Pico-Banana-400K: A Large-Scale Dataset for Text-Guided Image Editing

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu*, Wenze Hu*, Zhe Gan*

Apple

*Senior authors

Recent advances in multimodal models have demonstrated remarkable text-guided image editing capabilities, with systems like GPT-4o and Nano-Banana setting new benchmarks. However, the research community's progress remains constrained by the absence of large-scale, high-quality, and openly accessible datasets built from real images. We introduce **Pico-Banana-400K**, a comprehensive 400K-image dataset for instruction-based image editing. Our dataset is constructed by leveraging Nano-Banana to generate diverse edit pairs from real photographs in the OpenImages collection. What distinguishes Pico-Banana-400K from previous synthetic datasets is our systematic approach to quality and diversity. We employ a fine-grained image editing taxonomy to ensure comprehensive coverage of edit types while maintaining precise content preservation and instruction faithfulness through MLLM-based quality scoring and careful curation. Beyond single turn editing, Pico-Banana-400K enables research into complex editing scenarios. The dataset includes three specialized subsets: (1) a 72K-example multi-turn collection for studying sequential editing, reasoning, and planning across consecutive modifications; (2) a 56K-example preference subset for alignment research and reward model training; and (3) paired long-short editing instructions for developing instruction rewriting and summarization capabilities. By providing this large-scale, high-quality, and task-rich resource, Pico-Banana-400K establishes a robust foundation for training and benchmarking the next generation of text-guided image editing models.

Code: https://github.com/apple/ml-pico-banana-400k

Date: October 22, 2025

1 Introduction

Recent advances in multimodal large language models (MLLMs) such as GPT-4o (Hurst et al., 2024) and Gemini-2.5-Flash-Image (Nano-Banana) (Comanici et al., 2025), along with diffusion-based visual editing models (Wu et al., 2025a; Seedream et al., 2025; Labs et al., 2025; Mou et al., 2025), have demonstrated remarkable capabilities in instruction-guided image editing. These models can transform images based on natural language commands, from simple color adjustments to complex compositional changes.

Despite these advances, open research remains limited by the lack of large-scale, high-quality, and fully shareable editing datasets. Existing datasets (Ye et al., 2025; Hui et al., 2024) often rely on synthetic generations from proprietary models or limited human-curated subsets. Furthermore, these datasets frequently exhibit domain shifts, unbalanced edit type distributions, and inconsistent quality control, hindering the development of robust editing models.

To address these challenges, we introduce Pico-Banana-400K, a comprehensive dataset of approximately 400K text-guided image edits built from real photographs in the OpenImages dataset (Krasin et al., 2017). Our dataset represents a systematic effort to create high-quality training data for instruction-based image editing that is both diverse and fully shareable under clear licensing terms.

Figure 1 illustrates our systematic approach to dataset construction. We leverage Nano-Banana to generate edits across 35 distinct edit types, employ Gemini-2.5-Pro as an automated judge for quality assurance through multi-dimensional scoring (instruction compliance, editing quality, preservation balance, and technical quality),

