Edits in the Wild: A Multilingual Benchmark for Localizing On-Device Generative Image Edits on Consumer Platforms

Mohd. Farhan Israk Soumik[1] **Awal Ahmed Fime**[2]
Wanniarachchi Kankanamge Malithi Mithsara[1] **Arun Babu Aryal**[1]
Hoda Joudi Haghighi[1] **Bibek Koirala**[1] **Ahmed Imteaj**[2] **Abdur R. Shahid**[1]
[1]Southern Illinois University Carbondale [2]Florida Atlantic University
{mohdfarhanisrak.soumik, malithi.mithsara, arunbabu.aryal,
hoda.joudihaghighi, bibek.koirala}@siu.edu
{afime2025, aimteaj}@fau.edu shahid@cs.siu.edu

Abstract

Consumer devices now ship with generative image editing tools that can remove objects, move subjects, and create new visual content directly on phones, tablets, and laptops. Yet current benchmarks largely miss this setting: real on-device edits produced by black-box consumer generative AI editing tools, with variation across platforms, devices, edit operations, and post-edit compression. We introduce CONSUMERGENEDIT, a large-scale multilingual benchmark for language-guided localization of on-device generative image edits. The dataset contains 4,000 royalty-free source images and 48,000 manually edited images created on Apple, Google, and Samsung devices, covering object removal, object movement, and sketch-based object generation. To study robustness under realistic sharing conditions, we add 144,000 learned-compression variants at three quality levels. Each example includes platform, device, editing operation, compression metadata, bounding-box annotations, and edit descriptions in English, Bangla, and Persian/Farsi generated with five vision-language models (VLMs). We evaluate five separate benchmark VLMs on edit detection and edit-operation classification. By focusing on where and how consumer AI edits occur, rather than only whether an image is synthetic, CONSUMERGENEDIT provides a practical testbed for multilingual VLM grounding, image provenance, and robust interpretation of AI-edited media.

1 Introduction

Generative AI is rapidly becoming part of everyday visual communication, as consumer devices now allow users to create, modify, and reinterpret images through built-in AI editing tools (Bansal et al., 2024). Systems such as Apple Intelligence (Li et al., 2025a), Google Magic Editor (Alzayer et al., 2025), and Samsung Galaxy AI (Choi et al., 2024) enable users to remove objects, move subjects, and generate visual content directly from sketches on personal devices (Figure 1). While these capabilities expand creative expression and accessibility, they also make image provenance harder to verify and create new challenges for visual misinformation analysis and trustworthy interpretation of edited media (Li et al., 2025b).

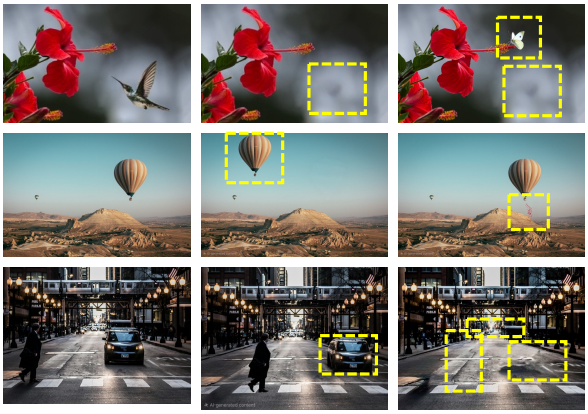


Figure 1: Examples of on-device generative image editing on iPhone 16 Pro and Samsung Galaxy S25 Ultra. Each row shows an original image (left) and edited outputs for object removal, object movement, and sketch-guided generation in standalone and cross-device settings. The figure illustrates the benchmark setup for VLM-based edit detection, edited-region localization, and multilingual captioning. Bounding boxes are illustrative and not model outputs.

This brings new challenges for vision-language models (VLMs) and multimodal AI systems. A useful model must do more than decide whether an image is real or synthetic: it must localize edits made by on-device generative tools, identify what content changed, and infer the editing context, including the platform, device family, operating system, and edit operation. It must also remain reliable across languages and post-edit compression settings. Prior work has studied synthetic image detection (Baraldi et al., 2025), instruction-based image editing and editing benchmarks (Meng et al., 2022; Hertz et al., 2023; Kawar et al., 2023; Brooks et al., 2023; Lee et al., 2023; Yang et al., 2025), and

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 37756–37789
October 24-29, 2026 ©2026 Association for Computational Linguistics

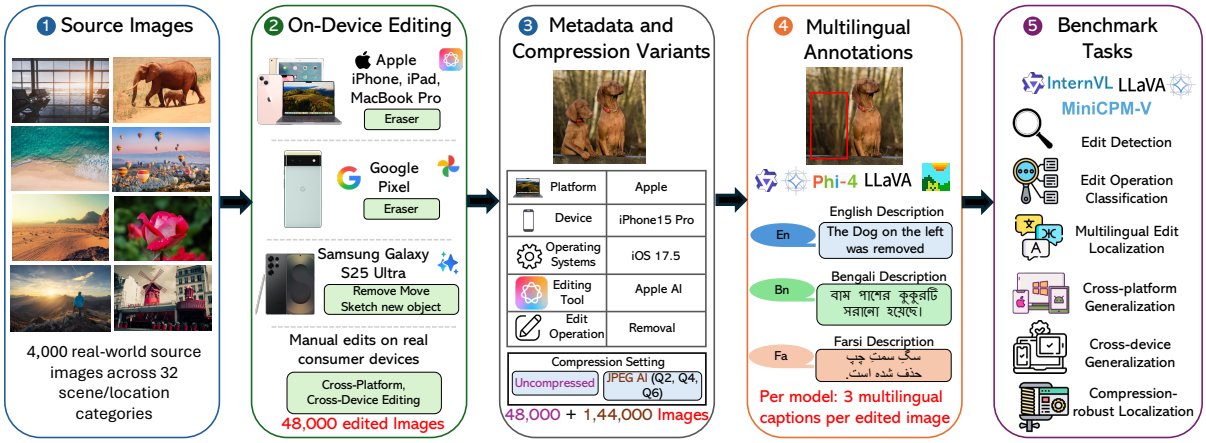


Figure 2: **ConsumerGenEdit** benchmark overview, covering source images, on-device consumer AI editing, platform and device metadata, Learned Image Compression variants, multilingual captions generated by VLMs, bounding-box annotations, and downstream VLM-based benchmark tasks. Q6, Q4, and Q2 correspond to approximately 0.83, 0.41, and 0.19 bpp and 35.7, 32.90, and 30 dB PSNR, respectively.

manipulation detection or localization (Ma et al., 2024). However, these settings do not fully capture consumer on-device editing, where edits are produced by black-box tools embedded in specific device ecosystems. Existing benchmarks also provide limited support for multilingual edit understanding (Zhang et al., 2023; Sheynin et al., 2024; Hui et al., 2025; Zhao et al., 2024; Schmidt et al., 2025) and rarely evaluate robustness after post-edit compression, despite compression being common in real-world image sharing. These gaps motivate a real-world, multilingual, and compression-aware benchmark for localizing and attributing generative image edits on consumer devices.

To address these gaps, we introduce CONSUMERGENEDIT, a multilingual benchmark for language-guided localization of on-device generative image edits.Moving beyond the question of edit detection, this benchmark evaluates whether vision-language models can identify what changed, where it changed, and how the edit was performed across consumer platforms, including under real-world compression and in the user's native language. The dataset contains 4,000 royalty-free source images and 48,000 manually edited images produced with consumer AI editing tools on Apple, Google, and Samsung devices. It includes edits from iPhone, iPad, MacBook Pro, Pixel, and Samsung Galaxy S25 Ultra, covering object removal, object movement, and sketch-based object generation. We further add 144,000 learned-compression variants generated with a neural image codec(Cheng et al., 2020) at three quality levels to support compression-robust evaluation. For each image, we provide platform and device metadata, edit-operation labels, compression metadata, bounding-box annotations, and edit descriptions in English, Bangla, and Persian/Farsi.

Our Contributions. We make the following contributions as depicted in Figure 2: ❶ We introduce CONSUMERGENEDIT, a multilingual benchmark of manually edited images produced with on-device consumer AI editing tools across Apple, Google, and Samsung devices. ❷ The dataset contains images edited with different operations on different consumer devices and operating systems, including object removal, object movement, and sketch-based object generation, enabling cross-platform, cross-device, and cross-edit evaluation. ❸ We provide Learned Compression(technological precursor to JPEG AI (Esenlik et al., 2026)) variants at multiple quality levels to study how post-edit compression affects edit detection, localization, and VLM robustness. ❹ We provide English, Bangla, and Persian/Farsi edit descriptions, generated by five highly capable VLMs to support language-guided edit localization across high-resource and lower-resource language settings. ❺ We evaluate five lightweight VLMs across zero-shot, few-shot, negative-example, reasoning-based, and region-aware prompting settings to test whether multilingual captions distilled from larger VLMs improve device-specific edit detection.

2 Related Work

Instruction-Based Generative Image Editing. Diffusion-based editing methods provide the technical basis for many consumer AI editing tools. SDEdit (Meng et al., 2022) introduced diffusion priors for guided synthesis from user inputs, while

Prompt-to-Prompt (Hertz et al., 2023) used cross-attention manipulation for text-only edits without spatial masks. Imagic (Kawar et al., 2023) enabled complex non-rigid edits through text-embedding optimization and diffusion fine-tuning, and InstructPix2Pix (Brooks et al., 2023) framed editing as supervised instruction following with synthetic image-instruction-edit triples. Consumer AI editing tools likely build on this line of work, but their implementations remain opaque, motivating output-based characterization of their edits.

Image Editing Benchmarks and Datasets. Several benchmarks evaluate instruction-based image editing. MagicBrush (Zhang et al., 2023) provides 10K manually annotated source-instruction-target triples across single-turn, multi-turn, mask-based, and mask-free edits. Emu Edit (Sheynin et al., 2024) covers object addition/removal, background alteration, style transfer, and color or texture changes. HQ-Edit (Hui et al., 2025) scales to roughly 200K edits using GPT-4V and DALL-E 3, while UltraEdit (Zhao et al., 2024) expands to approximately 4M samples through LLM-generated instructions, human curation, and real-image anchoring. Existing benchmarks largely rely on research models or automated pipelines, missing the platform-specific artifacts, color handling, edit semantics, and compression effects introduced by consumer tools such as Apple Intelligence, Samsung Galaxy AI, and Google's Magic Editor.

Vision-Language Models for Image Understanding. LLaVA (Liu et al., 2023) combines a CLIP-based visual encoder with an instruction-tuned language model, establishing a general-purpose visual assistant paradigm. Qwen-VL models (Wang et al., 2024; Bai et al., 2025b) extend this direction with dynamic-resolution processing, structured localization, and multi-image support. Pixtral 12B (Agrawal et al., 2024) and InternVL (Chen et al., 2024) further improve multimodal capability through native multimodal training and scalable visual foundation models. However, VLM performance varies across languages and cultural contexts. CulturalVQA (Nayak et al., 2024) reveals regional and cultural disparities in frontier VLMs, motivating multilingual evaluation in English, Bangla, and Persian/Farsi.

Image Manipulation and Forgery Detection. A related line of work studies manipulation detection and localization. ManTra-Net (Wu et al., 2019) learns generic manipulation traces through self-supervised training over diverse manipulation types. MVSS-Net (Dong et al., 2022) combines edge-supervised and noise-sensitive branches for semantic-agnostic localization. HiFi-Net (Guo et al., 2023) uses hierarchical fine-grained features to detect CNN-synthesized and edited images, while TruFor (Guillaro et al., 2023) fuses RGB and learned noise-fingerprint features with a transformer to produce localization maps and integrity scores. Robustness to recompression during online sharing has also been studied directly, with detectors designed to remain effective after social-network transmission (Wu et al., 2022).These methods often lack paired originals and focus on traditional manipulations, whereas consumer diffusion-based edits can blend into surrounding content and obscure forensic cues. Paired original and edited data is therefore essential for reference-based analysis and for exposing failures missed in unpaired settings.

AI-Generated Image and Edit Detection. Another line of work detects fully AI-generated images. Wang et al. (2020) showed that classifiers trained on one GAN can generalize across generators with appropriate augmentation. Ojha et al. (2023) found that frozen CLIP features with a linear probe improve cross-generator generalization, while DIRE (Wang et al., 2023) uses diffusion reconstruction error as a discriminative signal. GenImage (Zhu et al., 2023) provides a million-scale benchmark across eight generators. Closer to local editing, TGIF (Mareen et al., 2024) contains approximately 75K text-guided inpainting forgeries, while COCO-Inpaint (Yan et al., 2025), DiffSeg30k (Ci et al., 2025), and DEAL-300K (Zhang et al., 2025) extend diffusion-edit detection at different scales and supervision levels. Multimodal misinformation work also studies image manipulation in context: Akhtar et al. (2023) survey manipulation, out-of-context, and veracity classification; 5Pils (Tonglet et al., 2024) frames the task as image contextualization for fact-checking; MiRAGeNews (Huang et al., 2024) provides real and AI-generated news image-caption pairs; and Zeng et al. (2024) show that synthetic data and multimodal LLMs can support real-world fact-checking. Despite this progress, consumer-grade edits on real photographs remain underrepresented. A platform-grouped paired image benchmark with compression variants and multilingual descriptions addresses an increasingly common yet vastly understudied form of visual manipulation.

