignment to discover weak difference signals hidden inside overwhelming commonality. We propose ChainDiff, a controllable image-pair synthesis pipeline that combines dual-channel similarity filtering, a hard upper bound on the changed-region ratio, and chained editing with segment sampling to make sub-part-level fine-grainedness and per-pair change counts explicitly controllable. We further propose the Frequency-Domain Difference Attention (FDDA) module, which decouples multi-scale difference signals via a feature-domain discrete cosine transform (DCT) decomposition and models differences explicitly via dual-branch image-pair cross-attention, integrated into VLMs in a plug-in manner during supervised fine-tuning (SFT). We term the resulting models, trained on ChainDiff data with FDDA integrated, HarmonicDiff. Across three base VLMs and eight test sets, HarmonicDiff attains the best CIDEr among Multimodal LLM (MLLM) methods on five test sets; specialized non-MLLM baselines remain stronger in CIDEr only on synthetic CLEVR-Change.

1 Introduction

Image Difference Captioning (IDC) aims to precisely describe, in natural language, the differences between two highly similar images, with wide applications in remote-sensing change monitoring, medical image comparison, image editing auditing, and visual assistants. In recent years, research focus has shifted from coarse-grained differences such as object addition/removal to fine-grained IDC: the changed regions are confined to the object sub-part level, and descriptions must be precise at the attribute level, such as color, texture, pattern, and addition/removal of small objects. Two major bottlenecks stand in the way of this goal.

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.

Data side: fine-grained image-pair data is expensive to construct and hard to control. Existing IDC datasets fall into two families. *Human-collected* ones — Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018), Image-Edit-Request (Tan et al., 2019) and the real subset of OmniDiff (Liu et al., 2025) — rely on crowdsourcing or manual collection and annotation, so scaling them up is costly. *Automatically synthesized* ones — e.g., Img-Diff (Jiao et al., 2025), whose object-replacement pairs come from a segmentation-and-generation pipeline — scale cheaply but fix the change type and offer no guarantee on change magnitude. Across both families, the number of changes per pair cannot be adjusted on demand, region-level change annotations are either absent or limited to single-change replacement, and the change magnitude is uncontrollable, mixed with abundant global changes that go beyond the sub-part-level operating point we target (distributional details in Table A1). Models can thus neither learn pure fine-grained difference patterns nor obtain grounding supervision.

Model side: difference perception relies on implicit learning. Mainstream VLMs directly concatenate the two images' features into the language model (LM), leaving difference discovery entirely to implicit alignment; yet in fine-grained image pairs the changed-pixel ratio is small, so the differences are weak signals drowned in the commonality, entangled with multi-scale information and hard to model independently.

To address the above bottlenecks, this paper proposes a controllable fine-grained image-pair synthesis pipeline (Figure 2): dual-channel similarity filtering and an upper bound on the changed-region ratio enforce fine-grainedness, chained editing and segment sampling make the change count explicitly controllable, and pixel differencing automatically derives bboxes and generates region-aligned descriptions; we further propose the Frequency-Domain Difference Attention (FDDA) module (Figure 3), which models image-pair differences explicitly in a plug-in manner.

The main contributions of this paper are as follows.

ChainDiff data. From 300k internet images we synthesize 29,043 fine-grained image pairs with fully automatic annotations (examples in Figure 1): dual-channel similarity filtering plus a hard upper bound on the changed-region ratio keeps samples at object sub-part level; chained editing and segment sampling make the change count explicitly determined by the segment length; pixel differencing and connected-component analysis automatically derive the bbox of every changed region, and per-region descriptions are generated via Set-of-Marks visual prompting, endowing downstream models with grounding capability.

FDDA architecture. We propose a plug-and-play frequency-domain dual-branch difference attention module that decouples multi-scale signals via a feature-domain discrete cosine transform (DCT) decomposition, explicitly models differences with dual-branch image-pair cross-attention, and integrates stably into VLMs via gated fusion during supervised fine-tuning (SFT).

Empirical results. Across three base VLMs and eight test sets, HarmonicDiff attains the best CIDEr among MLLM methods on five test sets (OmniDiff-Real, OmniDiff-Render, CLEVR-DC, Spot-the-Diff, and Birds-to-Words); the ChainDiff data and the FDDA module contribute largely consistent gains, and combined, full HarmonicDiff improves CIDEr over either ablation condition on all eight test sets.

2 Related Work

2.1 Image Difference Captioning Datasets

The IDC task originates from change captioning research. Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018) crowdsources descriptions of adjacent surveillance-video frame pairs, but the changes are uncontrollable and lack localization annotation; CLEVR-Change (Park et al., 2019) synthesizes pairs with the CLEVR engine, accompanied by the robust change captioning model DUDA; CLEVR-DC (Kim et al., 2021) extends CLEVR-family evaluation with extreme viewpoint changes. Image-Edit-Request (IEdit) (Tan et al., 2019) collects real before/after edited pairs with users' editing requests, pioneering editing-driven data

construction; Birds-to-Words (Forbes et al., 2019) provides paragraph-level descriptions over bird photo pairs, targeting fine-grained natural-language comparison in a domain-specific setting. Recent work turns to fine-grained data: Img-Diff (Jiao et al., 2025) synthesizes object-replacement contrastive data; OmniDiff (Liu et al., 2025) covers real and 3D scenes with 12 change types and 60-word human annotations but without pixel-level control; OneDiff (Hu et al., 2024) builds DiffCap in a unified instruction format.

In terms of data controllability, none of the above datasets simultaneously provides the fine-grained hard guarantee, explicit change-count controllability, and built-in localization annotations. Rendering-based corpora such as CLEVR-Change (Park et al., 2019) can control the number of injected changes at render time, but their pairs lack real-scene distributions, fine-grainedness guarantees and localization-aligned descriptions; recent instruction-editing datasets support multi-turn edits but do not target sub-part-level fine-grainedness with a hard upper bound nor emit region-aligned bboxes per change. Our pipeline is, to our knowledge, the first to *jointly* guarantee sub-part-level fine-grainedness, explicit change-count control, and region-aligned bboxes on real internet images.

2.2 Image Difference Captioning Methods

Early methods explicitly fuse both images' features in encoder–decoder architectures (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019); we group recent methods into four (non-exhaustive) directions. **Feature interaction route:** starting from DUDA (Park et al., 2019), a sequence of specialized IDC models refine dual-image feature interaction with spatial or channel attention variants — including M-VAM (Shi et al., 2020), DUDA+TIRG (Hosseinzadeh & Wang, 2021), VACC (Kim et al., 2021), MCCFormers-D (Qiu et al., 2021), SRDRL (Tu et al., 2021b), R^3Net (Tu et al., 2021a), IFDC (Huang et al., 2022), NCT (Tu et al., 2023b), VARD (Tu et al., 2023a), SCORER (Tu et al., 2023c), SMART (Tu et al., 2024c), and DIRL (Tu et al., 2024b) — and CLIP4IDC (Guo et al., 2022) introduces CLIP priors. **Preference and contrastive optimization route:** Yao et al. align visual differences with text via self-supervised pretraining (Yao et al., 2022), and Huang et al. strengthen difference sensitivity with adversarial preference optimization (Huang et al., 2025). **MLLM route:** OneDiff (Hu et al., 2024) unifies multiple IDC benchmarks into a single instruction model; FINER-MLLM (Zhang et al., 2024) and M^3Diff (Liu et al., 2025) further scale up MLLM-based IDC, with M^3Diff introducing the multi-scale differential perception (MDP) module for feature-level difference modeling. **Multi-change and domain-specific routes:** CARD (Tu et al., 2024a) generates descriptions one by one for multi-change scenarios; on Birds-to-Words, domain-specific methods include Neural Naturalist (Forbes et al., 2019), Relational Speaker (Tan et al., 2019), L2C (Yan et al., 2021) and F&P (Lv et al., 2025).

FDDA follows a distinct design philosophy from the above: rather than encoding differences via explicit subtraction or aggregation across spatial features, FDDA decouples multi-scale signals via a feature-domain frequency decomposition on the vision output, applies bidirectional image-pair cross-attention within each frequency band, and integrates back into the vision→LM path via gated fusion during SFT. To our knowledge, feature-domain frequency decomposition has not been applied to explicit image-pair difference modeling in prior IDC work.