Dimension	Coverage in CONSUMERGENEDIT
❶ Source Images and Scale	
Image sources and coverage	4,000 royalty-free images from Pexels and Open Images Dataset V7, covering 32 visual categories
Dataset scale	48,000 on-device AI-edited images and 144,000 Learned Compressed edited images
Caption scale	15 captions per edited image, generated by 5 VLMs across 3 languages
❷ Consumer Editing Coverage	
Platforms and devices	Apple, Samsung, and Google devices: iPhone, iPad, MacBook Pro, Galaxy S25 Ultra, and Pixel Pro
Editing tools and settings	Apple Intelligence, Samsung Galaxy AI, and Google Magic Editor, with standalone and cross-device editing endpoints
Edit operations	Object removal, object movement, and sketch-based object generation
Original-edited pairs	48,000 uncompressed original-edited image pairs
❸ Compression and Annotation	
Compression codec and levels	Learned image compression (Cheng et al., 2020); Q6: light compression, 0.83 BPP / 35.7 dB PSNR; Q4: moderate compression, 0.41 BPP / 32.9 dB PSNR; Q2: heavy compression, 0.19 BPP / 30.0 dB PSNR
Compressed data	144,000 compressed pair instances and 144,000 compressed edited images
Languages and annotations	English, Bangla, and Persian/Farsi captions with edit descriptions, edit-operation labels, metadata, and edited-region bounding boxes
❹ Models and Benchmark Tasks	
Captioning VLMs and outputs	LLaVA-OV 7B, Pixtral 12B, Phi-4-RV 15B, Gemma 3 27B, and Qwen3-VL 32B; each produces English, Bangla, and Persian/Farsi captions per edited image
Detection VLMs and settings	Qwen2.5-VL, InternVL-2B, Phi-3.5-V, Gemma 3 4B, and MiniCPM-V 4B, evaluated under zero-shot and few-shot prompting
Benchmark tasks	Edit detection, edit-operation classification, multilingual localization, cross-platform and cross-device generalization, and compression-robust localization

Table 1: Overview of CONSUMERGENEDIT. The dataset includes source images, standalone and cross-device edits from consumer AI editing tools, Learned Compression variants, multilingual edit captions and benchmark tasks for edit detection and localization.

3 Dataset Construction

CONSUMERGENEDIT evaluates consumer generative image editing across devices, compression settings, and languages (Table 1). We construct the dataset through six steps: (1) source image collection, (2) standalone consumer editing, (3) cross-device editing, (4) original-edited pair construction, (5) JPEG AI compressed image generation, and (6) multilingual edit caption generation in English, Bangla, and Persian/Farsi. All edits are performed manually through native interfaces on real devices using Apple Intelligence, Samsung Galaxy AI, and Google Magic Editor, so the dataset reflects practical on-device editing workflows rather than API-based or fully automated editing pipelines. The image editing process was conducted between September 2025 and March 2026.

Source Images. The source contains 4,000 original photographs, equally collected from *Pexels.com* (Pexels, 2026) and *Open Images Dataset V7* (Kuznetsova et al., 2020; Google, 2026). Images cover 32 consumer photography categories, including natural scenes, urban environments, indoor spaces, roads, beaches, mountains, flowers, museums, landmarks, and travel locations. Each source image is assigned a unique identifier linking it to edited, compressed, and captioned variants for systematic evaluation.

Standalone Consumer Editing. Each source image is edited manually using consumer AI editing tools on real devices, forming editing endpoints across three consumer ecosystems and allow analysis across platforms, devices, and edit operations. Apple Intelligence is used on iPhone 16 Pro-Max, iPad Pro 11, and MacBook for object removal. Samsung Galaxy AI is used on a Galaxy S25 Ultra for object removal, object movement, and sketch-guided object generation. Google Magic Editor is used on a Pixel Pro device for object removal.

Cross-Device Editing. In addition to standalone edits, CONSUMERGENEDIT includes cross-device editing workflows. In this setting, an image edited on one device or platform is subsequently edited again using another device or platform. This reflects realistic sharing behavior, where images may move across ecosystems before publication or redistribution. The dataset includes five cross-device endpoints, covering combinations of Apple, Samsung, and Google editing tools. Together with the seven standalone endpoints, the complete editing layer contains 12 endpoints.

Pair Construction. Each edited image is paired with its corresponding original image using the preserved source identifier. This produces 48,000 original and edited image pairs which supports direct comparison between the original image and its edited counterpart for edit localization and caption generation.

Learned Image Compression (JPEG AI). Consumer images are often recompressed during upload, messaging, cloud synchronization, and sharing, which can weaken visual traces of generative editing. To model this realistic degradation setting, we compress every edited image using Cheng2020-Anchor (Cheng et al., 2020), a representative learned image codec implemented in CompressAI (Bégaint et al., 2020). Learned codecs are increasingly relevant to modern media pipelines, with neural image compression recently standardized as JPEG AI (ISO/IEC 6048-

1:2025) (Esenlik et al., 2026). We use three CompressAI quality levels following Kodak benchmark reporting: Q6 for light compression (≈0.83 BPP, 35.7 dB PSNR), Q4 for moderate compression (≈0.41 BPP, 32.9 dB PSNR), and Q2 for heavy compression (≈0.19 BPP, 30.0 dB PSNR). This produces 144,000 compressed edited images and corresponding compressed pair instances.

Edit-Localization Caption Generation. After constructing image pairs, we generate edit-localization captions with five VLMs: LLaVA-OV 7B, Pixtral 12B, Phi-4-RV 15B, Gemma 3 27B, and Qwen3-VL 32B (Li et al., 2024; Agrawal et al., 2024; Aneja et al., 2026; Team et al., 2025; Bai et al., 2025a). This is an edit-aware captioning task, not generic image captioning. Given an original image and its edited counterpart, each model is prompted in a zero-shot setting (Figures 7 and 8) to describe what changed and where it changed. The output links the edited region to the corresponding edit operation. We run the same procedure on uncompressed pairs and on compressed pairs at Q6, Q4, and Q2, allowing us to study how compression affects VLM edit understanding, including specificity loss, region mislocalization, and edit-operation errors.

Multilingual Captioning. We generate captions in English, Bangla, and Persian/Farsi. Bangla and Persian/Farsi are included because they use non-Latin scripts, have large speaker populations, and remain underrepresented in many vision-language benchmarks. For each language, the same five VLMs caption the same original-edited pairs under the same zero-shot paired-image prompt. The prompt specifies the target language and asks the model to compare the two images, identify the changed region, and describe the edit in that language. This keeps the image pairs, compression settings, and task definition fixed across languages, enabling direct comparison of VLM behavior across English, Bangla, and Persian/Farsi.

Benchmark Scope. Overall, CONSUMERGENEDIT supports evaluation along four axes: edit localization in consumer AI-edited images, robustness under JPEG AI compression, multilingual edit understanding, and cross-device editing behavior. By combining real-device edits, cross-device workflows, bitrate-controlled compression, and multilingual captions, the dataset provides a compact benchmark for analyzing how VLMs understand consumer generative edits under realistic visual and linguistic conditions.

4 Experiment Setup

All experiments were run on a server with one NVIDIA A100 GPU with 80GB VRAM, one NVIDIA A6000 GPU with 48GB VRAM, and an Intel Xeon CPU. The full pipeline required approximately 500 GPU-hours.

Captioning Vision-Language Models (VLMs). We evaluate five recent open-weight VLMs on the edit-description tasks: Qwen3-VL 32B, Gemma 3 27B, Phi-4-RV 15B, Pixtral 12B, and LLaVA-OneVision 7B. Each model receives the prompt template in Figure 7 (Appendix A.1.1) and generates a JSON object containing a free-text `caption`, a primary `edit_type` from a 13-token vocabulary, and a confidence value. We generate captions in English, Bangla, and Farsi. Decoding uses temperature 0.1, with retries at 0.3 and 0.5 when needed.

Benchmark VLMs. We evaluate five smaller VLMs under zero-shot and few-shot settings (Appendix A.1.3): Qwen2.5-VL, InternVL-2B, Phi 3.5 V, Gemma 3 4B, and MiniCPM-V 4B. The tasks cover binary edit detection, primary edit-type classification, and edited-region localization. In the zero-shot setting, models receive only the test image and task query. In the few-shot settings, the prompt includes demonstration examples before the test query, with separate configurations for positive edit pairs and positive-plus-negative examples.

Evaluation Metrics. Captioning quality is evaluated against uncompressed-input captions using `BERTScore F1` (Zhang et al., 2020) and `CIDEr` (Vedantam et al., 2015) for textual fidelity, `CLIPScore` drift (Hessel et al., 2021) for visual grounding, and length ratio for verbosity shift. CLIPScore drift is computed as the difference between uncompressed- and compressed-input CLIPScores, with values near zero indicating preserved grounding. We use language-specific backbones: English uses `RoBERTa-large` (Liu et al., 2019) and `ViT-L/14 CLIP` (Radford et al., 2021), while Bangla and Farsi use `XLM-RoBERTa-large` (Conneau et al., 2020) and multilingual CLIP (Carlsson et al., 2022). Cross-lingual agreement measures Bangla/Farsi consistency with English on the primary edit type, and cross-model agreement is measured using Fleiss' κ (Fleiss, 1971). Caption faithfulness is assessed in English using NLI-based contradiction probability, entailment asymmetry, and strict hallucination rate (He et al., 2021). Benchmark tasks report balanced accuracy and positive-prediction rate for binary edit detection.

Model	Pairwise agreement with model (%)					Per-model summary		
	Qwen3	Gemma	Phi-4	Pixtral	LLaVA	Mean	vs. Maj.	LOO-κ
Qwen3-VL 32B	100.0	55.0	58.2	65.6	60.6	59.9	69.2	0.201
Gemma 3 27B	55.0	100.0	44.0	54.1	50.4	50.9	56.4	0.271
Phi-4-RV 15B	58.2	44.0	100.0	58.6	57.3	54.5	60.1	0.239
Pixtral 12B	65.6	54.1	58.6	100.0	61.3	59.9	69.0	0.202
LLaVA-OV 7B	60.6	50.4	57.3	61.3	100.0	57.4	63.4	0.261

Table 2: Cross-VLM agreement on the primary `edit_type` label under the lightest compression tier. The left block reports pairwise agreement, while the right block summarizes model-level agreement and leave-one-out Fleiss' κ.

Model	L	Validity		Degeneracy		
		Val	Deg	Rep	D-2	gzip
Qwen3-VL	bn	98.7	1.3	0.19	0.95	2.56
	fa	99.6	0.4	0.28	0.94	1.90
Gemma-3	bn	100.0	0.0	0.10	0.99	2.07
	fa	99.9	0.1	0.16	0.98	1.59
Phi-4-RV	bn	97.0	3.0	0.20	0.93	2.32
	fa	98.4	1.6	0.25	0.94	1.61
Pixtral	bn	97.6	2.4	0.09	0.95	1.62
	fa	96.5	3.5	0.25	0.92	1.58
LLaVA-OV	bn	64.1	36.0	0.37	0.67	5.10
	fa	53.3	46.7	0.44	0.63	3.31

Table 3: Caption-quality diagnostics (uncompressed): validity (Val/Deg) and degeneracy signals (repeated-token fraction, distinct-2, gzip compressibility) (Welleck et al., 2020) (Holtzman et al., 2020)

5 Results and Evaluation

The results are organized into two parts. We first analyze the captioning layer of CONSUMERGENEDIT, focusing on cross-VLM agreement, multilingual caption quality, compression robustness, platform-level effects, and NLI-based caption errors. We then evaluate smaller benchmark VLMs on edit detection under reference-free, paired-image, zero-shot, and few-shot settings.

5.1 VLM Captioning Evaluation

Cross VLM Agreement. We measure cross-model agreement by comparing the primary `edit_type` assigned by five VLMs to the same image pair. Full five-model agreement is limited, while majority consensus is observed for most pairs. This suggests that consumer edits often lead to a dominant interpretation across models, but strict unanimity remains difficult. Table 2 shows that overall inter-VLM agreement is fair rather than strong. The highest pairwise agreement is observed between Qwen3-VL 32B and Pixtral 12B, followed by Pixtral 12B with LLaVA-OV 7B, and Qwen3-VL 32B with LLaVA-OV 7B. The weakest alignment occurs between Gemma 3 27B and Phi-4-RV 15B. At the model level, Qwen3-VL 32B and Pixtral 12B are the most consistent with the panel, while Gemma 3 27B has the lowest mean agreement and weakest agreement with the majority label. The leave-one-out κ values further suggest that Gemma 3 contributes the most disagreement relative to the other models. Overall, consumer edit categorization is partially consistent across VLMs but remains model-dependent, which motivates reporting cross-VLM agreement rather than relying on a single VLM as the only annotator.

Multilingual Caption Quality on Uncompressed Edited Images. Table 3 reports caption-quality diagnostics on uncompressed edited images. Most models produce valid and diverse Bangla and Persian/Farsi captions, with degeneration rates below 4.3%. LLaVA-OV 7B is the main exception, showing substantially lower valid-output rates and degeneration on roughly one-third to one-half of captions. The same pattern holds across Q2, Q4, and Q6 (Fig. 3), suggesting a model-intrinsic failure rather than a compression-induced effect

Captioning Robustness Under Learned JPEG AI Image Compression. Figure 3 shows that caption fidelity declines as compression increases across English, Bangla, and Farsi. Because BERTScore F1 and CIDEr decrease together, this pattern reflects genuine caption drift rather than metric-specific noise. The model ranking is also stable across languages: Pixtral 12B preserves the highest surface fidelity across compression levels, while Qwen3-VL 32B shows the lowest overlap scores. At the same time, visual grounding is less affected than surface form. CLIPScore drift remains small and approaches zero at lighter compression levels, suggesting that neural recompression changes caption wording more than image-text alignment. Caption length is also mostly stable, except for LLaVA-OneVision 7B, which produces longer captions under heavier compression, especially in Bangla and Farsi. Cross-lingual agreement shows a complementary pattern: agreement with English captions rises from roughly 27%–28% for LLaVA-OneVision 7B to 71%–74% for Qwen3-VL 32B, with Farsi consistently matching or exceeding Bangla. Thus,larger models show greater cross-lingual consistency despite more paraphrastic captions and lower BERTScore/CIDEr.