2.3 Frequency-Domain Methods

Frequency-domain transforms have been widely applied in vision and sequence models — FNet (Lee-Thorp et al., 2022), AFNO (Guibas et al., 2022), GFNet (Rao et al., 2021), FFC (Chi et al., 2020) and FcaNet (Qin et al., 2021) aim at efficiency or global mixing on single-image features. Although GFNet and FcaNet already use frequency-domain filtering and DCT channel attention in the feature domain, both still operate on single-image features; to our knowledge, no prior work uses feature-domain frequency decomposition to explicitly model differences between image pairs. In remote-sensing change detection, multi-scale bi-temporal difference analysis has been widely explored at both the pixel level (e.g., wavelet/Fourier decomposition) and the feature level (e.g., transformer-based change detectors) (Liu et al., 2022); these approaches produce dense change maps rather than natural-language descriptions. FDDA differs from all of the above in (i) operating on learned

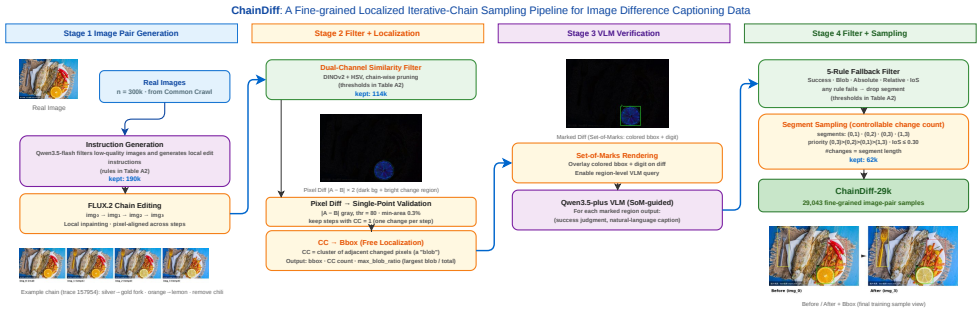

Figure 2: Overview of the ChainDiff pipeline. A four-stage flow from 300k source images to 29,043 fine-grained image pairs (retained scales shown in the figure); the detailed engineering steps are given in the appendix.

semantic features of the VLM during SFT post-training, (ii) targeting language output, and (iii) using dual-branch cross-attention to couple the two images at each frequency band.

3 Dataset

3.1 Pipeline Overview and Data Source

Our data pipeline starts from 300k raw images collected from Common Crawl, goes through instruction generation, chained editing and multi-level filtering, and ultimately produces 29,043 high-quality fine-grained image pairs; the overall structure and the retained scale of each stage are shown in Figure 2.

3.2 Instruction Generation and Chained Editing

VLM instruction generation and editing-type constraints. Each image is fed into Qwen3.5-Flash, which simultaneously produces four fields: a quality verdict, a short caption, an entity list and 1–3 editing instructions (field definitions and low-quality rejection types in the appendix); 190k of the 300k source images are retained. Editing instructions are restricted to a whitelist of sub-part-level operations valuable for fine-grained perception; global scene changes, spatial displacement, pose changes and text modification are explicitly forbidden (full whitelist and reject list in Appendix A.1).

Chained editing. We adopt FLUX.2 Edit (Black Forest Labs, 2025) and execute instructions in a chained synthesis manner: the input of the i-th edit is the output image of the $(i-1)$-th, not the original image. Each source image yields one editing chain of up to 4 images; adjacent images in the chain naturally form a difference pair with exactly one local change, and step-by-step quality verification is straightforward.

3.3 Fine-Grainedness and Controllable Change Count Guarantee

Dual-channel similarity filtering. A single metric cannot cover all editing types: DINOv2 (Oquab et al., 2024) semantic similarity captures changes in structure, shape and layout but is insensitive to pure color changes, which hue–saturation–value (HSV) color-histogram similarity complements. Consecutive chain steps must simultaneously pass the semantic-similarity and the weighted-similarity thresholds (Table A2); a failed step discards itself and all subsequent edits, retaining 114k records in total.

Single-point editing verification and fine-grained upper bound. For each pair of adjacent images, a full-resolution pixel difference is computed followed by connected-component labeling; only steps with exactly one connected component are kept, and any single changed region whose area exceeds 10% of the whole image is rejected (`max_blob_ratio` $\leq$ 0.10), enforcing sub-part-level fine-grainedness by definition (implementation and thresholds in the appendix).

Segment sampling and explicitly controllable change count. A 4-image chain yields 1/2/3-step candidate segments plus a fallback segment salvaging failed first edits; at most one segment is selected per source image by priority (Table A2). The change count of an output sample equals the segment length, so the change-count distribution can be set on demand.

3.4 Automatic Bounding Boxes and Captions

Automatic bbox derivation. The connected-component analysis above directly yields the bounding box of every changed region, requiring no manual annotation; the process is pure pixel computation and scales to the million level.

Set-of-Marks-guided caption generation. Numbered colored bboxes (Yang et al., 2023) are overlaid on the after image, which is fed with the before image into Qwen3.5-Plus, generating a difference description per changed location; for multi-change samples the descriptions are concatenated region by region, and numbers correspond one-to-one to bboxes, enforcing text–localization alignment. The same call also judges whether each edit was correctly executed (a `success` field), intercepting failed edits or erroneous modification of neighboring objects without additional API calls.

3.5 Fallback Filtering and Dataset Statistics

The segment-level fallback filtering covers five orthogonal dimensions — Success, intersection-over-smaller (IoS), Blob, absolute cleanliness and relative cleanliness; triggering any of them discards the whole segment. After filtering, the 62k segments yield a final set of 29,043 image pairs (27,344 train / 1,699 test), with a pyramid-like change-count distribution: single-change 57.6%, 2- / 3-change 34.7% / 7.7%. Full rule definitions and thresholds are in Appendix A.2 and Table A2.

Fine-grainedness verification. The pixel-level changed ratio diff_ratio statistics (Table A1) confirm that ChainDiff pairs concentrate at the sub-part level, well below OmniDiff and CLEVR-Change on the median and substantially smaller than Birds-to-Words and IEdit which represent whole-image or large-magnitude edits. Full commentary in Appendix A.3.

4 Methodology

This section introduces the Frequency-Domain Difference Attention (FDDA) module, a plug-in inserted between the vision encoder and the language model that models image-pair differences explicitly at the feature level; the overall structure is shown in Figure 3.

FDDA addresses two problems of the standard concatenation paradigm: difference signals are hidden inside overwhelming commonality, and multi-scale difference components (global color/layout shifts vs. local sub-part attributes) are entangled in the same patch features. FDDA introduces frequency-domain decomposition to decouple multi-scale signals and image-pair cross-attention to establish difference perception explicitly.

4.1 Feature-Domain Frequency Decomposition

Reshape the vision encoder's output on image I into a 2D patch-feature grid $F \in \mathbb{R}^{H\times W\times D}$; low-frequency components carry overall structure and global semantics, and high-frequency components carry discontinuities between adjacent patches such as semantic boundaries and local attribute changes. FDDA applies a per-channel 2D discrete cosine transform (DCT) to obtain $\hat{F} = \mathrm{DCT}_{2D}(F)$; DCT is chosen over the fast Fourier transform (FFT) for its real-valued orthogonality and empirical margin over FFT (Table A5).