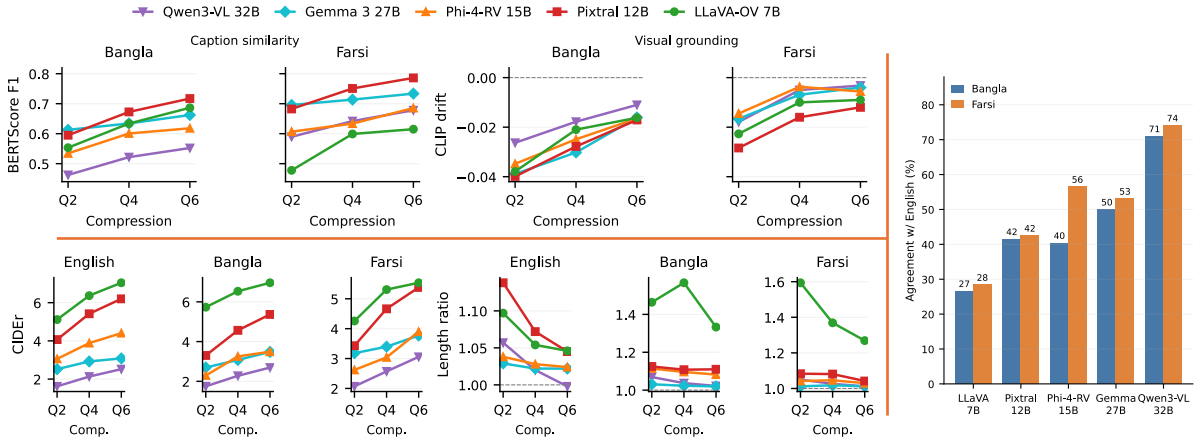


Figure 3: Surface-form effects across English, Bangla, and Farsi. Top-Left: Caption Similarity using BERTScore Drift. Top-Right: Visual Grounding with ClIPscore drift. Bottom-Left: CIDEr (n-gram overlap, compressed vs. uncompressed caption); Bottom-Right: caption length ratio (compressed / uncompressed); The barplot describes caption agreement of studied VLMs in Multilingual setting.

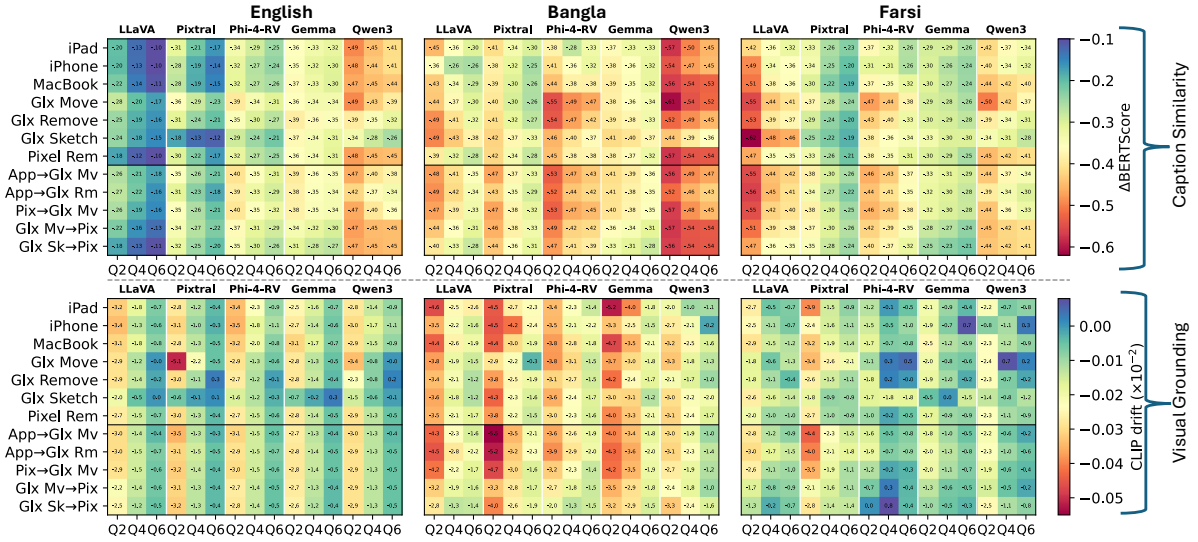


Figure 4: Per-platform BERTScore and CLIPScore drift ($\times 10^{-2}$) by language, averaged across the five models. (Top) BERTScore Drift (Bottom) Clipscore drift

Platform-Level Compression Effects. Figure 4 adds the platform-level view of compression effects. The heatmap shows that compression effects are primarily quality-driven and largely device-invariant. Both ΔBERTScore and CLIP drift move toward zero from Q2 to Q6 across editing endpoints, while platform-specific differences remain small. Neither the on-device editing tool nor the cross-device transfer pipeline substantially changes compression sensitivity. The largest degradation occurs for Qwen3-VL in Bangla and Farsi at Q2, consistent with the per-model ordering.

NLI-Based caption Error Analysis. To characterize how captions describing and localizing edits fail, an NLI-based error analysis has been performed which can be observed in Fig. 5. This error analysis is restricted to English only, owing to NLI coverage constraints for Bangla and Persian. Fig. 5 further shows that heavier compression increases contradiction probability, with Phi-4-RV showing the largest spread. Entailment asymmetry remains near zero for most models, while strict hallucination stays low and does not increase systematically. Overall, compression mainly causes omission and semantic drift rather than fabricated content.

5.2 Caption Correctness Against Human References

The preceding analysis establishes that the five captioning VLMs agree with each other only partially (Table 2) and that caption surface form drifts under compression (Figure 3). Modest inter-model agreement alone, however, does not reveal whether any individual model is correct or not. To separate correctness from consistency, model captions

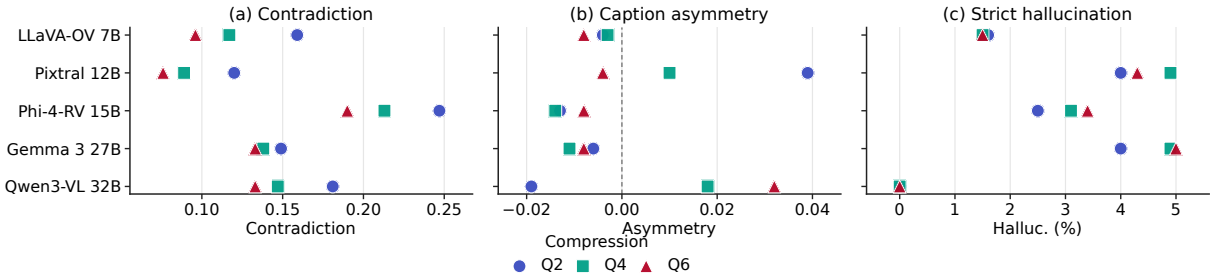


Figure 5: NLI-based error analysis across compression levels (English). (a) contradiction probability; (b) length-adjusted caption asymmetry (architecture-dependent crossover); (c) strict hallucination

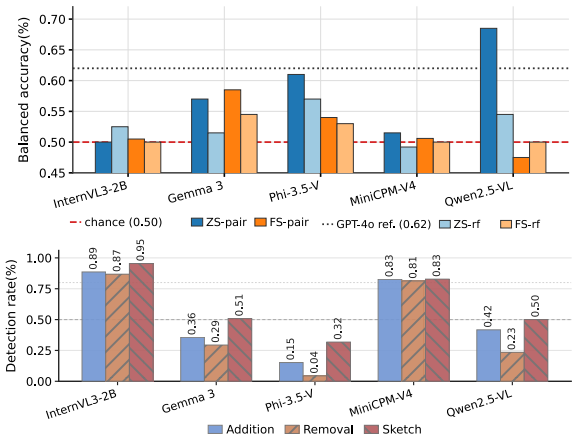


Figure 6: Performance of lightweight VLMs on edit detection on ConsumerGenEdit Dataset. (Top) Binary Detection Accuracy (Bottom) Detection rate by Edit type.

Model	Correctness vs. human ref. (%) Obj.	Op.	Full	Summary % Ceil.
English	*ceiling 39.6, majority baseline 62.2*			
Gemma 3 27B	62.8	55.8	35.8	90.4
Qwen3-VL 32B	77.9	42.3	32.7	82.5
LLaVA-OV 7B	55.6	49.9	28.0	70.7
Pixtral 12B	44.2	57.3	27.8	70.2
Phi-4-RV 15B	67.1	34.6	23.3	58.8
Bangla	*ceiling 37.2, majority baseline 61.3*			
Qwen3-VL 32B	85.5	46.4	39.3	105.6
Gemma 3 27B	70.4	20.8	14.7	39.5
Phi-4-RV 15B	49.2	26.6	13.5	36.3
Pixtral 12B	40.7	15.5	7.1	19.1
LLaVA-OV 7B	13.1	4.2	0.4	1.1
Farsi	*ceiling 36.0, majority baseline 47.6*			
Qwen3-VL 32B	81.5	43.1	35.9	99.7
Gemma 3 27B	66.9	30.0	21.8	60.6
Phi-4-RV 15B	42.7	35.3	17.5	48.6
Pixtral 12B	33.9	30.2	12.5	34.7
LLaVA-OV 7B	13.1	7.9	1.6	4.4

Table 4: Caption correctness against human references (%). *Obj.*: model names the object the annotator identified; *Op.*: same edit type; *Full*: both; *% Ceil.*: fully correct rate relative to the human agreement ceiling. Each language band reports its ceiling and the majority-class baseline (operation accuracy of always predicting the most common edit type).

are evaluated against independently written human references.

Protocol. A representative human evaluation was conducted with four independent annotators over 1,500 original-edited image pairs, comprising 500 pairs each in English, Bangla, and Farsi. The pairs were sampled across all platforms, edit types, and compression tiers. For each pair, the edited object and corresponding editing operation were independently identified by the annotators. Their consensus was used as the perceptual reference for evaluating model captions according to three criteria: ***object correctness***, which required the model to identify the object identified by the annotators using synonym-aware matching; ***operation correctness***, which required the corresponding edit type to be identified, with generative fill following object removal counted as removal; and ***full correctness***, which required both the object and operation to be correctly identified. Agreement among annotators, evaluated using the same criteria, was used to estimate the perceptual ceiling of the task. Text-similarity metrics such as BERTScore and CIDEr were not used for model ranking, as they primarily reflected caption length rather than edit fidelity on this dataset, favoring concise captions (e.g., Pixtral 12B) despite lower object recall while penalizing more descriptive captions with stronger object identification (e.g., Qwen3-VL 32B).

Table 4 shows that, across all three languages, no model exceeds the majority-class baseline for operation accuracy. In other words, none of the models identifies the editing operation more reliably than a constant prediction. The fully correct rate reaches at most 35.8% in English, 39.3% in Bangla, and 35.9% in Farsi. Qwen3-VL 32B and Phi-4-RV 15B abstain by predicting "No significant change" on 15-37% of the pairs, while Gemma 3 27B and Pixtral 12B rarely abstain. Qwen3-VL achieves the highest object identification accuracy in all three languages, although it does not provide an answer for roughly one third of the pairs. Gemma 3 27B has the highest aggregate correctness in English,

Model	Fully correct (%)		Significance	
	Uncomp.	Comp.	Δ (pp)	p-value
Gemma 3 27B	35.79	32.11	−3.68	0.04
Qwen3-VL 32B	32.72	29.65	−3.07	0.19
LLaVA-OV 7B	28.02	25.36	−2.66	0.07
Pixtral 12B	27.81	25.56	−2.25	0.08
Phi-4-RV 15B	23.31	24.74	+1.43	0.46

Table 5: Compression effect on caption correctness (English). Δ: change in fully correct rate; p: McNemar exact test (McNemar, 1947).

largely because it consistently produces an answer. In Bangla and Farsi, however, Qwen3-VL performs substantially better than the other models and approaches the human agreement level, while LLaVA-OV 7B performs poorly in both languages,consistent with the pattern observed in Table 3.

Compression. Scoring the same image pairs against the same human references under learned image compression isolates the effect of compression from image difficulty (Table 5). Fully correct rates decrease for four of five models by 2-4% points, with the largest drop at Q2 and little change at Q6. The decrease is significant for Gemma 3 27B (−3.68pp, McNemar p=0.04). For Qwen3-VL, abstention increases from 34.8% to 41.1%, while object identification remains similar (77.9% to 78.7%). Thus, the loss in full correctness is mainly associated with more abstentions rather than incorrect object identification. This result is consistent with the NLI-based analysis in section 5.1, which also indicates that compression primarily leads to caption omission.

5.3 Benchmarking VLMs for Edit Detection

Caption-Free and Caption-With Detection. Binary edit detection is evaluated across five lightweight VLMs under paired and reference-free settings, with performance shaped by both model capacity and access to the original image (Figure 6, top). In the reference-free setting, model performance remains close to the 50% random baseline, indicating limited ability to detect edits from the edited image alone. Providing the original image as a reference is therefore the strongest factor improving detection, although this benefit is concentrated among higher-capacity models. In the zero-shot paired setting, Qwen2.5-VL and Phi-3.5-V show comparatively stronger performance level while Gemma 3 performs modestly beyond the random baseline, whereas InternVL-2B and MiniCPM-V4 remain baseline-level across settings.

Few-Shot Prompting Effects. Few-shot with reference edit description demonstrations do not provide a consistent corrective, it improves Gemma 3 only marginally, but reduce performance for Qwen2.5-VL and Phi-3.5-V, suggesting that additional exemplar context can introduce confusion rather than improve discrimination.

Prediction Bias in Low-Performing Models. The edit-type breakdown first reveals a clear prediction bias (Figure 6, bottom). InternVL-2B and MiniCPM-V4 produce uniformly high detection rates (0.81–0.98) with little variation across edit types. This reflects an always-positive bias rather than reliable edit discrimination, consistent with their near-random performance.

Operation-Level Detection Difficulty. Among the models that show meaningful discrimination, namely Gemma 3, Phi-3.5-V, and Qwen2.5-VL, edit operations differ in difficulty. Sketch-based edits are the most detectable, removal edits are the least detectable, and addition edits fall between the two. This ordering reflects the visual evidence left by each operation: sketch edits introduce conspicuous synthetic strokes, whereas removal edits often require the model to infer an absence rather than recognize an added artifact. Overall, binary edit detection remains challenging for lightweight open-weight VLMs, motivating fine-tuning and supporting captioning as a more tractable evaluation axis.