Frequency banding. For each frequency location (u, v) we define a normalized radial index

$$r(u,v) = \frac{1}{\sqrt{2}}\sqrt{\left(\frac{u}{H-1}\right)^2 + \left(\frac{v}{W-1}\right)^2} \in [0,1], \tag{1}$$

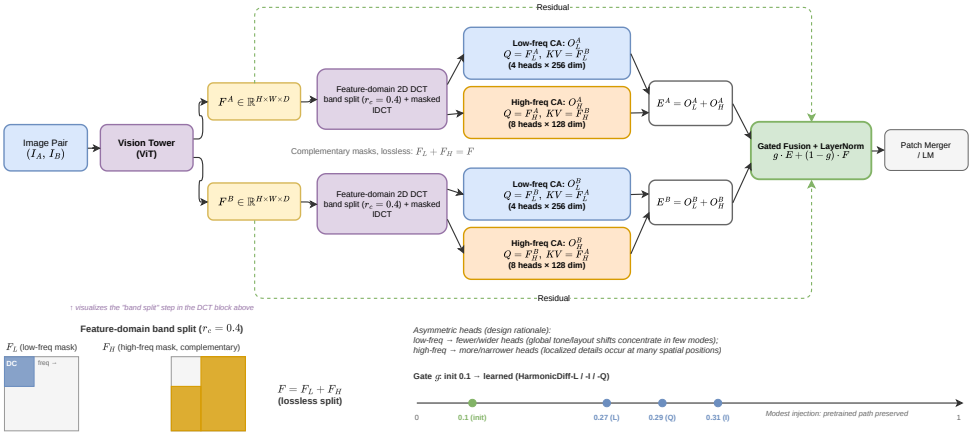

Figure 3: Structure of the FDDA module. The image-pair features are decomposed by a feature-domain 2D DCT into low/high-frequency branches, restored by masked IDCT, and each branch executes A↔B bidirectional cross-attention; the outputs are summed and mixed with the original features in a residual manner through a learnable gate, followed by LayerNorm. The decomposition is lossless.

and choose a cutoff $r_c = 0.4$ to obtain complementary low/high-frequency masks $M_L = \mathbf{1}[r \le r_c]$ and $M_H = \mathbf{1}[r > r_c]$. Masked inverse DCTs restore each band to the spatial domain:

$$F_L = \mathrm{IDCT}_{2D}(\hat{F} \odot M_L), \qquad F_H = \mathrm{IDCT}_{2D}(\hat{F} \odot M_H). \tag{2}$$

The decomposition is lossless ($F_L + F_H = F$), and returning to the spatial domain enables position-aligned image-pair cross-attention. FDDA acts on learned features rather than pixels, so "high frequency" here refers to sharp changes across adjacent patches (semantic boundaries and local attribute discontinuities), naturally aligning with the sub-part-level differences the pipeline targets.

4.2 Dual-Branch Image-Pair Cross-Attention and Gated Output

Let the decomposition of an image pair (I_A, I_B) yield four spatial-domain tensors: $F_L^A, F_L^B, F_H^A, F_H^B$. FDDA uses two structurally complementary branches to model differences at the two scales separately; within each branch, A↔B bidirectional cross-attention is executed:

$$O_X^A = \mathrm{CrossAttn}(Q{=}F_X^A,\ K{=}V{=}F_X^B), \qquad O_X^B = \mathrm{CrossAttn}(Q{=}F_X^B,\ K{=}V{=}F_X^A), \quad X \in \{L, H\} \tag{3}$$

Asymmetric head configuration. A key design choice of FDDA is to make the two branches structurally asymmetric so they match the statistics of the signal they carry: the low-frequency branch uses *fewer but wider* heads (4 heads × 256 dims), since low-frequency differences are typically global tone or layout shifts whose modes concentrate in a small number of high-dimensional subspaces; the high-frequency branch uses *more but narrower* heads (8 heads × 128 dims), since high-frequency differences are localized detail changes at many spatial positions and benefit from a larger number of attention modes. The total attention capacity is kept comparable across branches. Q/K/V projections map the D-dimensional features into the branch's inner-product space and another projection maps the output back, so the same FDDA inserts into any base without modification. Our internal ablation (Table A5) shows that either fully symmetric alternative (both branches 4×256 or both 8×128) underperforms the asymmetric configuration, confirming that the head-count / head-dim trade-off should be aligned with the frequency band.

The two branch outputs are summed per patch position ($E^A = O_L^A + O_H^A$, $E^B = O_L^B + O_H^B$) — summation is preferred over concatenation because the components are spatially aligned views of the same patch. A learnable scalar gate g then mixes the enhanced and original features, followed by LayerNorm: $\mathrm{FDDA}(F^A, F^B)^Z = \mathrm{Norm}(g \cdot E^Z + (1-g) \cdot F^Z)$ for $Z \in \{A, B\}$. g is initialized to 0.1 so early training stays close to the pretrained path; the converged g is 0.27

/ 0.31 / 0.29 on the LLaVA-OneVision-7B / InternVL3-8B / Qwen3.5-4B bases respectively, indicating a moderate residual injection. FDDA adds only ∼8.4M parameters and one DCT plus two masked IDCTs per image, negligible relative to the base VLM.

4.3 Integration with the Training Framework

FDDA mounts at runtime into the visual feature stream of the base VLM (vision encoder → FDDA → patch merger → language model, Figure 3); no modeling code of the base is modified and `from_pretrained` loading stays fully compatible. Verified on three structurally different bases — LLaVA-OneVision-7B (SigLIP + Qwen2-7B), InternVL3-8B (InternViT + Qwen2.5-7B), Qwen3.5-4B (native multimodal) — FDDA is insensitive to vision-encoder structure and tiling strategy.

5 Experiments

5.1 Experimental Setup

Base models. We experiment on three mainstream open-source VLMs — LLaVA-OneVision-7B, InternVL3-8B and Qwen3.5-4B — abbreviated LLaVA-OV / InternVL3 / Qwen3.5-4B in the tables; with FDDA plugged in they become HarmonicDiff-L / -I / -Q.

Training data. The training data composition follows M^{3}Diff (Liu et al., 2025) for fairness: six public IDC datasets (the Real and Render subsets of OmniDiff merged as one), totaling ∼145k image pairs, plus the 27,344 ChainDiff training pairs; the only difference from Liu et al. (2025) is that we do not apply its 2× OmniDiff upsampling. M^{3}Diff's base, LLaVA-OneVision-7B (SigLIP + Qwen2-7B) (Liu et al., 2025), is the same base model as our HarmonicDiff-L, so the two share an identical backbone and training recipe and are directly comparable in Tables 1–2. All public data is converted into a unified "image pair + difference description" format, each CLEVR-family caption forming an independent training sample as in Liu et al. (2025) (per-dataset composition in the appendix).

Experimental conditions. Three progressive conditions with identical bases and hyper-parameters isolate data and module contributions: **w/o ChainDiff & FDDA** trains only on the six public datasets; **w/o FDDA** adds the ChainDiff training set (data effect); **HarmonicDiff** further introduces FDDA (module effect).

Training hyper-parameters. Aligned with the fine-tuning recipe of M^{3}Diff (Liu et al., 2025): AdamW optimizer, cosine annealing with linear warm-up (warm-up ratio 0.03), global batch size 256, trained for 1 epoch; the vision encoder and the vision–language projection layer are fully fine-tuned (2e-6 / 1e-5), the language model uses LoRA (1e-5), and FDDA is trained with full parameters from scratch (5e-5).

Evaluation benchmarks. We evaluate on the official test sets of six public benchmarks plus our own ChainDiff test set (1,699 manually verified pairs), following each benchmark's official split and standard evaluation protocol throughout: CLEVR-Change (Park et al., 2019), CLEVR-DC (Kim et al., 2021), Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018), Image-Edit-Request (Tan et al., 2019), Birds-to-Words (Forbes et al., 2019), and OmniDiff (Liu et al., 2025) (Real / Render). Per-benchmark characteristics (e.g., distractor changes, single- vs multi-change protocols) are summarized in Table A11 in the appendix.

Evaluation metrics. Following the OmniDiff (Liu et al., 2025) protocol, we report BLEU-4 (B), METEOR (M), ROUGE-L (R), CIDEr-D (C) and SPICE (S), computed offline with the standard Microsoft COCO caption evaluation toolkit under greedy decoding; metrics not reported by a comparison method are marked —.

5.2 Main Results

We compare the three bases in three conditions against MLLM baselines (FINER-MLLM (Zhang et al., 2024), OneDiff (Hu et al., 2024), M^{3}Diff (Liu et al., 2025)); dedicated non-MLLM IDC models are included in the appendix tables where such results have been

reported. Bold and underline mark the best and second-best among public comparison methods and HarmonicDiff, with ablation rows excluded from the ranking; Table 3 follows a separate convention and compares only against the prior MLLM best (see its caption). We report the two OmniDiff subsets (Tables 1–2), the public-benchmark overview (Table 3), and our own test set in the appendix (Table A3).

OmniDiff-Real.