6 Conclusion

In this paper, we introduce a large-scale benchmark for evaluating how VLMs interpret images edited with consumer on-device generative AI tools across detection, localization, and multilingual captioning tasks. We study object removal, object movement, and sketch-guided object generation across major device categories and ecosystems. The benchmark includes learned compressed images to assess robustness under realistic degradation conditions such as social media sharing and cross-platform transmission. It also provides VLM-generated edit descriptions in English, Bangla, and Farsi, covering both resource-rich and resource-constrained language settings. For benchmarking, we evaluate VLMs under zero-shot and few-shot settings across edit detection, edit-type classification, localization, and captioning-based analysis. Our findings show a clear gap between the rapid adoption of consumer generative editing tools and the robustness of current VLM-based auditing pipelines. By combining edit-level labels, localization, caption robustness, cross-language agreement, and compression sensitivity, the benchmark enables evaluation of VLMs beyond English-only and uncompressed inputs.

7 Limitations

This work has several limitations. First, our language coverage is limited. We include English as a resource-rich language and Bangla and Farsi as resource-constrained languages, which provides a useful setting for cross-language evaluation. However, these three languages do not capture the full diversity of languages, scripts, visual description practices, and cultural contexts that may affect multilingual image editing and captioning.

Second, our model coverage is necessarily limited. We evaluate five VLMs for caption generation and five VLMs for benchmark tasks. Although this allows us to compare multiple model families and sizes, the findings may not generalize to all current or future vision-language models, especially closed-source systems whose preprocessing, safety filters, and model updates are not publicly documented.

Third, our image sources do not fully cover real-world visual diversity. In particular, we do not explicitly evaluate culturally diverse image contexts or face-centered editing scenarios. This limits our ability to draw conclusions about how the evaluated systems behave on images involving culturally specific objects, scenes, identities, or sensitive human attributes.

Fourth, our compression analysis uses algorithmically applied compression levels rather than compression produced by real-world applications. While this controlled setup helps isolate the effect of compression severity, it does not capture the full behavior of social media platforms, messaging applications, cloud storage services, or repeated upload-download workflows that may resize, recompress, or otherwise transform images.

Fifth, our cross-device editing pipelines represent only a subset of possible user workflows. Real users may edit images across multiple devices, applications, file formats, and sharing platforms, sometimes with several rounds of saving, uploading, screenshotting, and recompression. These more complex workflows may introduce additional artifacts or failures not captured in our benchmark.

Sixth, our NLI-based error analysis is limited to English because the NLI tool used in our analysis does not support Bangla and Farsi. Therefore, we could not apply the same error categorization to these languages, and the reported NLI analysis should be interpreted as an English-only diagnostic.

Seventh, the captioning analysis is currently limited to a completed subset of the benchmark: 2,000 original source images, 24,000 edited images, and 72,000 learned compression edited images. Thus, captioning results should be interpreted as subset-level evidence rather than as complete coverage of the full benchmark.

Finally, the results are platform- and time-dependent. Both consumer AI editing tools and VLMs evolve rapidly through model updates, interface changes, and backend preprocessing changes. Therefore, our findings should be interpreted as a snapshot of the evaluated systems at the time of data collection rather than as permanent claims about these platforms.

8 Ethical Considerations

Image sources and licensing. All source images are collected from royalty-free or openly licensed sources, including Pexels and Open Images Dataset V7. We use these sources because they provide redistribution-friendly images and help avoiding copyright risk in dataset construction.

Dataset release. Because the edited images are produced using proprietary consumer AI editing tools, the release of edited or compressed images will follow the terms of the corresponding platforms and source-image licenses. When direct redistribution is restricted, we will release derived resources, including image identifiers, platform and device metadata, edit-operation labels, compression settings, captions, bounding-box annotations, evaluation splits, and reproduction scripts where permitted.

Sensitive content and representation. At this stage, the dataset is not designed to study identity, culture, race, gender, or facial attributes. We avoid making demographic or cultural claims from the images and do not use the benchmark for facial recognition, identity inference, or demographic classification. This limits some fairness and cultural-bias analysis, but it also reduces the risk of unsupported claims about people depicted in the images.

Annotation reliability. The dataset fields come from three sources. Edit operations, platform, and device metadata are recorded during the human editing process. Edited-region hints are generated using pixel differences and are provided only as auxiliary localization signals, not as ground truth. Edit captions and `edit_type` labels are generated

by VLMs and may include localization errors, inaccurate descriptions, or language-specific inconsistencies. To address this limitation, a 1,500-pair subset includes independently authored human reference captions §5.2, which provide verified references for evaluation and an estimate of the perceptual correctness ceiling.

Dual-use risk. The dataset concerns AI-edited images and therefore has potential dual-use implications. While the intended use is to support edit detection, localization, provenance analysis, and robustness evaluation, similar resources could be misused to study how edits evade detection. We frame the benchmark around detection, localization, and responsible interpretation of consumer AI edits rather than improving deceptive editing.

Use of generative AI in writing. During paper preparation, we used generative AI tools for limited writing assistance, including grammar correction, short-form phrasing support, polishing, and formatting. All technical claims, experimental design decisions, results, and interpretations were reviewed and finalized by the authors.

9 Acknowledgment

This material is partly based upon work supported by the U.S. NSF Grant CRII-IIS-RI-2553868, NSF NAIRR support (NAIRR260238). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. 2025. Magic fixup: Streamlining photo editing by watching dynamic videos. *ACM Transactions on Graphics*, 44(5):1–25.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. 2026. Phi-4-reasoning-vision-15b technical report. *arXiv preprint arXiv:2603.03975*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Gaurang Bansal, Aditya Nawal, Vinay Chamola, and Norbert Herencsar. 2024. Revolutionizing visuals: The role of generative ai in modern image generation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11):1–22.

Lorenzo Baraldi, Davide Bucciarelli, Federico Betti, Marcella Cornia, Nicu Sebe, and Rita Cucchiara. 2025. What changed? detecting and evaluating instruction-guided image edits with multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16217–16226.

Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. 2020. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 6848–6854.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198. Oral presentation.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7939–7948.

Pyeongjun Choi, Jeongsoo Kim, and Jeongho Kwak. 2024. Impact of joint heat and memory constraints of mobile device in edge-assisted on-device artificial intelligence. In *Proceedings of the 2nd International Workshop on Networked AI Systems*, pages 31–36.

Hai Ci, Ziheng Peng, Pei Yang, Yingxin Xuan, and Mike Zheng Shou. 2025. DiffSeg30k: A multi-turn diffusion editing benchmark for localized AIGC detection. *arXiv preprint arXiv:2511.19111*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. 2022. MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Semih Esenlik, Yaojun Wu, Zhaobin Zhang, Ye-Kui Wang, Kai Zhang, Li Zhang, João Ascenso, and Shan Liu. 2026. An overview of the jpeg ai learning-based image coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):2520–2537.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Google. 2026. Open images dataset v7 and extensions. `https://storage.googleapis.com/openimages/web/index.html`. Accessed: 2026-05-21.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20606–20615.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3155–3165.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRAGeNews: Multimodal realistic AI-generated news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16436–16448, Miami, Florida, USA. Association for Computational Linguistics.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2025. HQ-Edit: A high-quality and high-coverage dataset for general image editing. In *International Conference on Learning Representations (ICLR)*.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, and 1 others. 2023. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Ethan Li, Anders Boesen Lindbo Larsen, Chen Zhang, Xiyou Zhou, Jun Qin, Dian Ang Yap, Narendran Raghavan, Xuankai Chang, Margit Bowler, Eray Yildiz, and 1 others. 2025a. Apple intelligence foundation language models: Tech report 2025. *arXiv preprint arXiv:2507.13575*.

Fanxiao Li, Jiaying Wu, Canyuan He, and Wei Zhou. 2025b. CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9342–9354, Vienna, Austria. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*. Oral presentation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, and 1 others. 2024. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37:134591–134613.

Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallendael, Peter Lambert, and Symeon Papadopoulos. 2024. TGIF: Text-guided inpainting forgery dataset. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489.

Pexels. 2026. Free stock photo and video license. `https://www.pexels.com/license/`. Accessed: 2026-05-21.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

Fabian David Schmidt, Florian Schneider, Chris Biemann, and Goran Glavaš. 2025. MVL-SIB: A massively multilingual vision-language benchmark for cross-modal topical matching. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16285–16312, Vienna, Austria. Association for Computational Linguistics.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2024. Emu Edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8871–8879.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. "Image, tell me your story!" predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7845–7864, Miami, Florida, USA. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *In-*

ternational Conference on Learning Representations (ICLR).

Haiwei Wu, Jiantao Zhou, Jinyu Tian, Jun Liu, and Yu Qiao. 2022. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 17:443–456.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9543–9552.

Haozhen Yan, Yan Hong, Jiahui Zhan, Yikun Ji, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. 2025. COCO-Inpaint: A benchmark for image inpainting detection and manipulation localization. *arXiv preprint arXiv:2504.18361*.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, and Christopher Clark. 2025. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17486–17505, Vienna, Austria. Association for Computational Linguistics.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. 2024. Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10467–10484, Miami, Florida, USA. Association for Computational Linguistics.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Rui Zhang, Hongxia Wang, Hangqing Liu, Yang Zhou, and Qiang Zeng. 2025. DEAL-300K: Diffusion-based editing area localization with a 300K-scale dataset and frequency-prompted baseline. *arXiv preprint arXiv:2511.23377*.

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

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. UltraEdit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. GenImage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

A Appendix

A.1 Prompt Templates

This appendix lists the exact prompts used to elicit edit descriptions from the eight evaluated vision-language models (Section 4). All models receive the same template, populated at runtime with four variables: the editing platform, an edit-type prior derived from that platform, the pixel-diff region summary, and the target output language. We use temperature 0.1 for the first attempt with two retries at 0.3 and 0.5; `max_new_tokens=512`; for Ollama deployments `num_ctx=8192`.

A.1.1 Edit Description Prompt

Figure 7 shows the prompt template used to generate edit descriptions. Curly-brace tokens (e.g. `{platform_info}`) are runtime fillers; their possible values are listed in Table 6. The trailing `/no_think` token is a Qwen3-family soft switch that suppresses chain-of-thought emission; other models (Qwen2.5-VL, MiniCPM-V, Pixtral, InternVL, Gemma, Phi-4-RV, LLaVA) ignore it harmlessly.

A.1.2 Reduced-Schema Variant for LLaVA-OneVision

LLaVA-OneVision 7B is the smallest model in our evaluation suite. In preliminary testing it produced inconsistent outputs when given the full three-field schema in Figure 7: roughly one in five responses either omitted the `confidence` field, emitted it as free text rather than one of the three allowed tokens, or broke the JSON structure entirely. To keep the comparison meaningful rather than dominated by parse failures, we use a reduced two-field schema (`caption` and `edit_type`, without `confidence`) for this model alone, shown in Figure 8. The edit-type vocabulary is also reduced from thirteen to twelve tokens by dropping the residual `other` label, which LLaVA-OneVision overused as a default.

All seven other models (Qwen3-VL 32B, Pixtral 12B, Gemma 3 27B, MiniCPM-V 4.5, Phi-4-RV 15B) use the main prompt in Figure 7. When reporting cross-model comparisons in Section 4, we evaluate the `caption` and `edit_type` fields, which are common to both prompts; the `confidence` field is reported only for the seven models that produce it.

Table 6: Runtime fillers in the prompt template (Figure 7).

Filler	Possible values
`{platform_info}`	*Apple:* "The edit was performed using: Apple Intelligence (Clean Up / Rewrite tools on iOS/macOS)." *Galaxy:* "The edit was performed using: Samsung Galaxy AI (Generative Edit / Object Eraser)." *Pixel:* "The edit was performed using: Google Pixel (Magic Eraser / Magic Editor / Best Take)."
`{tool_priors}`	Soft hints inferred from the platform name. For Clean Up / Eraser tools: "Note: this tool typically performs object_removal or generative_fill." For Magic Editor / Generative Edit: a wider set including `object_replacement`, `background_change`, `sky_replacement`. For Best Take: `face_edit`.
`{bbox_info}`	One of three forms, derived from OpenCV pixel differencing: (i) "Pixel analysis found N edited region(s) in this $W \times H$ image:" followed by per-region lines specifying vertical and horizontal position (top/middle/bottom × left/center/right), area as a percentage of the image, and pixel coordinates; (ii) for global edits: "This appears to be a GLOBAL edit — $p\%$ of pixels changed."; (iii) for null diffs: "Pixel analysis found no clear edited region."
`{lang_label}`	`English`, or `Bangla` (বাংলা), or `Farsi` (فارسی) — the target output language for the `caption` field. Native-script glosses are included to anchor the model in the requested script.

A.1.3 Binary Edit-Detection Prompt (Ablation Study)

In addition to the open-ended description prompt above, we run a second-track binary classification ablation in which a VLM is asked to label a single image as either `Edited` or `Original`. This ablation spans five models (Qwen2.5-VL, Gemma 3 4b, Intern VL-2B, MiniCPM-V4 4B, Phi-3.5V) and three shot-stack configurations (zero-shot, few-shot with positive pair shots only, and few-shot with positive pair shots plus negative shots). To isolate the effect of the shot stack, all five models receive the *same* neutral system prompt and the same query text; the only variable across the three modes is the demonstration stack that precedes the query.

Figure 9 shows the four text components used to construct the conversation: a system prompt that defines the task, a query text appended to the test image, and two demonstration texts that label few-

shot examples as edited pairs or as clean negatives.

Answer parsing. VLM outputs to this prompt are converted to one of {`Edited`, `Original`, `None`} by a tolerant parser. The parser first tries the cleaned leading token against two vocabularies (`edited`, `manipulated`, `altered`, `modified`, `fake` for `Edited`; `original`, `real`, `authentic`, `unedited`, `genuine`, `unmodified`, `untouched` for `Original`). If neither matches, it scans the first sentence for positive markers while excluding negation patterns ("not edited", "no signs of editing"). As a last fallback, bare `yes`/`no` responses are mapped to `Edited`/`Original` respectively. Outputs that match none of these patterns are recorded as `None` and excluded from accuracy/precision/recall but counted in the per-model failure rate.

A.1.4 Output Parsing and Failure Handling

We parse responses with a three-stage extractor: (1) strip `<think>...</think>` blocks and Markdown fences; (2) attempt `json.loads` on the cleaned string; (3) on failure, fall back to a greedy regex for the outermost `{...}` block. If all three fail, the example is retried with a higher temperature (0.3 on the first retry, 0.5 on the second). After three failures the response is logged as `[Empty response]` / `error` and excluded from caption-quality metrics but still counted in the per-model failure rate reported in Section 4.

A.1.5 Edit-Type Vocabulary Rationale

The thirteen-token `edit_type` vocabulary was chosen to cover the editing operations exposed by the three consumer platforms in our study, plus a residual `other` class and an explicit `none` option for the no-edit case. `object_removal` and `generative_fill` cover the dominant Apple Clean Up / Pixel Magic Eraser operations; `object_replacement`, `background_change`, and `sky_replacement` cover the more aggressive Magic Editor and Galaxy Generative Edit operations; `face_edit` captures Pixel Best Take; `color_adjustment`, `style_transfer`, and `enhancement` capture global tone- and filter-style edits; `text_edit` covers caption or sign rewriting on the small fraction of images containing text. The vocabulary is held in English (rather than translated) so that label distributions remain directly comparable across the English, Bangla, and Farsi caption settings.

	Error rate (%)		
Quality	**Deletion**	**Spurious**	**Loc. shift**
Q2 (heavy)	4.42	3.63	1.21
Q4	2.90	1.73	0.71
Q6 (light)	2.04	1.21	0.44

Table 7: Pixel-level edit-localization errors under Cheng2020-anchor compression. *Deletion*: reference box unmatched at IoU ≥ 0.5; *Spurious*: compressed box with maximum IoU < 0.1; *Loc. shift*: match with IoU $\in [0.1, 0.5)$. Per-model variation is below 0.1 percentage points.

B Pixel-Level Edit-Localization Errors Under Compression

The captioning pipeline first localizes edited regions using an OpenCV-based pixel-difference stage. We evaluate errors at this stage separately from the VLM results in §5.1.

Method. Bounding boxes from the uncompressed pairs are used as references and matched to boxes from the corresponding compressed pairs using the Hungarian algorithm over the IoU matrix (Kuhn, 1955). A *deletion* is a reference box with no match at IoU ≥ 0.5; a *spurious detection* is a compressed box with maximum IoU < 0.1 against any reference box; and a *localization shift* is a match with IoU $\in [0.1, 0.5)$. Deletion and shift rates are normalized by the number of reference boxes, while spurious-detection rates are normalized by the number of compressed boxes.

Results. All error rates increase with compression strength (Table 7). From Q6 to Q2, deletions increase from 2.04% to 4.42%, spurious detections from 1.21% to 3.63%, and localization shifts from 0.44% to 1.21%. Per-model variation is below 0.1 percentage points, indicating that these errors arise primarily from the localization stage. Despite the increase, absolute error rates remain low, supporting the reliability of the pixel-difference stage under compression.

B.1 Examples of Original and Edited Images

Table 9, 10, 11, 12, 13 show the set of sample images from our dataset, showcasing both the original photos and their edited versions generated using Apple Intelligence tools on three different devices: iPad, iPhone, and MacBook.

Prompt: Edit Description (template)

```
You are an expert visual diff analyst. You will receive two images:
- IMAGE 1: the ORIGINAL unedited photo.
- IMAGE 2: the EDITED version, modified by a smartphone AI tool.
{platform_info}{tool_priors}
{bbox_info}

# Task
Identify what changed in IMAGE 2 relative to IMAGE 1. Use the pixel-diff
regions above as a hint about WHERE to look, but verify visually  pixel diff
can be noisy (alignment artifacts, JPEG re-encoding) and can also miss subtle
semantic edits.

# Output
Respond with a single JSON object and nothing else. No markdown fences, no
preamble, no explanation outside the JSON.

{
  "caption": "<one to three sentences, written in {lang_label}, describing
             concretely what changed>",
  "edit_type": "<one or more values from the English vocabulary below,
              comma-separated>",
  "confidence": "<high | medium | low>"
}

# edit_type vocabulary (use these English tokens exactly, do not translate)
object_removal, object_addition, object_replacement, background_change,
color_adjustment, style_transfer, face_edit, text_edit, generative_fill,
enhancement, sky_replacement, none, other

# Rules
- The caption value MUST be written in {lang_label}. The JSON keys and the
  edit_type values MUST stay in English exactly as listed above.


- Be concrete: name objects, colors, and positions (e.g. bottom-left, behind
  the person). Avoid vague phrasing.
- If multiple edits exist, describe each, ordered by visual prominence
  (largest or most noticeable first).
- Maximum 3 sentences. Stay grounded in what is actually visible.
- If IMAGE 2 looks identical to IMAGE 1, set edit_type to "none", set
  confidence to "high", and write a caption (in {lang_label}) stating that
  no visible edit was found.
- If you are uncertain whether a region truly changed, set confidence to
  "low" and reflect that uncertainty in the caption.
- Do not describe IMAGE 1 in isolation. The caption is about THE DIFFERENCE
  between the two images.
- Do not speculate about how the edit was made or the editor's intent  only
  describe what changed.

Output the JSON now. /no_think
```

Figure 7: Prompt template used to elicit edit descriptions from all eight evaluated VLMs. Curly-brace tokens are runtime fillers (see Table 6). The same template is reused across English, Bangla, and Farsi by changing only `{lang_label}`; the JSON keys and the edit-type vocabulary remain in English so that outputs are directly comparable across languages.

Prompt: Edit Description (LLaVA-OneVision variant)

```
You are looking at two images. IMAGE 1 is the original photo.
IMAGE 2 is an AI-edited version.



{region_section}
{platform_hint}



Compare IMAGE 2 to IMAGE 1 and identify what specifically changed in
IMAGE 2 relative to IMAGE 1. Be concrete: name objects, locations


(top/middle/bottom, left/center/right), and the nature of the change.

Respond ONLY with valid JSON in this exact format:


{
  "caption": "<2-4 sentences describing what changed, what was
             added/removed/altered, and where>",


  "edit_type": "<one or more of: object_removal, object_addition,
               object_replacement, generative_fill, sky_replacement,


               background_change, face_edit, text_edit,
               color_adjustment, style_transfer, enhancement, none,
               comma-separated>"


}

Use 'none' for edit_type only if IMAGE 2 appears truly identical to


IMAGE 1. Otherwise pick the most specific edit_type(s) that apply.
Multiple labels are allowed (comma-separated).



Output the JSON object and nothing else. Do not add commentary,
headers, or markdown fences. Respond in {target_language}.
```

Figure 8: Reduced-schema prompt used only for LLaVA-OneVision 7B. Compared to Figure 7, this variant (i) omits the `confidence` field, (ii) drops the `other` edit-type label, (iii) omits the tool-specific edit-type priors, and (iv) omits the Qwen-specific `/no_think` directive. The fillers `{region_section}` and `{platform_hint}` carry the same content as `{bbox_info}` and `{platform_info}{tool_priors}` in Figure 7, modulo the dropped tool priors.

Prompt: Binary Edit Detection (ablation)

Instruction — reference-free modes (`zeroshot_rf`, `fewshot_rf`)

```
You are inspecting a single photograph to
decide whether it has been digitally altered
by an AI photo-editing tool. An ALTERED photo
may contain removed or added objects,
AI-generated fill regions, or global tonal
changes (brightness, contrast, color,
sharpness) beyond a normal camera capture.
Look for inconsistencies in lighting, texture,
edges, repetition, or synthetically
smoothed/filled regions.
```

Instruction — pair modes (`zeroshot_pair`, `fewshot_pair`)

```
You are comparing two photographs of the same
scene. The FIRST is the reference; decide
whether the SECOND has been digitally altered
relative to it. Examine both carefully and
focus on any region where the second differs
from the first — an added or removed object,
an AI-filled area, or a tonal/color shift. If
you find such a region, answer EDITED; if the
two show the same scene with no alteration,
ORIGINAL.
```

Answer cue (appended to every query; label order randomized per item)

```
Answer with exactly one word, no explanation:
EDITED or ORIGINAL.
```

Few-shot demonstration prefix (`fewshot_rf`, `fewshot_pair`)

```
You will learn what AI photo edits look like
from labeled example pairs (original first,
edited second). Each note names the region
that changed, so you learn where to look.
Some scenes are unedited.
[EXAMPLE 1 - addition] (original then edited)
Edit: ⟨Gemma3-27B region caption, addition⟩
[EXAMPLE 2 - removal] (original then edited)
Edit: ⟨Gemma3-27B region caption, removal⟩
[EXAMPLE 3 - sketch] (original then edited)
Edit: ⟨Gemma3-27B region caption, sketch⟩
[EXAMPLE 4 - none] (original then the same
original)
Edit: none - the photo is unaltered.
```

The query image (single, or original+candidate) follows the prefix; the answer cue is appended last.

Figure 9: Text components for the binary edit-detection ablation: a single user turn (no system role); the instruction varies by reference mode, the answer cue is constant with per-item label-order randomization, and few-shot modes prepend three labeled exemplar pairs (Gemma3-27B region captions) plus one constructed unedited example. Thinking-mode models also receive a `/no_think` tag.

Pair 1 **Platform:** Apple iPad **Category:** Beach **Consensus:** `object_removal`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	In the bottom-right area of the image, a small dark object resembling a pebble or debris has been removed from the sand. The surrounding sand texture appears slightly smoothed and filled in, consistent with generative fill. No other visible changes are present.	In the bottom-right area of the image, a small dark object resembling a pebble or debris in the sand has been removed in IMAGE 2. The surrounding sand texture appears slightly smoothed and filled in, consistent with generative fill. No other visible changes are present.	In the bottom-right area of the image, a small dark object resembling a pebble or debris in the sand has been removed in IMAGE 2. The surrounding sand texture appears slightly smoothed and blended to match the rest of the beach, indicating a localized fill. No other visible changes are present.	The edited image shows minor alterations in the sand texture in the bottom-right and middle-right regions, where small dark specks or debris present in the original image have been removed, resulting in a smoother, more uniform sand surface. These changes are subtle and localized, consistent with object removal or generative fill to clean up small imperfections.
Gemma3	A small dark object, possibly a piece of debris, was removed from the sand in the bottom-right corner of the image. The area now appears as smooth sand, consistent with the surrounding texture.	A small dark object, possibly a piece of debris, was removed from the sand in the bottom-right corner of the image. The area now appears as smooth sand, consistent with the surrounding texture.	A small dark object near the bottom-right corner of the image has been removed. The area now blends seamlessly with the surrounding sand.	Small dark spots were removed from the sand in the bottom-right and middle-right of the image. These appear to be small pebbles or debris that were digitally erased.
Pixtral	A small object or disturbance in the bottom-right sand area was removed.	A small object or disturbance in the bottom-right sand area was removed.	A small object or disturbance in the sand at the bottom-right was removed.	Small objects or details in the sand were removed from the middle-right and bottom-right areas.
LLaVA-OV	The image has been edited to remove a small pixelated area in the bottom-right corner. The rest of the image remains unchanged, showing a sandy beach with footprints and the ocean in the background under a clear sky.	The image has been edited to remove a small pixelated area in the bottom-right corner. The rest of the image remains unchanged, showing a sandy beach with waves and a clear sky.	The image has been edited to remove a small region of the bottom-right corner, which appears to be a small object or artifact that was present in the original image.	The image has been edited to remove a small object in the bottom-right corner, which appears to be a small rock or pebble. The rest of the image remains unchanged, showing a sandy beach with waves in the background under a clear sky.
Phi-4-RV	In the bottom-right corner of the image, a small dark spot has been removed, leaving the sand texture uninterrupted. The rest of the beach scene remains unchanged.	In the bottom-right corner of the image, a small patch of sand has been removed, revealing a slightly darker, more uniform sandy texture beneath. The edit is subtle and localized, affecting only a tiny area of the beach scene.	In the bottom-right corner of the image, a small patch of sand has been removed, revealing a slightly darker, more uniform sandy texture beneath. The edit is subtle and localized, affecting only a tiny area of the beach scene.	In the AI-edited image, two small patches of sand in the bottom-right and middle-right areas have been removed, leaving behind slightly smoother, less textured sand. The overall scene remains a beach with waves and sky, but these minor edits subtly alter the natural grain of the sand.

Figure 10: Pair 1: **Apple iPad /** Beach. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_removal` as the primary edit_type in every condition (Table 8).