Table 1: Performance comparison on the OmniDiff-Real test set. No dedicated non-MLLM IDC baseline has been reported on this benchmark.

Backbone	Method	B	M	R	C	S
MLLM	M^{3}Diff (ICCV'25) (Liu et al., 2025)	14.3	<u>18.9</u>	32.9	31.3	—
LLaVA-OV	w/o ChainDiff & FDDA	13.2	18.1	32.3	28.9	21.8
	w/o FDDA	13.4	18.6	32.7	30.5	21.9
	HarmonicDiff-L	14.1	18.3	34.0	31.8	22.3
InternVL3	w/o ChainDiff & FDDA	13.6	17.4	33.1	34.4	22.1
	w/o FDDA	14.1	17.0	34.0	33.3	21.7
	HarmonicDiff-I	<u>14.7</u>	17.8	<u>36.1</u>	<u>34.6</u>	<u>22.6</u>
Qwen3.5-4B	w/o ChainDiff & FDDA	17.6	20.4	36.5	46.0	25.8
	w/o FDDA	17.4	20.5	36.6	46.3	25.7
	HarmonicDiff-Q	**18.2**	**21.3**	**37.9**	**46.9**	**26.2**

OmniDiff-Render.

Table 2: Performance comparison on the OmniDiff-Render test set. No dedicated non-MLLM IDC baseline has been reported on this benchmark.

Backbone	Method	B	M	R	C	S
MLLM	M^{3}Diff (ICCV'25) (Liu et al., 2025)	15.7	19.9	35.3	28.3	—
LLaVA-OV	w/o ChainDiff & FDDA	12.4	17.3	33.8	27.9	24.6
	w/o FDDA	13.2	17.6	33.7	27.8	25.3
	HarmonicDiff-L	14.8	19.3	34.8	29.4	**26.8**
InternVL3	w/o ChainDiff & FDDA	18.7	19.5	36.3	28.1	24.2
	w/o FDDA	19.0	19.4	36.0	30.1	24.6
	HarmonicDiff-I	<u>20.4</u>	<u>21.1</u>	<u>37.5</u>	<u>31.7</u>	25.9
Qwen3.5-4B	w/o ChainDiff & FDDA	20.4	20.5	36.4	34.9	26.4
	w/o FDDA	20.8	21.0	37.1	35.0	26.3
	HarmonicDiff-Q	**21.9**	**21.7**	**37.9**	**35.5**	<u>26.6</u>

Our own ChainDiff test set. As shown in Table A3 (appendix), public-data-only models score around 20 CIDEr on our own test set; injecting ChainDiff alone already lifts them into the 58–62 CIDEr band, with a further small FDDA gain on top. Because the test-set references were generated by the same captioner (Qwen3.5-Plus) used during training, we further re-evaluate against GPT-4-regenerated references (Table A4); HarmonicDiff CIDEr is largely preserved, indicating that the reported gain reflects genuine capability on the fine-grained distribution rather than reference-style memorization.

Overview of the public benchmarks. Table 3 summarizes five public benchmarks (other than the two OmniDiff subsets, reported in Tables 1–2) against a per-metric "Prior MLLM best" baseline computed over FINER-MLLM, OneDiff and M^{3}Diff. HarmonicDiff attains the best CIDEr among MLLM methods on both OmniDiff subsets and on CLEVR-DC, Spot-the-Diff, and Birds-to-Words; it remains competitive with the prior MLLM best on CLEVR-Change and IEdit. On CLEVR-Change specialized non-MLLM methods still lead on this synthetic benchmark (Appendix Table A6); on IEdit, whose large-amplitude edits lie outside ChainDiff's target distribution, HarmonicDiff edges past the prior MLLM best on BLEU-4, ROUGE-L and SPICE (taking the best of the three bases) but stays below it on METEOR and CIDEr.

5.3 Ablation Study

The three progressive conditions form a joint ablation of data and module (full rows in Tables 1–3 and Tables A3–A10, appendix); we summarize the trends here. *ChainDiff data* (w/o FDDA vs. w/o ChainDiff & FDDA) yields directionally positive gains on nearly all public benchmarks across the three bases, and delivers a large jump (two- to three-fold) on

Table 3: Five-metric overview on five public benchmarks (CLEVR-Change, CLEVR-DC, Spot-the-Diff, IEdit, and Birds-to-Words). "Prior MLLM best" is the best per-metric value *among existing MLLM methods* (FINER-MLLM, OneDiff, M^{3}Diff); specialized non-MLLM IDC methods (e.g., CLIP4IDC on CLEVR-Change, F&P on Birds-to-Words) are listed in Appendix Tables A6–A10. Best / second-best per benchmark × metric are marked bold / underline. SPICE is not reported by prior MLLM methods on Birds-to-Words; HarmonicDiff SPICE values are shown for reference.

Benchmark	Method	B	M	R	C	S
CLEVR-Change	Prior MLLM best	57.1	**41.9**	75.3	**137.2**	33.8
	HarmonicDiff-L	**57.3**	40.7	74.6	129.8	33.5
	HarmonicDiff-I	54.7	39.2	75.5	126.3	29.1
	HarmonicDiff-Q	56.5	40.7	**77.9**	135.5	**34.2**
CLEVR-DC	Prior MLLM best	60.6	37.6	73.0	109.4	21.3
	HarmonicDiff-L	61.5	37.6	74.0	110.9	21.1
	HarmonicDiff-I	62.7	37.7	73.1	114.8	**22.9**
	HarmonicDiff-Q	**63.2**	**38.4**	**74.9**	**116.6**	22.3
Spot-the-Diff	Prior MLLM best	14.4	15.4	37.6	71.1	24.9
	HarmonicDiff-L	**15.3**	**15.8**	**37.8**	**71.8**	**25.0**
	HarmonicDiff-I	13.6	15.1	36.1	65.9	23.3
	HarmonicDiff-Q	12.9	15.4	34.9	64.7	22.9
IEdit	Prior MLLM best	33.6	**26.5**	59.7	**136.6**	27.5
	HarmonicDiff-L	**33.8**	26.0	**60.0**	134.3	27.1
	HarmonicDiff-I	29.2	24.8	58.5	126.3	25.3
	HarmonicDiff-Q	25.1	25.6	58.3	134.7	**27.7**
Birds-to-Words	Prior MLLM best	27.1	25.8	49.6	28.4	—
	HarmonicDiff-L	31.0	26.0	50.8	27.1	17.9
	HarmonicDiff-I	30.2	26.7	**52.0**	**29.1**	**18.7**
	HarmonicDiff-Q	**31.7**	**27.1**	49.8	27.9	18.2

our own fine-grained test set. *FDDA module* (HarmonicDiff vs. w/o FDDA), applied on top of ChainDiff, yields further positive CIDEr gains on all three bases across all eight test sets, indicating that frequency-domain difference modeling is broadly effective on structurally different bases. The two contributions combine additively: full HarmonicDiff does not degrade CIDEr on any of the eight test sets relative to either ablation condition.

FDDA internal ablation. We further ablate the three internal design axes of FDDA — cutoff r_c, head configuration, and frequency transform — on the Qwen3.5-4B base (Appendix Table A5). All examined variants underperform the baseline configuration by only a small margin on our own test set, indicating that the baseline is a stable operating point rather than a knife-edge optimum; among the three axes, the head configuration and the DCT-vs-FFT choice have a comparatively larger effect than the cutoff.

6 Conclusion

We approach fine-grained image difference captioning from both the data and model sides: the ChainDiff pipeline synthesizes controllable fine-grained image pairs with fully automatic annotations, and the plug-and-play FDDA module models pair differences explicitly via feature-domain DCT decomposition and dual-branch cross-attention. Across three bases and eight test sets, HarmonicDiff attains the best CIDEr among MLLM methods on five test sets; on our own fine-grained test set, the large jump in CIDEr is largely preserved under a GPT-4-regenerated reference, suggesting the gain reflects real capability rather than same-captioner stylistic memorization.

Limitations. Edit types are bounded by FLUX.2 (no spatial displacement, pose or text edits); transfer to remote sensing / medical imaging is not evaluated here; and FDDA is pair-only by design, activating only when the input contains image pairs, so it does not benefit single-image tasks.