Pair 2 **Platform:** Apple iPhone **Category:** Building **Consensus:** `object_removal`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	In the bottom-left corner of the image, a dark, partially obscured object resembling a street sign or pole has been removed. The area is now filled with a seamless continuation of the sky and building reflections, blending with the surrounding glass facade. The edit creates a cleaner composition by eliminating the foreground obstruction.	In the bottom-left corner of IMAGE 2, a dark, angular structure resembling a building or overhang has been removed, leaving a cleaner view of the sky and adjacent skyscraper. The area now shows a continuation of the blue sky and clouds, seamlessly blending with the surrounding architecture. This change is consistent with object removal and generative fill to reconstruct the background.	In the bottom-left corner of IMAGE 2, a dark, angular structure resembling a building or architectural element has been removed. The area now shows a continuation of the sky and clouds, seamlessly blending with the surrounding glass facade reflections. The rest of the image remains unchanged.	In the bottom-left region, a dark vertical structure resembling a pole or signpost has been removed, leaving a cleaner view of the sky and building facade. In the middle-center region, a small dark object, possibly a camera or sensor, mounted on the building's edge has been removed, resulting in a smoother architectural line. Both edits appear to be object removals using generative fill to blend the background.
Gemma3	A small portion of the building in the bottom-left corner has been removed. The area now appears to blend with the surrounding sky and building facade.	A small portion of the sidewalk and a person's leg have been removed from the bottom-left corner of the image. The area is now filled with a continuation of the building's facade and the ground.	A small portion of the bottom-left corner of the image has been removed. Specifically, a section of the building and the sidewalk area have been replaced with a continuation of the sky and adjacent building facade.	A small dark object was removed from the bottom-left corner of the image, near the base of the building. Additionally, a small dark spot on the middle building has been filled in with the surrounding blue sky reflection.
Pixtral	An object was removed from the bottom-left corner of the image.	An object was removed from the bottom-left corner of the image.	An object was removed from the bottom-left corner of the image.	In IMAGE 2, there are slight enhancements and adjustments in the bottom-left and middle-center regions, likely involving object removal and color adjustments.
LLaVA-OV	The image has been edited to remove a small section of the bottom-left corner, which appears to be a part of a building or structure. The removed area is approximately 6.5% of the image's total area.	The image has been edited to remove a small section of the bottom-left corner, which appears to be a part of a building or structure that was not intended to be in the frame.	The image has been edited to remove a small section of the bottom-left corner, which appears to be a small object or detail that was not part of the original image. The rest of the image remains unchanged, showing a cityscape with tall buildings and a clear sky.	The image has been edited to remove a small section of the building on the left side, resulting in a more seamless integration with the surrounding architecture.
Phi-4-RV	In the bottom-left corner of the image, a small, dark object that resembles a traffic cone has been removed. The surrounding pavement area now appears uninterrupted and slightly brighter, as if the object's shadow and presence have been digitally erased.	In the bottom-left corner of the image, a small, dark object that resembles a traffic cone has been removed. The surrounding pavement area now appears uninterrupted and slightly brighter, as if the object's shadow has also been eliminated.	In the bottom-left corner of the image, a small, dark object that resembles a traffic cone has been removed. The surrounding pavement area now appears uninterrupted and slightly brighter, as if the object's shadow has also been eliminated.	In the AI-edited image, a small rectangular object that was present in the bottom-left corner of the original photo has been removed. Additionally, a tiny detail in the middle-center of the image has been altered, likely to smooth or clean up the scene.

Figure 11: Pair 2: **Apple iPhone /** Building. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_removal` as the primary edit_type in every condition (Table 8).

Pair 3 **Platform:** Apple MacBook **Category:** Cappadocia **Consensus:** `object_removal`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	In the top-center region of the image, a hot air balloon with a blue and white striped pattern has been removed. The area where the balloon was located now shows a clear patch of sky, seamlessly blending with the surrounding background. No other visible changes are present in the rest of the image.	In the top-center of the image, a hot air balloon with a blue and white striped pattern has been removed. The sky in that area now appears smoother and more uniform, with no visible balloon or its shadow, indicating a clean-up of the object.	In the top-center of the image, a hot air balloon with a blue and white striped pattern has been removed. The sky in that area now appears smoother and more uniform, with no visible balloon or its shadow, indicating a clean-up of the object.	In the top-center of the image, a hot air balloon with a blue and white striped pattern has been removed. The area where the balloon was located now shows only the sky and distant mountains, with no visible balloon or tether. The surrounding balloons and landscape remain unchanged.
Gemma3	A hot air balloon near the top-center of the image has been removed. The area where the balloon was is now filled with the sky color.	A hot air balloon near the top-center of the image has been removed. The area where the balloon was is now filled with a similar sky color, blending with the background.	A hot air balloon near the top-center of the image has been removed. The area where the balloon was is now filled with the sky color.	A hot air balloon near the top-center of the image has been removed. The area where the balloon was is now filled with a similar sky color.
Pixtral	An object was removed from the top-center of the image.	A hot air balloon located at the top-center of the image was removed.	A hot air balloon located at the top-center of the image was removed.	A hot air balloon located at the top-center of the image was removed.
LLaVA-OV	The image has been edited to remove a small region in the top-center area, which appears to be a small object or detail that was not part of the original scene.	The image has been edited to remove a small region in the top-center area, which appears to be a small object or detail that was not part of the original scene.	The image has been edited to remove a small region in the top-center area, which appears to be a small object or detail that was not part of the original scene.	The image has been edited to remove a small region in the top-center area, which appears to be a small object or detail that was not part of the original scene.
Phi-4-RV	In the edited image, a single hot air balloon that was originally positioned in the top-center of the sky has been removed. The surrounding balloons and landscape remain unchanged, creating a noticeable gap where the balloon once floated.	In the edited image, a single hot air balloon that was originally positioned in the top-center of the sky has been removed. The rest of the scene, including the landscape and the remaining balloons, remains unchanged.	In the edited image, a single hot air balloon that was originally positioned in the top-center of the sky has been removed. The rest of the scene, including the landscape and the remaining balloons, remains unchanged.	In the edited image, a single hot air balloon that was originally positioned in the top-center of the sky has been removed. The surrounding balloons and landscape remain unchanged, creating a noticeable gap where the balloon once floated.

Figure 12: Pair 3: **Apple MacBook /** Cappadocia. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_removal` as the primary edit_type in every condition (Table 8).

Pair 4 **Platform:** Galaxy Move **Category:** Colorado **Consensus:** `sky_replacement`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	The sky in the top-center region has been replaced with a different cloud formation, appearing more diffuse and less vibrant than the original fiery sunset. The snow-capped mountains remain, but the clouds above them are now softer and less defined, with a lighter, more uniform tone. The rest of the image, including the road, trees, and foreground, is unchanged.	The sky in the top-center region has been replaced with a lighter, more uniform cloud layer, removing the vibrant orange and red sunset hues present in the original. The snow-capped mountains are now partially obscured by this new cloud cover, altering the overall lighting and atmosphere. A small, faint white mark near the top-left edge of the image has also been removed.	The sky in the top-center region has been replaced with a more dramatic, cloud-filled sunset with enhanced orange and red tones, while the snow-capped mountains remain visible but appear slightly less defined. A small, dark, irregular shape near the top-left edge of the image has been removed, leaving a smoother transition between the sky and the mountain ridge.	The sky in the top-center region has been replaced with a more uniform, less detailed orange and red gradient, losing the intricate cloud textures and subtle color variations present in the original. The snow-capped mountain peaks appear slightly less defined and more blended into the sky, suggesting a generative fill that smoothed the transition. The overall lighting on the mountains and trees below appears slightly dimmer and less dynamic compared to the original.
Gemma3	The sky has been dramatically altered, with the original blue sky replaced by a vibrant, fiery orange and red sunset. The mountains in the background appear more pronounced and snow-covered due to the new sky. The overall scene has a warmer, more dramatic color cast.	The sky in the edited image has been dramatically altered with a vibrant orange and red color scheme, replacing the more muted tones of the original. Additionally, a small portion of the trees on the top left have been subtly altered, though the exact nature of the change is difficult to discern.	The sky has been dramatically altered with a vibrant orange and red color scheme, replacing the original, more subdued sky. Additionally, a small portion of the trees on the top left have been subtly altered, possibly to blend with the new sky.	The sky in the edited image has been dramatically altered to feature a vibrant, fiery orange and red sunset with more dramatic cloud formations. The original image had a much more subdued, pale blue sky. This appears to be a complete sky replacement.
Pixtral	The top-center of the image has been altered, with changes to the sky and clouds.	The sky in the top-center region has been altered, showing more vibrant and intense colors.	The sky in the top-center region has been altered, showing more vibrant and intense colors.	The top-center portion of the image has been altered, with changes to the sky and clouds.
LLaVA-OV	The sky has been altered to have a more uniform color, and the clouds have been removed in IMAGE 2 compared to IMAGE 1.	The sky has been altered to a more uniform color, and the overall image appears slightly brighter and more saturated.	The sky has been altered to a more uniform color, and the overall image appears slightly brighter and more saturated.	The sky has been altered to have a more uniform color, and the overall image has a slightly different color tone, giving it a more surreal appearance.
Phi-4-RV	The sky in the upper portion of the image has been replaced with a different sunset scene, featuring more clouds and a slightly different color palette. The mountains and landscape below remain unchanged.	The sky in the upper portion of the image has been replaced with a more dramatic, fiery sunset. The clouds and color gradients have been altered to appear more intense and vibrant, particularly in the top-center and top-left regions.	The sky in the upper portion of the image has been replaced with a more dramatic, fiery sunset. The mountains in the background appear slightly darker and more pronounced against the new sky.	The sky in the second image has been replaced with a more dramatic, fiery sunset. The clouds are more pronounced and the colors are more intense, with deeper oranges and reds. The mountains and landscape below remain unchanged.

Figure 13: Pair 4: **Galaxy Move** / Colorado. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `sky_replacement` as the primary edit_type in every condition (Table 8).

Pair 5 **Platform:** Galaxy Remove **Category:** Airport **Consensus:** `object_removal`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	The two gray suitcases in the bottom-center of the image have been removed in IMAGE 2. The area where the suitcases were is now filled with the airport floor and background, seamlessly blending with the surrounding environment. The rest of the scene, including the seating, windows, and airplane view, remains unchanged.	The two gray suitcases in the bottom-center of the image have been removed in IMAGE 2. The area where the suitcases were is now filled with the airport floor and background, seamlessly blending with the surrounding environment. The rest of the scene, including the seating, windows, and airplane view, remains unchanged.	The silver suitcase with an extended handle in the bottom-center of the image has been removed in IMAGE 2. The area where the suitcase was located is now filled with the airport floor and background, seamlessly blending with the surrounding environment. No other visible changes are present.	The two gray suitcases in the bottom-center of the image have been removed in IMAGE 2. The area where the suitcases were is now filled with the airport floor and background, seamlessly blending with the surrounding environment. The rest of the scene, including the seating, windows, and airplane view, remains unchanged.
Gemma3	A person was removed from the bottom-center of the image, near the seating area. The space where the person stood is now empty floor space.	A person has been removed from the bottom-center of the image, near the seating area. The space where the person stood is now filled with the floor pattern, suggesting a generative fill operation.	A person has been removed from the seating area in the center of the image. The space where the person was sitting is now empty.	A person has been removed from the bottom-center of the image, near the luggage. The space where the person was standing is now empty floor. The AI tool appears to have filled in the area with the surrounding floor texture.
Pixtral	A suitcase was removed from the bottom-center of the image.	A suitcase was removed from the bottom-center of the image.	A suitcase was removed from the bottom-center of the image.	A suitcase was removed from the bottom-center of the image.
LLaVA-OV	The image has been edited to remove the luggage in the foreground. The chairs and the background remain unchanged.	The image has been edited to remove the luggage in the foreground. The chairs and the background remain unchanged.	The image has been edited to remove the luggage in the foreground. The chairs and the background remain unchanged.	The image has been edited to remove the luggage in the foreground. The chairs and the background remain unchanged.
Phi-4-RV	The two suitcases in the foreground of the airport terminal have been removed in the edited image. The chairs and the airplane in the background remain unchanged.	The two suitcases in the foreground have been removed. The floor area where they once stood is now empty, and the seating area appears slightly more open.	The two suitcases in the foreground of the airport terminal have been removed in the edited image. The chairs and the airplane in the background remain unchanged.	The two suitcases in the foreground of the airport terminal have been removed in the edited image. The chairs and the airplane in the background remain unchanged.

Figure 14: Pair 5: **Galaxy Remove /** Airport. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_removal` as the primary edit_type in every condition (Table 8).

Pair 6 **Platform:** Galaxy Sketch **Category:** Art **Consensus:** `object_addition`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	A white sailboat with a gray and white sail has been added to the center of the image, positioned over the red-orange gradient background. The sailboat is oriented diagonally, with its bow pointing toward the bottom-right corner. The addition is isolated and does not alter the surrounding gradient or texture.	A white sailboat with a gray and white sail has been added to the center of the image, positioned over the red-orange gradient background. The sailboat is oriented diagonally, with its bow pointing toward the bottom-right corner. The addition is isolated to the middle-center region, with no other visible changes to the background or color gradient.	A white sailboat with a grey and white sail has been added to the center of the image, positioned over the red-orange gradient background. The sailboat is oriented diagonally, with its bow pointing toward the bottom-right corner. The addition is isolated to the middle-center region, with no other visible changes to the background or color gradient.	A white sailboat with a tall sail has been added to the center of the image, positioned over the red-orange gradient background. The sailboat is detailed with a blue circular logo on the sail and appears to be floating on the colorful surface. The surrounding gradient background remains unchanged.
Gemma3	A small sailboat has been added to the center of the image. The boat is white with a tall mast and a triangular sail. It appears to be floating on water, though no water is explicitly visible.	A white sailboat has been added to the center of the image. The boat is positioned against the gradient background, appearing to be on the water. The edit appears to be a generative fill.	A sailboat has been added to the center of the image. The boat is white with a blue sail and appears to be floating on water.	A small sailboat has been added to the center of the image. The boat is white with a blue sail and appears to be floating on water.
Pixtral	A sailboat with a white sail and red hull was added in the middle of the image.	A sailboat was added in the middle of the image.	A sailboat with a white sail and red hull was added in the middle of the image.	A sailboat was added in the middle of the image.
LLaVA-OV	A sailboat has been added to the center of the image, against a background of horizontal stripes in shades of yellow, orange, red, and pink.	A sailboat has been added to the center of the image, against a background of horizontal stripes in shades of yellow, orange, red, and pink.	A sailboat has been added to the center of the image, against a background of horizontal stripes in shades of yellow, orange, red, and pink.	A sailboat has been added to the center of the image, against a background of a sunset gradient from yellow to red to orange.
Phi-4-RV	A small sailboat has been added to the center of the image, floating on the gradient background. The boat is white with a single mast and sail, positioned slightly to the right of the vertical center.	A small sailboat has been added to the center of the image, floating on the gradient background. The boat is white with a single mast and sail, positioned slightly to the right of the vertical center.	A small sailboat has been added to the center of the image, floating on the gradient background. The boat is white with a single mast and sail, positioned slightly to the right of the vertical center.	A small sailboat has been added to the center of the image, floating on the gradient water. The boat is white with a single mast and sail, positioned slightly to the right of the vertical centerline.

Figure 15: Pair 6: **Galaxy Sketch** / Art. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_addition` as the primary edit_type in every condition (Table 8).