References

Black Forest Labs. The FLUX family of image generation and editing models (FLUX.1, 2024; FLUX.2, 2025). `https://bfl.ai`, 2025. No formal paper; cited as a technical release.

Lu Chi, Borui Jiang, and Yadong Mu. Fast Fourier convolution. In *Proceedings of NeurIPS*, pp. 4479–4488, 2020.

Maxwell Forbes, Patsy Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. In *Proceedings of EMNLP-IJCNLP*, pp. 708–717, 2019. arXiv:1909.04101. Proposes the Birds-to-Words dataset.

John Guibas, Mohammad Amin Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive Fourier neural operators: Efficient token mixers for transformers. In *Proceedings of ICLR*, 2022. arXiv:2111.13587.

Zixin Guo, Tzu-Jui Julius Wang, and Jorma Laaksonen. CLIP4IDC: CLIP for image difference captioning. In *Proceedings of AACL-IJCNLP (Short Papers)*, pp. 33–42, 2022. arXiv:2206.00629.

Mehdi Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *Proceedings of CVPR*, pp. 2725–2734, 2021. Reports the DUDA+TIRG results.

Erdong Hu, Longteng Guo, Tiancheng Yue, Zhuo Zhao, Jian Xue, and Jiaming Liu. OneDiff: A generalist model for image difference captioning. In *Proceedings of ACCV*, pp. 2439–2455, 2024. arXiv:2407.05645.

Qian Huang, Yuxin Liang, Jiawei Wei, Chen Yi, Hongxia Liang, Ho-Fung Leung, and Quan Li. Image difference captioning with instance-level fine-grained feature representation. *IEEE Transactions on Multimedia*, 24:2004–2017, 2022.

Zihan Huang, Junda Wu, Rohan Surana, Tong Yu, David Arbour, Ritwik Sinha, and Julian McAuley. Image difference captioning via adversarial preference optimization. In *Proceedings of EMNLP*, pp. 33758–33770, 2025.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of EMNLP*, 2018. arXiv:1808.10584.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Bolin Ding, Yaliang Li, and Ying Shen. Img-Diff: Contrastive data synthesis for multimodal large language models. In *Proceedings of CVPR*, pp. 9296–9307, 2025. arXiv:2408.04594.

Hyeonje Kim, Jinsung Kim, Heeseung Lee, Sungroh Park, and Gwanhyun Kim. Viewpoint-agnostic change captioning with cycle consistency. In *Proceedings of ICCV*, 2021.

James Lee-Thorp, Joshua Ainslie, Illia Eckstein, and Santiago Ontañón. FNet: Mixing tokens with Fourier transforms. In *Proceedings of NAACL*, 2022. arXiv:2105.03824.

Chenyang Liu, Renqiao Zhao, Hongping Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. The RSICC dataset and RSICCformer.

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 ICCV*, pp. 21440–21449, 2025. arXiv:2503.11093.

Fang Lv, Rui Wang, Liang Jing, and Lei Liu. Find and perceive: Tell visual change with fine-grained comparison. In *Proceedings of IJCAI*, 2025. doi: 10.24963/ijcai.2025/654. Proposes the F&P method.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. arXiv:2304.07193.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of ICCV*, pp. 4624–4633, 2019. arXiv:1901.02527. The CLEVR-Change dataset and the DUDA model.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *Proceedings of ICCV*, 2021. arXiv:2012.11879.

Yuji Qiu, Shota Yamamoto, Kodai Nakashima, Ryosuke Suzuki, Kazuhiro Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of ICCV*, pp. 1971–1980, 2021. Proposes the MCCFormers model.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Proceedings of NeurIPS*, 2021. arXiv:2107.00645.

Xu Shi, Xu Yang, Jiaxian Gu, Shafiq Joty, and Jianshu Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Proceedings of ECCV, Part XIV*, pp. 574–590, 2020. arXiv:2009.14352. Proposes the M-VAM model.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. In *Proceedings of ACL*, 2019. arXiv:1906.07689. Proposes the Image-Edit-Request dataset.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhuoran Yu. R3Net: Relation-embedded representation reconstruction network for change captioning. In *Proceedings of EMNLP*, 2021a. arXiv:2110.10328.

Yunbin Tu, Ting Yao, Liang Li, Jianxin Lou, Shengxiang Gao, Zhuoran Yu, and Chenggang Yan. Semantic relation-aware difference representation learning for change captioning. In *Findings of ACL-IJCNLP*, pp. 63–73, 2021b. Proposes the SRDRL model.

Yunbin Tu, Liang Li, Li Su, Jiali Du, Ke Lu, and Qingming Huang. Viewpoint-adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32:2620–2635, 2023a. Proposes the VARD model.

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, 25:9518–9529, 2023b. Proposes the NCT model.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of ICCV*, pp. 2805–2815, 2023c. Proposes the SCORER model.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Context-aware difference distilling for multi-change captioning. In *Proceedings of ACL*, 2024a. arXiv:2405.20810. Proposes the CARD model.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In *Proceedings of ECCV*, pp. 311–328, 2024b. Proposes the DIRL model.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. SMART: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4926–4943, 2024c. Journal extension of SRDRL.

An Yan, Xun Wang, Tsu-Jui Fu, and William Yang Wang. L2C: Describing visual differences needs semantic understanding of individuals. In *Proceedings of EACL*, 2021. arXiv:2102.01860.

Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

Lewen Yao, Wei Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *Proceedings of AAAI*, 2022.

Xudong Zhang, Hongwei Wen, Jianlong Wu, Peng Qin, Hao Xue, and Liqiang Nie. Differential-perceptive and retrieval-augmented MLLM for change captioning. In *Proceedings of ACM MM*, pp. 4148–4157, 2024. Proposes the FINER-MLLM model.

A Pipeline Parameters and Verification Details

A.1 Editing-Type Whitelist

Editing instructions produced by the Qwen3.5-Flash instruction generator (Section 3.2) are restricted to four sub-part-level operations valuable for fine-grained perception, and are rejected if they belong to any of the four forbidden categories:

- **Whitelisted operations.** (i) attribute modification (color / material / pattern of an existing object part); (ii) object removal; (iii) object addition; (iv) same-category replacement.
- **Rejected operations.** (i) spatial displacement of existing objects; (ii) pose or gesture change; (iii) text overlay / OCR-style modification; (iv) global scene changes such as lighting, weather, background or overall style.

The rejection list is enforced by the Qwen3.5-Flash prompt and further verified downstream by the single-point editing verification (Section 3.3), which discards any step producing more than one connected changed region or a region exceeding the `max_blob_ratio` upper bound (Table A2).

A.2 Fallback Filtering Rules

The segment-level fallback filtering combines five orthogonal criteria; triggering any of them discards the entire segment.

- **Success.** The per-region `success` field returned by the Set-of-Marks captioning call (Section 3.4) must be `True` for every changed region in the segment; failed or distorted edits are dropped.
- **IoS** (Intersection over Smaller). If any two bboxes in the segment satisfy IoS > 0.30, the two edits touch the same region and the segment is discarded (threshold in Table A2).
- **Blob.** No single connected component may exceed 10% of the image (`max_blob_ratio` ≤ 0.10), rejecting global redrawing and large-object addition/removal.
- **Absolute cleanliness.** fg $-$ blob ≤ 0.03, where fg is the total foreground (pixel-difference) area and blob the area covered by the union of segment bboxes; the residual outside the editing regions catches secondary unintended changes.
- **Relative cleanliness.** The ratio of foreground to bbox-covered blob area must be ≤ 1.5, filtering fragmented samples with scattered diffs.

A.3 Fine-Grainedness Verification (extended commentary)

Table A1 reports pixel-level diff_ratio quantiles for the seven datasets. ChainDiff's median (0.041) sits well below OmniDiff (0.140) and CLEVR-Change (0.089), and its p75 (0.070) stays below the operating threshold, so fine-grainedness is a distribution-level invariant rather than an average property, consistent with the hard upper bound `max_blob_ratio` ≤ 0.10 of Section 3.3. Spot-the-Diff's median (0.0170) is even lower because of how the dataset is collected: its pairs are adjacent surveillance-video frames, so changes are inherently subtle; a valid pair typically contains exactly one change, and a substantial portion of its adjacent-frame pairs contain almost no visible change at all. This also caps what automatic metrics can achieve on this benchmark and depresses all methods' scores there. Birds-to-Words (median 0.9578) and IEdit (median 0.3785) correspond respectively to whole-image comparisons and large-magnitude edits, both outside the fine-grained target distribution — included in Table A1 for reference.

Table A1: Statistics of the pixel-level changed ratio diff_ratio of ChainDiff and existing IDC datasets (fine-grainedness verification in Section 3.5).

Dataset	n	mean	p25	p50	p75
Spot-the-Diff	10,554	0.0182	0.0100	0.0170	0.0246
ChainDiff (ours)	**27,344**	**0.0503**	**0.0224**	**0.0411**	**0.0697**
CLEVR-Change	71,636	0.1015	0.0555	0.0885	0.1331
OmniDiff	14,027	0.2920	0.0272	0.1398	0.5124
CLEVR-DC	43,200	0.3147	0.2181	0.3025	0.4011
Image-Edit-Request	3,061	0.4223	0.0574	0.3785	0.7731
Birds-to-Words	2,953	0.9216	0.9081	0.9578	0.9829

Note: n counts the training-split samples (test set excluded), consistent with the official scales in Tables A11 and A12.

Table A2: Key parameters of the ChainDiff pipeline.

Stage	Parameter	Value	Role
Similarity filtering	Semantic similarity s_{sem}	> 0.70	Semantic structure does not collapse
Similarity filtering	Weighted similarity $0.5s_{\text{sem}} + 0.5s_{\text{color}}$	> 0.75	Constrains the overall editing magnitude
Single-point verification	Pixel difference threshold thr	80	Filters out JPEG compression noise
Single-point verification	Connected-component area threshold	0.3% of total pixels	Filters out fragments
Single-point verification	Number of connected components	$= 1$	Guarantees single-point editing
Fine-grained upper bound	`max_blob_ratio`	$\leq 10\%$ of the whole image	Object sub-part-level fine-grainedness
Segment sampling	Sampling priority (segment start/end indices in the 4-image chain)	$(0,3) > (0,2) > (0,1) > (1,3)$	3-step segments preferred; (1, 3) salvages first-step failures
Segment sampling	In-segment bbox overlap IoS	≤ 0.30	Avoids conflicting edits in the same region
Fallback filtering	fg − blob	≤ 0.03	Absolute cleanliness
Fallback filtering	fg / blob	≤ 1.5	Relative cleanliness

Here IoS (Intersection over Smaller) is the intersection area of the two boxes divided by the area of the smaller box.

Implementation and threshold calibration of single-point editing verification. Per adjacent chain pair, a full-resolution pixel difference is computed: per-channel absolute differences are multiplied by 2, converted to grayscale and binarized at threshold 80 (average per-channel difference ≥ 40) to filter JPEG noise; 8-connectivity labeling plus a 0.3%-area fragment filter keeps only steps with exactly one connected component. The threshold was calibrated on 500 records (993 steps): thr80/r0.003 retains 389 CC=1 samples with 1.9% noise; a larger area threshold drops retention to 322. All ∼260k editing steps are processed with this verification.

Supplementary pipeline details (Sections 3.2–3.5). Qwen3.5-Flash outputs four fields: `Quality` (ACCEPT/REJECT), `ShortCaption`, `Entity` and 1–3 `Edit` entries; low-quality images (pure text/document, UI screenshots, extremely blurry, no valid foreground, dense stacks of small objects) are rejected directly. Set-of-Marks generates a 12–30-word description per changed location. The five fallback rules: **Success** (failed/distorted edits), **IoS** (multiple edits in the same region), **Blob** (global redrawing and large-object addition/removal), **absolute cleanliness** (secondary changes outside the editing region), **relative cleanliness** (scattered diffs; thresholds in Table A2).

FDDA engineering implementation details (Section 4.3). Image pairs are arranged within a batch in interleaved order $[A_1, B_1, A_2, B_2, \dots]$ and FDDA splits with stride 2 (a trailing odd image degenerates to identity). Differing feature-grid sizes are zero-padded to the union grid, then cropped back for gated fusion; mismatched tile counts fall back to identity.

B Detailed Experimental Results

Tables A3–A10 give the detailed comparisons and analyses: Tables A3–A5 report analyses on our own test set (main results, cross-captioner validation, and the FDDA internal ablation), and Tables A6–A10 cover the five public benchmarks with the full list of dedicated (non-MLLM) IDC baselines, corresponding to the main-text tables. The annotation protocol is consistent with the main text (ablation rows are not annotated); bold and underline mark the best and second-best among public comparison methods and HarmonicDiff, with ablation rows excluded from the ranking (same convention as the main text).

Table A3: Performance comparison on the 1,699 pairs of our own ChainDiff test set. This test set has no public baseline, so no best/second-best annotation is made; the comparison of the three conditions simultaneously serves the ablation analysis in Section 5.3.

Backbone	Method	B	M	R	C	S
LLaVA-OV	w/o ChainDiff & FDDA	4.1	17.2	28.2	19.7	15.4
	w/o FDDA	19.1	33.1	35.1	58.1	34.3
	HarmonicDiff-L	20.5	33.8	36.6	59.3	35.7
InternVL3	w/o ChainDiff & FDDA	5.1	13.2	29.4	20.3	15.1
	w/o FDDA	18.3	33.5	37.8	61.7	35.7
	HarmonicDiff-I	19.1	33.6	38.1	62.9	37.0
Qwen3.5-4B	w/o ChainDiff & FDDA	5.7	19.1	30.4	22.1	16.9
	w/o FDDA	20.9	34.3	36.2	60.1	35.7
	HarmonicDiff-Q	21.7	35.8	37.7	62.3	37.2

Table A4: Cross-captioner validation on our own ChainDiff test set: the 1,699 test-set reference captions are regenerated by GPT-4 (instead of Qwen3.5-Plus used in training and in Table A3), and the three conditions are re-evaluated. Comparing against Table A3, HarmonicDiff CIDEr drops by only 0.3, 1.4, and 0.5 for L / I / Q respectively, indicating that the gains reflect genuine capability rather than reference-style memorization.

Backbone	Method	B	M	R	C	S
LLaVA-OV	w/o ChainDiff & FDDA	4.0	16.9	27.8	19.6	15.3
	w/o FDDA	18.8	33.0	34.7	57.8	34.2
	HarmonicDiff-L	20.3	33.7	36.4	59.0	35.5
InternVL3	w/o ChainDiff & FDDA	5.0	13.1	29.4	20.1	15.0
	w/o FDDA	18.2	33.4	37.7	61.6	35.2
	HarmonicDiff-I	18.9	33.4	37.8	61.5	35.6
Qwen3.5-4B	w/o ChainDiff & FDDA	5.5	18.9	30.2	21.9	16.5
	w/o FDDA	20.7	34.1	35.7	60.0	35.2
	HarmonicDiff-Q	21.6	35.5	37.4	61.8	37.1

Table A5: FDDA internal ablation on our own ChainDiff test set (Qwen3.5-4B base). The baseline configuration uses cutoff $r_c = 0.4$, asymmetric heads (low: 4 heads $\times$ 256 dim, high: 8 heads $\times$ 128 dim), DCT decomposition, and $g_{\text{init}} = 0.1$. Each variant modifies one axis while holding the others at their baseline values. All examined variants perform within 1.4 CIDEr of the baseline; the head configuration and the DCT-vs-FFT choice contribute most to the final performance.

Variant	B	M	R	C	S
Baseline (HarmonicDiff-Q)	**21.7**	**35.8**	**37.7**	**62.3**	**37.2**
cutoff $r_c = 0.2$	21.4	35.5	37.1	61.9	36.7
cutoff $r_c = 0.6$	21.6	35.7	37.4	62.2	36.8
symmetric-large (4×256 / 4×256)	21.1	35.4	37.2	61.1	36.5
symmetric-small (8×128 / 8×128)	21.0	34.7	36.2	60.9	36.1
FFT (instead of DCT)	20.8	35.1	35.8	61.7	36.2

Table A6: Performance comparison on the CLEVR-Change test set.