Pair 7 **Platform:** Google Pixel **Category:** Animal **Consensus:** `object_removal`

	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	In the edited image, the small yellow bird flying in the top-right corner has been removed. Additionally, the fishbowl with goldfish in the bottom-right has been replaced with a blank white area, and a small pink mouse near the fishbowl is also gone.	In IMAGE 2, the small yellow bird flying in the top-right has been removed. Additionally, the goldfish bowl in the bottom-right has been altered: the fish are gone, and the bowl appears empty with a faint white reflection. A small white object near the bowl, possibly a mouse or debris, has also been removed.	In IMAGE 2, the small yellow bird flying in the top-right has been removed. Additionally, the goldfish bowl in the bottom-right has been altered: the fish inside are no longer visible, and the bowl appears empty. A small white object near the bowl, possibly a mouse or a piece of debris, has also been removed.	In IMAGE 2, the small yellow bird flying in the top-right has been removed. Additionally, the goldfish bowl in the bottom-right has been altered: the fish inside are no longer visible, and the bowl appears empty. A small white object near the bowl, possibly a mouse or a piece of debris, has also been removed.
Gemma3	A yellow bird was removed from the top-right of the image. A goldfish bowl was removed from the bottom-right, and a mouse was removed from the bottom-right as well.	A yellow bird was removed from the top-right of the image. A goldfish bowl was removed from the bottom-right, and a mouse was removed from the bottom-right as well.	A small yellow bird was removed from the top-right of the image. A goldfish bowl was removed from the bottom-right, and a mouse was removed from the bottom-right as well.	A yellow bird was removed from the top-right of the image. A goldfish bowl was removed from the bottom-right, and a mouse was removed from the bottom-right as well.
Pixtral	In IMAGE 2, a small object in the top-right corner was removed, and minor adjustments were made to the bottom-right area.	In IMAGE 2, a small object in the top-right corner was removed, and minor adjustments were made to the bottom-right area.	In IMAGE 2, three small objects were removed from the top-right, bottom-right, and bottom-right regions of the image. These objects were likely minor elements that were not central to the composition.	In IMAGE 2, a small object in the top-right corner was removed, and minor adjustments were made to the background in the bottom-right area.
LLaVA-OV	The bird that was perched on the dog's back in the original image has been removed in the edited image.	The bird that was perched on the dog's back in the original image has been removed in the edited image.	The bird that was on the dog's back in the original image has been removed in the edited image.	The bird that was perched on the dog's back in the original image has been removed in the edited image.
Phi-4-RV	In the second image, the goldfish in the fishbowl on the right side of the group has been removed. The bowl itself remains, but the fish is no longer visible. Additionally, the small mouse on the floor near the bowl has been erased. The rest of the animals and background remain unchanged.	In the second image, the goldfish bowl with two goldfish in the bottom-right corner has been removed. Additionally, the small gray mouse sitting on the floor near the bowl is no longer present.	In the second image, the goldfish bowl and the small mouse on the floor in the bottom-right corner have been removed. The rest of the scene, including the dog, cats, birds, and other animals, remains unchanged.	In the second image, the goldfish bowl and the small mouse on the floor in the bottom-right corner have been removed. The rest of the scene, including the dog, cats, parrot, and other animals, remains unchanged.

Figure 16: Pair 7: **Google Pixel /** Animal. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_removal` as the primary edit_type in every condition (Table 8).

Pair 8 **Platform:** Apple AI → Galaxy Sketch **Category:** Room **Consensus:** `object_addition`

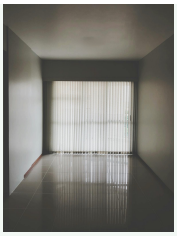
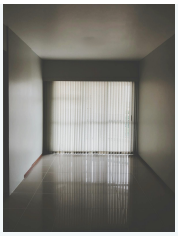
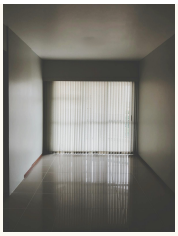
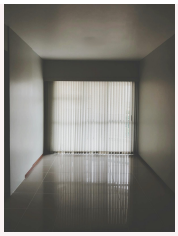
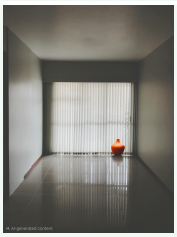
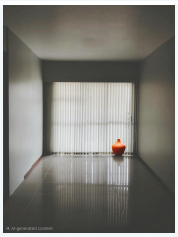
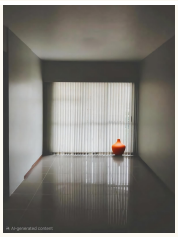
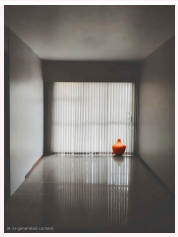
	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	An orange vase has been added to the floor in the middle-right area, near the window with vertical blinds. The vase is positioned on the reflective tiled floor, casting a subtle reflection. The rest of the scene, including the walls, ceiling, and window, remains unchanged.	A bright orange vase has been added to the floor in the middle-right area of the room, positioned in front of the vertical blinds. The vase is not present in the original image and casts a subtle reflection on the glossy tiled floor. The rest of the scene, including the lighting and shadows from the blinds, remains unchanged.	A bright orange vase has been added to the floor in the middle-right area of the room, near the window with vertical blinds. The vase is positioned on the reflective tiled floor and casts a subtle shadow. The rest of the scene, including the walls, ceiling, and window, remains unchanged.	A bright orange vase has been added to the floor in the middle-right area, near the window with vertical blinds. The vase is positioned on the reflective tiled floor, casting a subtle shadow and reflection. The rest of the scene, including the walls, ceiling, and window, remains unchanged.
Gemma3	A small orange vase has been added to the floor in the center of the room, between the blinds and the right wall. The vase is positioned slightly forward from the wall and is reflecting light.	A small, orange vase has been added to the floor in the center of the room, between the blinds and the right wall. The vase is positioned in front of the wall and reflects light from the window.	A small, orange vase has been added to the floor in the center of the room, between the blinds and the right wall. The vase is reflected in the tiled floor.	A small, orange vase has been added to the floor in the middle-right of the room. The vase is positioned in front of the window and reflects light onto the tiled floor.
Pixtral	An orange vase was added to the middle-right section of the floor in front of the window blinds.	An orange vase was added to the middle-right section of the floor in front of the window blinds.	An orange vase was added to the middle-right section of the floor in front of the window blinds.	An orange vase was added to the middle-right section of the floor in front of the window blinds.
LLaVA-OV	An orange vase has been added to the center-right of the image, on the floor in front of the window.	An orange vase has been added to the center-right of the image, on the floor in front of the window.	An orange vase has been added to the center-right of the image, on the floor in front of the window.	An orange vase has been added to the center-right of the image, on the floor in front of the window.
Phi-4-RV	A single orange vase has been added to the floor in front of the window blinds, near the right wall. The vase is centered horizontally in the lower half of the image.	A single orange vase has been added to the floor in front of the window blinds, positioned slightly to the right of center.	A single orange vase has been added to the floor in front of the window blinds, near the right wall. The rest of the room remains unchanged.	A single orange vase has been added to the floor in front of the window blinds, near the right wall. The rest of the room remains unchanged.

Figure 17: Pair 8: **Apple AI → Galaxy Move /** Room. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_addition` as the primary edit_type in every condition (Table 8).

Pair 9 **Platform:** Apple AI → Galaxy Sketch **Category:** Sky **Consensus:** `object_addition`

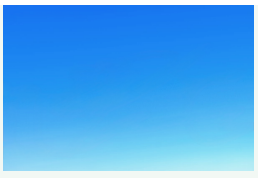
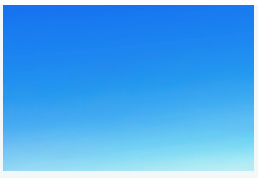
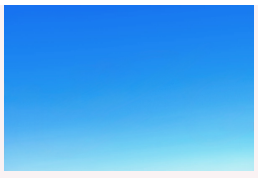
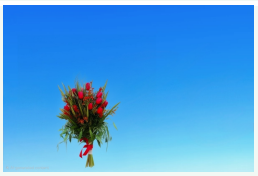
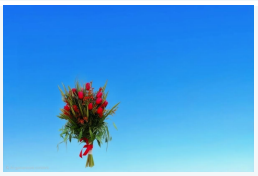
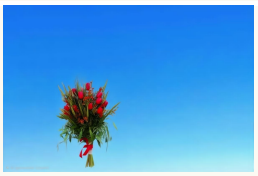
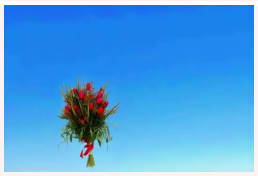
	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	A bouquet of red tulips with green foliage and a red ribbon has been added to the bottom-left portion of the image, floating against the blue sky gradient. The bouquet is positioned in the lower-left quadrant, occupying the region previously empty in the original image.	A bouquet of red tulips with green foliage and a red ribbon has been added to the bottom-center of the image, floating against the blue sky gradient. The bouquet is positioned in the lower half of the frame, centered horizontally, and appears to be suspended in mid-air with no visible support.	A bouquet of red tulips with green foliage and a red ribbon has been added to the bottom-center of the image, floating against the blue sky gradient. The bouquet is positioned in the lower half of the frame, centered horizontally, and appears to be suspended in mid-air with no visible support.	A bouquet of red flowers with green stems and a red ribbon has been added to the bottom-center of the image, floating against the blue gradient background. The bouquet is positioned in the lower half of the frame and was not present in the original image.
Gemma3	A bouquet of red and white flowers with a ribbon has been added to the bottom-left corner of the image. The bouquet appears to be composed of tulips and other foliage.	A bouquet of red and green flowers has been added to the bottom-center of the image. The flowers appear to be roses and other foliage, hanging downwards.	A bouquet of red and green flowers has been added to the bottom-center of the image. The flowers appear to be a mix of buds and fully bloomed blossoms with long stems and leaves.	A bouquet of red and green flowers has been added to the bottom-center of the image. The flowers appear to be a mix of roses and foliage, with a ribbon tied around the stems.
Pixtral	A bouquet of red flowers tied with a red ribbon was added at the bottom-left corner against a clear blue sky.	A bouquet of red flowers tied with a red ribbon was added at the bottom center of the image.	A bouquet of red flowers tied with a red ribbon was added at the bottom-left corner against a clear blue sky.	A bouquet of red flowers tied with a red ribbon was added at the bottom center of the image.
LLaVA-OV	A bouquet of red flowers has been added to the bottom of the image, against a clear blue sky.	A bouquet of red flowers has been added to the bottom center of the image, against a clear blue sky background.	A bouquet of red flowers has been added to the bottom left corner of the image, against a clear blue sky.	A bouquet of red flowers with green leaves and stems is added to the bottom center of the image, against a clear blue sky background.
Phi-4-RV	A bouquet of red tulips with green foliage and a red ribbon has been added to the bottom-left corner of the sky background. The flowers appear to be digitally inserted, contrasting with the clear blue sky.	A bouquet of red flowers with green foliage and a red ribbon has been added to the bottom-center of the sky background. The flowers appear to be floating in the air.	A bouquet of red flowers with green foliage and a red ribbon has been added to the bottom-left corner of the sky. The flowers appear to be floating against the blue background.	A bouquet of red flowers with green foliage and a red ribbon has been added to the bottom-center of the sky. The flowers appear to be floating against the blue background.

Figure 18: Pair 9: **Apple AI → Galaxy Remove /** Sky. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_addition` as the primary edit_type in every condition (Table 8).

Pair 10 **Platform:** Pixel → Galaxy Sketch **Category:** Desert **Consensus:** `object_addition`

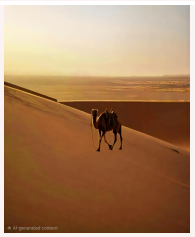
	Uncompressed	Q6_high	Q4_medium	Q2_low
Original				
Edited				
Qwen3-VL	A camel with a saddle and harness has been added to the center of the sand dune in IMAGE 2, walking uphill. The camel is positioned in the middle-center of the image, where the pixel-diff region is located. The surrounding desert landscape and lighting remain unchanged.	A camel with a saddle and a lead rope has been added to the center of the sand dune in IMAGE 2, standing on the slope facing left. The camel is not present in IMAGE 1, and the surrounding sand appears undisturbed around it, suggesting a generative addition. The background and lighting remain unchanged.	A camel with a saddle and a lead rope has been added to the center of the sand dune in IMAGE 2, standing on the slope facing left. The camel is not present in IMAGE 1, and its addition is the only visible change.	A camel has been added to the center of the sand dune in IMAGE 2, standing on the slope with a harness and a rope. The surrounding desert landscape and lighting remain unchanged from IMAGE 1. The addition is clearly visible in the middle-center region.
Gemma3	A camel with a saddle and rider has been added to the desert landscape. The camel is positioned in the middle of the image, walking across the sand.	A camel with a saddle and rider has been added to the desert landscape. The camel is positioned in the middle of the image, walking across the sand.	A camel and rider have been added to the desert landscape in the middle-center of the image. The camel is light brown and appears to be walking across the sand.	A camel and rider have been added to the desert landscape in the middle-center of the image. The camel is light brown and appears to be walking across the sand.
Pixtral	A camel was added in the middle of the desert landscape, walking down the sand dune.	A camel was added in the middle of the desert landscape, walking down the sand dune.	A camel was added in the middle of the desert landscape, walking down the sand dune.	A camel was added in the middle of the desert landscape, walking down the sand dune.
LLaVA-OV	A camel has been added to the image, walking on the sand dune in the middle-center region.	A camel has been added to the image, walking on the sand dune in the middle-center region.	A camel has been added to the image, walking on the sand dune in the middle-center region.	A camel has been added to the image, walking on the sand dune in the middle-center region.
Phi-4-RV	A camel with a saddle has been added to the middle-center of the desert landscape, standing on a sand dune. The rest of the scene, including the vast sand dunes and distant horizon, remains unchanged.	A camel with a saddle has been added to the middle-center of the desert landscape, standing on a sand dune. The rest of the scene, including the vast sand dunes and distant horizon, remains unchanged.	A camel with a saddle has been added to the middle-center of the desert landscape, standing on a sand dune. The rest of the scene, including the vast sand dunes and distant horizon, remains unchanged.	A camel has been added to the middle-center of the desert landscape, standing on a sand dune. The rest of the scene, including the sand dunes and distant horizon, remains unchanged.