Backbone	Method	B	M	R	C	S
—	DUDA (ICCV'19) (Park et al., 2019)	47.3	33.9	—	112.3	24.5
—	M-VAM (ECCV'20) (Shi et al., 2020)	50.3	37.0	69.7	114.9	30.5
—	M-VAM+RAF (ECCV'20) (Shi et al., 2020)	51.3	37.8	70.4	115.8	30.7
—	IFDC (TMM'22) (Huang et al., 2022)	49.2	32.5	69.1	118.7	—
—	DUDA+TIRG (CVPR'21) (Hosseinzadeh & Wang, 2021)	51.2	37.7	70.5	115.4	31.1
—	VACC (ICCV'21) (Kim et al., 2021)	52.4	37.5	—	114.2	31.0
—	SRDRL (Findings'21) (Tu et al., 2021b)	54.9	40.2	73.3	122.2	32.9
—	R^{3}Net (EMNLP'21) (Tu et al., 2021a)	54.7	39.8	73.1	123.0	32.6
—	MCCFormers-D (ICCV'21) (Qiu et al., 2021)	52.4	38.3	—	121.6	26.8
—	CLIP4IDC (AACL'22) (Guo et al., 2022)	56.9	38.4	<u>76.4</u>	**150.7**	—
—	NCT (TMM'23) (Tu et al., 2023b)	55.1	40.2	73.8	124.1	32.9
—	VARD-Trans (TIP'23) (Tu et al., 2023a)	55.4	40.1	73.8	126.4	32.6
—	SCORER (ICCV'23) (Tu et al., 2023c)	56.3	<u>41.2</u>	74.5	126.8	33.3
—	SMART (TPAMI'24) (Tu et al., 2024c)	56.1	40.8	74.2	127.0	33.4
—	DIRL (ECCV'24) (Tu et al., 2024b)	54.6	38.1	71.9	123.6	31.8
MLLM	FINER-MLLM (MM'24) (Zhang et al., 2024)	55.6	36.6	72.5	<u>137.2</u>	26.4
MLLM	M^{3}Diff (ICCV'25) (Liu et al., 2025)	<u>57.1</u>	**41.9**	75.3	130.5	<u>33.8</u>
LLaVA-OV	w/o ChainDiff & FDDA	55.2	39.3	74.2	128.3	30.3
	w/o FDDA	56.6	39.4	74.1	128.4	31.7
	HarmonicDiff-L	**57.3**	40.7	74.6	129.8	33.5
InternVL3	w/o ChainDiff & FDDA	53.7	38.6	73.3	125.8	28.4
	w/o FDDA	53.2	38.7	74.1	125.9	28.5
	HarmonicDiff-I	54.7	39.2	75.5	126.3	29.1
Qwen3.5-4B	w/o ChainDiff & FDDA	55.8	38.4	75.3	132.2	32.5
	w/o FDDA	55.4	39.1	76.6	133.7	32.1
	HarmonicDiff-Q	56.5	40.7	**77.9**	135.5	**34.2**

Table A7: Performance comparison on the CLEVR-DC extreme-viewpoint-change test set.

Backbone	Method	B	M	R	C	S
—	DUDA (ICCV'19) (Park et al., 2019)	40.3	27.1	—	56.7	16.1
—	DUDA+CC (ICCV'21) (Kim et al., 2021)	41.7	27.5	—	62.0	16.4
—	M-VAM (ECCV'20) (Shi et al., 2020)	40.9	27.1	—	60.1	15.8
—	M-VAM+CC (ICCV'21) (Kim et al., 2021)	41.0	27.2	—	62.0	15.7
—	VA (ICCV'21) (Kim et al., 2021)	44.5	29.2	—	70.0	17.1
—	VACC (ICCV'21) (Kim et al., 2021)	45.0	29.3	—	71.7	17.6
—	MCCFormers-D (ICCV'21) (Qiu et al., 2021)	46.9	31.7	—	71.6	14.6
—	NCT (TMM'23) (Tu et al., 2023b)	47.5	32.5	65.1	76.9	15.6
—	VARD-Trans (TIP'23) (Tu et al., 2023a)	48.3	32.4	—	77.6	15.4
—	SCORER (ICCV'23) (Tu et al., 2023c)	49.4	33.4	66.1	83.7	16.2
—	DIRL (ECCV'24) (Tu et al., 2024b)	51.4	32.3	66.3	84.1	16.8
MLLM	M^{3}Diff (ICCV'25) (Liu et al., 2025)	60.6	37.6	73.0	109.4	21.3
LLaVA-OV	w/o ChainDiff & FDDA	60.9	37.4	73.6	108.5	20.7
	w/o FDDA	60.1	37.3	73.5	108.5	20.9
	HarmonicDiff-L	61.5	37.6	<u>74.0</u>	110.9	21.1
InternVL3	w/o ChainDiff & FDDA	61.2	36.8	73.9	113.3	22.1
	w/o FDDA	61.9	36.4	73.1	113.6	22.5
	HarmonicDiff-I	<u>62.7</u>	<u>37.7</u>	73.1	<u>114.8</u>	**22.9**
Qwen3.5-4B	w/o ChainDiff & FDDA	63.7	37.7	74.4	114.1	21.9
	w/o FDDA	63.9	37.1	75.6	114.4	22.1
	HarmonicDiff-Q	**63.2**	**38.4**	**74.9**	**116.6**	<u>22.3</u>

Table A8: Performance comparison on the Spot-the-Diff test set (single-change protocol; see Section 5.1).

Backbone	Method	B	M	R	C	S
—	DUDA (ICCV'19) (Park et al., 2019)	8.1	11.8	29.1	32.5	—
—	M-VAM (ECCV'20) (Shi et al., 2020)	10.1	12.4	31.3	38.1	—
—	M-VAM+RAF (ECCV'20) (Shi et al., 2020)	11.1	12.9	33.2	42.5	17.1
—	IFDC (TMM'22) (Huang et al., 2022)	8.7	11.7	30.2	37.0	—
—	DUDA+TIRG (CVPR'21) (Hosseinzadeh & Wang, 2021)	8.1	12.5	29.9	34.5	—
—	VACC (ICCV'21) (Kim et al., 2021)	9.7	12.6	32.1	41.5	—
—	SRDRL (Findings'21) (Tu et al., 2021b)	—	13.0	31.0	35.3	18.0
—	R^3Net (EMNLP'21) (Tu et al., 2021a)	—	13.1	32.6	36.6	18.8
—	MCCFormers-D (ICCV'21) (Qiu et al., 2021)	10.0	12.4	—	43.1	18.3
—	CLIP4IDC (AACL'22) (Guo et al., 2022)	11.6	14.2	35.0	47.4	—
—	VARD-Trans (TIP'23) (Tu et al., 2023a)	—	12.5	29.3	30.3	17.3
—	SCORER (ICCV'23) (Tu et al., 2023c)	10.2	12.2	—	38.9	18.4
—	SMART (TPAMI'24) (Tu et al., 2024c)	—	13.5	31.6	39.4	19.0
—	DIRL (ECCV'24) (Tu et al., 2024b)	10.3	13.8	32.8	40.9	19.9
MLLM	FINER-MLLM (MM'24) (Zhang et al., 2024)	12.9	14.7	35.5	61.8	22.1
MLLM	OneDiff (ACCV'24) (Hu et al., 2024)	12.8	14.6	35.8	56.6	—
MLLM	M^3Diff (ICCV'25) (Liu et al., 2025)	<u>14.4</u>	<u>15.4</u>	<u>37.6</u>	<u>71.1</u>	<u>24.9</u>
LLaVA-OV	w/o ChainDiff & FDDA	12.2	14.6	35.9	69.3	23.4
	w/o FDDA	14.7	14.7	36.9	71.6	24.7
	HarmonicDiff-L	**15.3**	**15.8**	**37.8**	**71.8**	**25.0**
InternVL3	w/o ChainDiff & FDDA	12.7	14.1	35.3	63.5	21.7
	w/o FDDA	12.4	14.0	35.2	63.1	21.8
	HarmonicDiff-I	13.6	15.1	36.1	65.9	23.3
Qwen3.5-4B	w/o ChainDiff & FDDA	11.7	14.7	34.5	62.8	21.1
	w/o FDDA	11.4	14.1	34.6	63.0	21.3
	HarmonicDiff-Q	12.9	<u>15.4</u>	34.9	64.7	22.9

Table A9: Performance comparison on the Image-Edit-Request (IEdit) test set.

Backbone	Method	B	M	R	C	S
—	DUDA (ICCV'19) (Park et al., 2019)	6.5	12.4	37.3	22.8	—
—	MCCFormers-D (ICCV'21) (Qiu et al., 2021)	8.3	14.3	39.2	30.2	—
—	CLIP4IDC (AACL'22) (Guo et al., 2022)	8.2	14.6	40.4	32.2	—
—	NCT (TMM'23) (Tu et al., 2023b)	8.1	15.0	38.8	34.2	12.7
—	VARD-Trans (TIP'23) (Tu et al., 2023a)	10.0	14.8	39.0	35.7	—
—	SCORER (ICCV'23) (Tu et al., 2023c)	10.0	15.0	39.6	33.4	—
—	SMART (TPAMI'24) (Tu et al., 2024c)	10.5	15.2	39.1	37.8	—
—	DIRL (ECCV'24) (Tu et al., 2024b)	10.9	15.0	41.0	34.1	—
MLLM	FINER-MLLM (MM'24) (Zhang et al., 2024)	14.1	15.9	40.4	53.5	15.9
MLLM	OneDiff (ACCV'24) (Hu et al., 2024)	29.6	25.1	55.6	109.6	—
MLLM	M^3Diff (ICCV'25) (Liu et al., 2025)	<u>33.6</u>	**26.5**	<u>59.7</u>	**136.6**	<u>27.5</u>
LLaVA-OV	w/o ChainDiff & FDDA	32.2	25.8	58.7	130.9	26.2
	w/o FDDA	33.0	26.7	59.7	132.4	26.9
	HarmonicDiff-L	**33.8**	<u>26.0</u>	**60.0**	134.3	27.1
InternVL3	w/o ChainDiff & FDDA	27.8	23.1	56.4	123.3	24.8
	w/o FDDA	28.8	24.9	58.7	126.1	24.9
	HarmonicDiff-I	29.2	24.8	58.5	126.3	25.3
Qwen3.5-4B	w/o ChainDiff & FDDA	25.5	23.8	56.2	126.9	25.4
	w/o FDDA	25.6	25.7	58.3	131.0	26.1
	HarmonicDiff-Q	25.1	25.6	58.3	<u>134.7</u>	**27.7**

Table A10: Performance comparison on the Birds-to-Words test set. This benchmark is an additional one included by this work beyond the OmniDiff (Liu et al., 2025) evaluation suite. † L2C+CUB additionally uses the CUB dataset as external supervision. SPICE is not reported by prior methods on this benchmark.

Backbone	Method	B	M	R	C	S
—	Neural Naturalist (EMNLP'19) (Forbes et al., 2019)	22.0	—	43.0	25.0	—
—	Relational Speaker (ACL'19) (Tan et al., 2019)	21.5	22.4	43.4	5.8	—
—	DUDA (ICCV'19) (Park et al., 2019)	23.9	21.9	44.3	4.6	—
—	L2C (EACL'21) (Yan et al., 2021)	31.3	—	45.3	15.1	—
—	L2C+CUB (EACL'21) (Yan et al., 2021)†	31.8	—	45.6	16.3	—
—	IDC-PCL (AAAI'22) (Yao et al., 2022)	31.0	23.4	49.1	25.3	—
—	F&P (IJCAI'25) (Lv et al., 2025)	**33.4**	25.7	**52.2**	28.1	—
MLLM	OneDiff (ACCV'24) (Hu et al., 2024)	27.1	25.8	49.6	28.4	—
LLaVA-OV	w/o ChainDiff & FDDA	30.6	26.5	48.2	25.4	17.2
	w/o FDDA	31.0	26.5	50.4	26.3	17.5
	HarmonicDiff-L	31.0	26.0	50.8	27.1	17.9
InternVL3	w/o ChainDiff & FDDA	29.9	25.9	49.9	27.4	18.1
	w/o FDDA	30.7	26.0	51.6	28.7	18.3
	HarmonicDiff-I	30.2	26.7	52.0	**29.1**	**18.7**
Qwen3.5-4B	w/o ChainDiff & FDDA	31.0	26.1	50.0	26.7	17.7
	w/o FDDA	31.5	26.9	50.3	27.1	17.9
	HarmonicDiff-Q	31.7	**27.1**	49.8	27.9	18.2

Manual quality assurance of our own test set. Each of the 1,699 test pairs was manually selected from pipeline candidates: actual edits and captions were cross-checked, keeping only samples whose changes truly exist and whose descriptions strictly align with the images. The references therefore contain no incorrect changes or hallucinations; the "same pipeline, same captioning model" caveat in Section 5.2 concerns only stylistic fit, not reference correctness.

Training setup. All training runs on a single node with 8×NVIDIA A100 GPUs; hyperparameters follow M^{3}Diff (Liu et al., 2025) (Section 5.1) plus LoRA rank 128, alpha 256, dropout 0.05. The vision encoder and the vision–language projection layer are fully fine-tuned (2e-6 / 1e-5), the language model uses LoRA (1e-5), and FDDA (if enabled) is trained with full parameters from scratch (5e-5); FDDA's gate g is initialized to 0.1; its converged values are 0.27/0.31/0.29 for HarmonicDiff-L / -I / -Q. Training precision is bf16 with DeepSpeed ZeRO-2. The three conditions are trained separately: w/o ChainDiff & FDDA (public data only) takes ∼14 hours, and introducing FDDA leaves the training time essentially unchanged — the 8.4M-parameter module's overhead is negligible. Evaluation uses greedy decoding with a maximum generation length of 1024.

Table A11: Evaluation benchmark overview. Unless noted, we follow each benchmark's official split and standard protocol.

Benchmark	Type	Scale	Split / Protocol	Notes
CLEVR-Change (Park et al., 2019)	3D-rendered	79,606 pairs	train/val/test = 67,660/3,976/7,970	Contains distractor changes besides real changes
CLEVR-DC (Kim et al., 2021)	3D-rendered	48,000 pairs	Official 85/5/10 split	Introduces extreme viewpoint changes
Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018)	Surveillance-video adjacent frames	13,192 pairs	Official 8:1:1; single-change protocol	Only the primary change is described
Image-Edit-Request (Tan et al., 2019)	Real edited pairs	3,939 pairs	train/val/test = 3,061/383/495	—
Birds-to-Words (Forbes et al., 2019)	Bird photo pairs	3,347 pairs	Official split; paragraph-level captions	Additional; whole-image-level change region (diff$_r$atio median 0.9578, Table A1)
OmniDiff (Liu et al., 2025)	Real + 3D-rendered	Real 8,609 / Render 6,989 pairs	Official split; multi-change protocol	Test sets Real 803 / Render 767
ChainDiff (ours)	Real internet images	29,043 pairs	train/test = 27,344/1,699	Test set manually curated and cross-checked; references contain no change errors or hallucinations (see quality assurance above)

Table A12: Training data composition. Kept identical to M^{3}Diff (Liu et al., 2025) for fairness; all public data is converted to a unified "image pair + difference description" instruction format, each CLEVR-family caption forming an independent training sample, as in Liu et al. (2025).

Dataset	Type	Training scale
Spot-the-Diff	Real scene	Official train split: 10,554 pairs
Image-Edit-Request	Real scene	Train set: 3,061 pairs
Birds-to-Words	Real scene	2,953 pairs
OmniDiff	Real + 3D-rendered	Real and Render merged & deduplicated train+val: 14,027 pairs
CLEVR-Change	3D-rendered	train+val: 71,636 pairs
CLEVR-DC	3D-rendered	train+val: 43,200 pairs
ChainDiff (ours)	Real scene	27,344 pairs

Total ∼145k public image pairs plus the 27,344 ChainDiff training pairs.

C Use of LLMs

Large language models (LLMs) were used in three limited capacities during this work: (i) assisting with literature search and the identification of relevant prior work; (ii) AI-assisted coding for training, evaluation, and data-pipeline scripts, with all algorithms, designs, and results conceived, reviewed, and verified by the authors; and (iii) minor polishing of the manuscript's English prose. LLMs were not used to generate research ideas, experimental results, or any scientific claims in this paper. The authors take full responsibility for the content of this work.