Figure 19: Pair 10: **Pixel → Galaxy Move /** Desert. Each column shows the *actual input* the VLMs received at that compression level: row 1 is the original (before-edit) image, row 2 is the edited (after-edit) image, and rows 3–7 are the natural-language descriptions each of the five VLMs emitted. Captions in a given column describe the images in the same column. All five VLMs assigned `object_addition` as the primary edit_type in every condition (Table 8).

Table 8: Ten qualitative pairs where all five VLMs (Qwen3-VL, Gemma3, Pixtral, LLaVA-OneVision, Phi-4-RV) assigned the same primary `edit_type` under the uncompressed input *and* all three JPEG-AI compression levels (Q2, Q4, Q6) – i.e. unanimous 5/5 agreement in 4/4 conditions for every pair. The **Consensus** cell is shaded by edit type: green = `object_removal`, cyan = `object_addition`, maroon = `sky_replacement`. The selection covers all seven standalone platforms (Apple iPad/iPhone/MacBook, Galaxy Move/Remove/Sketch, Google Pixel) plus three cross-device pipelines, and spans ten distinct scene categories. Thumbnails are from the uncompressed pair.

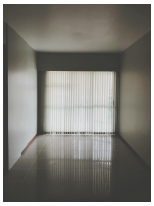
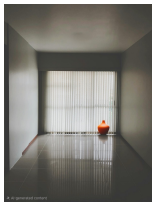
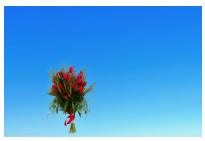
#	Original	Edited	Platform / Category	Consensus	Secondary-label models
1			Apple iPad / Beach	`object_removal`	Gemma3, Phi-4-RV, Qwen3-VL
2			Apple iPhone / Building	`object_removal`	Gemma3, Phi-4-RV, Pixtral, Qwen3-VL
3			Apple MacBook / Cappadocia	`object_removal`	Gemma3, Qwen3-VL
4			Galaxy Move / Colorado	`sky_replacement`	Gemma3, LLaVA-OV, Phi-4-RV, Pixtral, Qwen3-VL
5			Galaxy Remove / Airport	`object_removal`	Gemma3, Qwen3-VL
6			Galaxy Sketch / Art	`object_addition`	Gemma3, Phi-4-RV, Qwen3-VL
7			Google Pixel / Animal	`object_removal`	Pixtral, Qwen3-VL
8			Apple AI → Galaxy Sketch / Room	`object_addition`	Gemma3, Qwen3-VL
9			Apple AI → Galaxy Sketch / Sky	`object_addition`	Gemma3, Phi-4-RV, Pixtral, Qwen3-VL
10			Pixel → Galaxy Sketch / Desert	`object_addition`	Gemma3, Qwen3-VL

Table 9: Dataset Samples (Airport to Cappadocia)

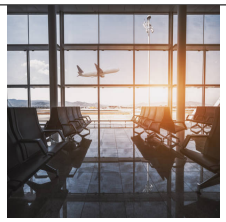
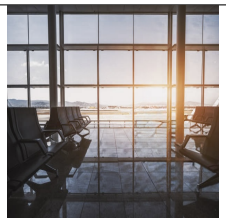
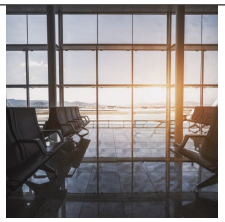
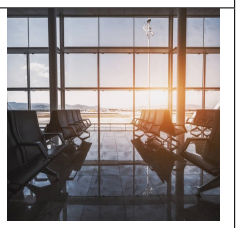
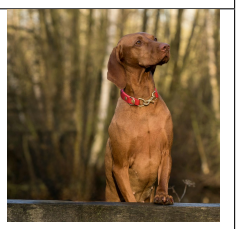
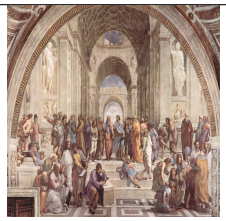
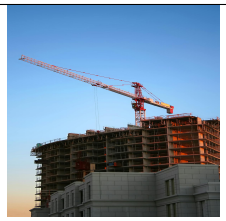
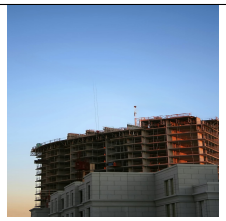
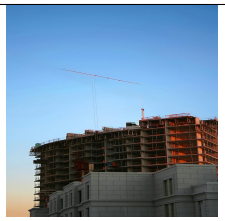
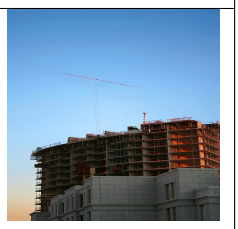
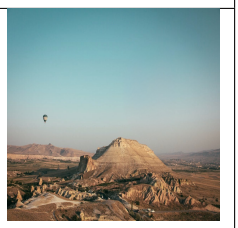
Dataset Folder	Original	iPad Edit	iPhone Edit	MacBook Edit
Airport Images				
Animal Images				
Art Images				
Beach Images				
Building Images				
Cappadocia Images				

Table 10: Dataset Samples (Coastal Life to Flower)

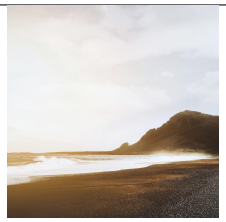
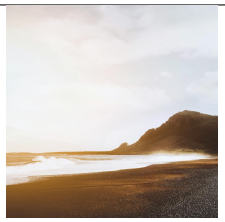
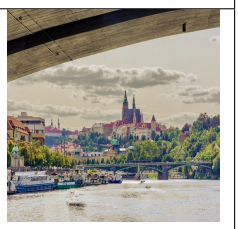
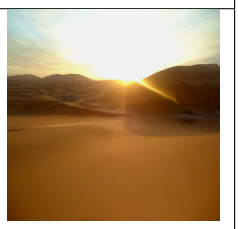
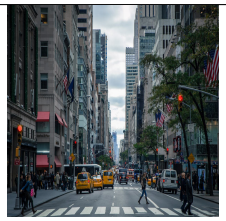
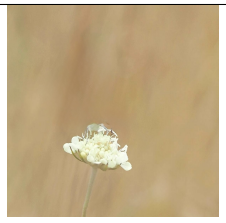
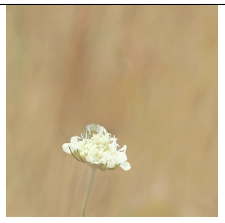
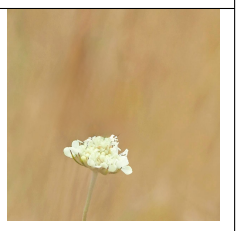
Dataset Folder	Original	iPad Edit	iPhone Edit	MacBook Edit
Coastal Life Images				
Colorado Images				
Czech Images				
Desert Images				
Downtown Images				
Flower Images				

Table 11: Dataset Samples (Forest to Madrid)

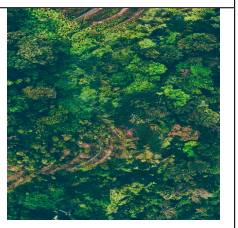
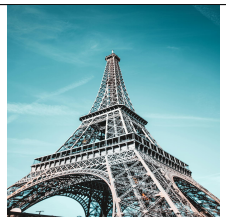
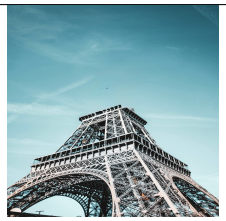
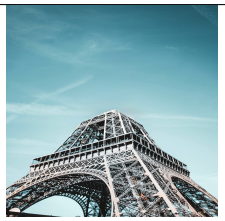
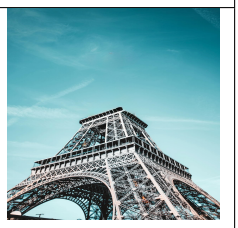
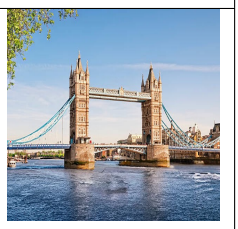
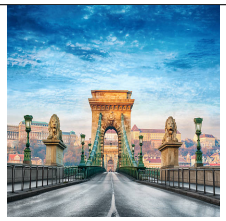
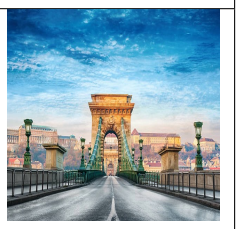
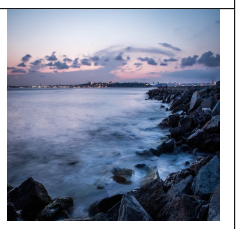
Dataset Folder	Original	iPad Edit	iPhone Edit	MacBook Edit
Forest Images				
France Images				
Geology & London Bridge Images				
Hungary Images				
Lighthouse Images				
Madrid Images				

Table 12: Dataset Samples (Mostar to Road)

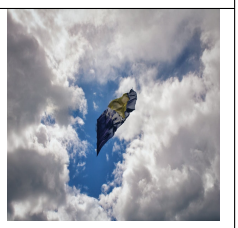
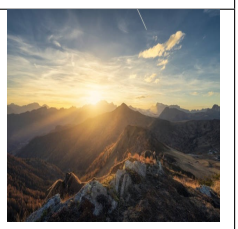
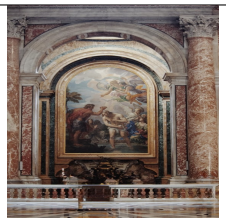
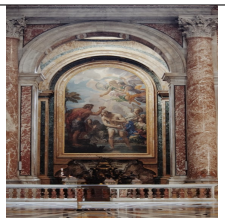
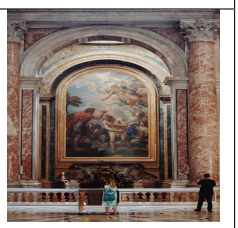
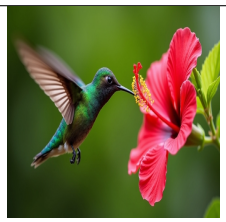
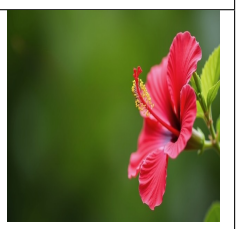
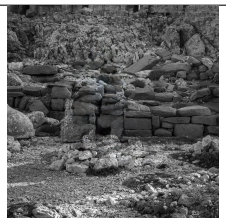
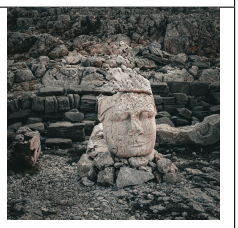
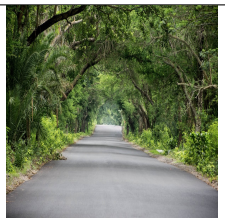
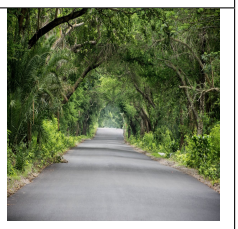
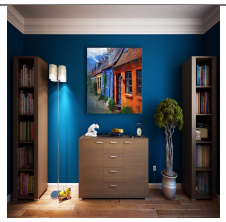
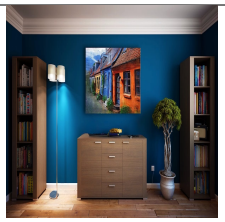
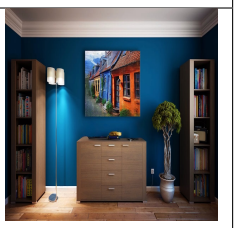
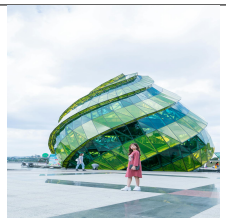
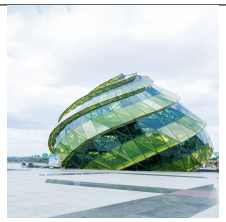
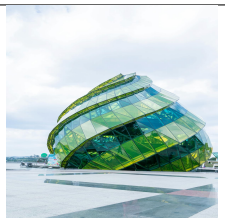
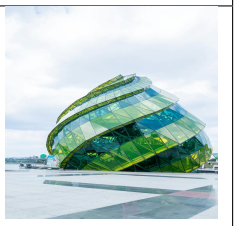
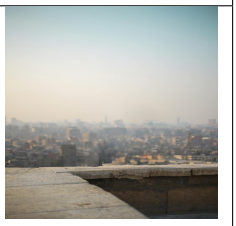
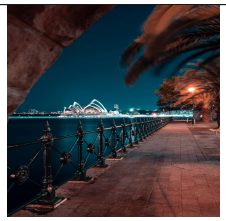
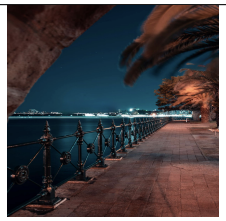
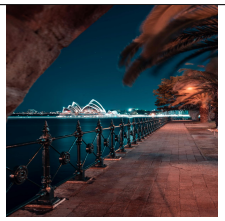
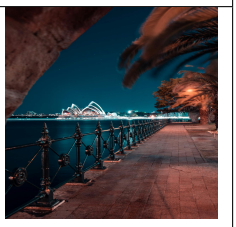
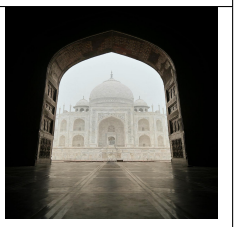
Dataset Folder	Original	iPad Edit	iPhone Edit	MacBook Edit
Mostar Images				
Mountain Images				
Museum Images				
Nature Images				
Night Images				
Nimrut Images				
Road Images				

Table 13: Dataset Samples (Room to Winter)

Dataset Folder	Original	iPad Edit	iPhone Edit	MacBook Edit
Room Images				
Scenery Images				
Sky Images				
Sydney Images				
Taj Mahal Images				
Winter Images			