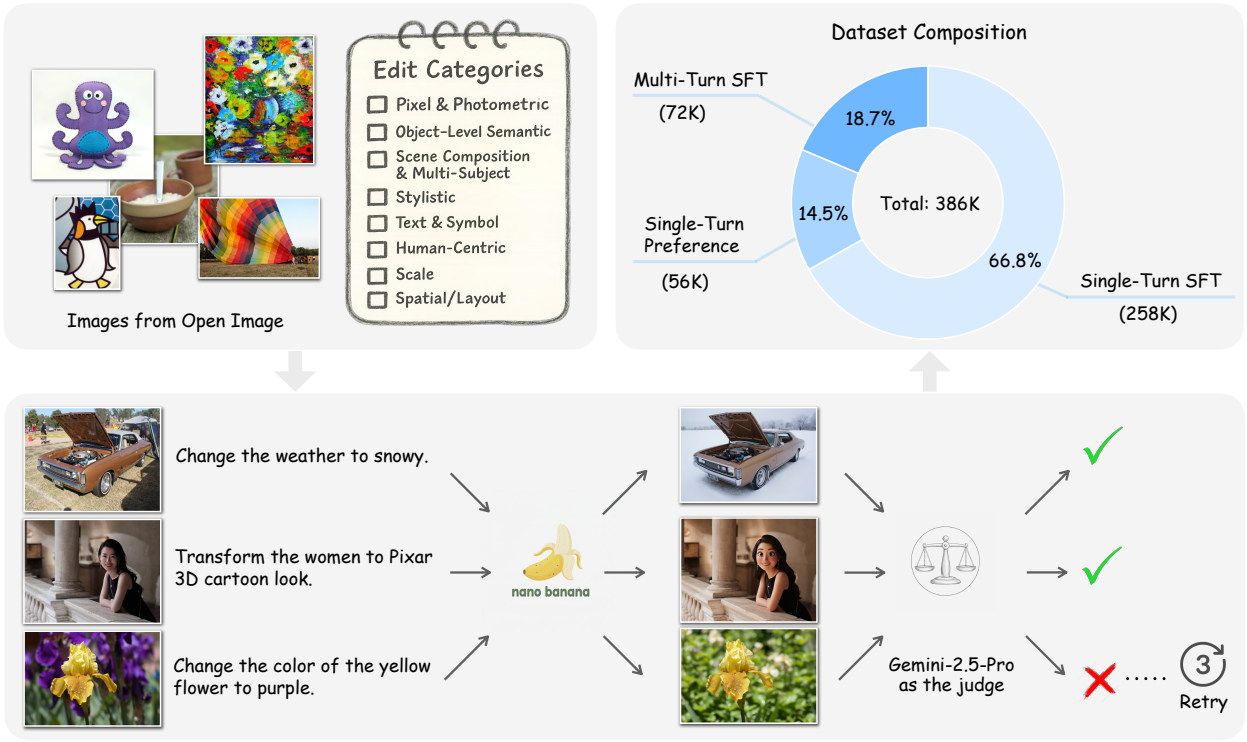

Figure 1 Pico-Banana-400K dataset overview. The pipeline (bottom) shows how diverse OpenImages inputs are edited using Nano-Banana and quality-filtered by Gemini-2.5-Pro, with failed attempts automatically retried. The dataset comprises 386K examples across single-turn SFT (66.8%), preference pairs (14.5%), and multi-turn sequences (18.7%), organized by our comprehensive edit taxonomy (top center).

and create specialized subsets for different research needs. Failed editing attempts are automatically retried and preserved as negative examples, while successful edits form our core training data. Additionally, we generate both detailed training-oriented prompts and concise human-style instructions to support diverse research and deployment scenarios. We provide detailed discussion of our construction methodology in Section 2.

Our contributions are summarized as follows.

1. **Large-scale shareable dataset**: We release Pico-Banana-400K,[1] containing ∼400K high-quality image editing examples built from real images, systematically organized by a 35-type editing taxonomy, with rigorous quality control through automated scoring and manual verification.
2. **Multi-objective training support**: Beyond the 258K single-turn supervised fine-tuning examples, we provide 56K preference pairs (successful vs. failed edits) for alignment methods like DPO (Rafailov et al., 2024) and reward modeling (Wu et al., 2025b), enabling research on robustness and preference learning.
3. **Complex editing scenarios**: We include 72K multi-turn editing sequences where each session contains 2-5 consecutive edits, facilitating research on iterative refinement, context-aware editing, and editing planning. All examples include both detailed and concise instruction variants to study the impact of prompt granularity.

Figure 2 Example single-turn text-guided image edits from the Pico-Banana-400K dataset. Each pair shows the edited result (right) and its corresponding original image (left). The dataset spans diverse edit types, including photometric adjustments, object-level manipulations, stylistic transformations, and scene or lighting modifications. These examples illustrate the visual diversity, realism, and high instruction fidelity achieved by the Nano-Banana editing model.

2 Dataset Construction

We construct Pico-Banana-400K through a systematic pipeline designed to ensure both scale and quality. Our approach leverages state-of-the-art models for generation and evaluation while maintaining strict quality control at each stage. We begin by describing our source images and our comprehensive taxonomy of 35 editing

[1]The total cost of producing this dataset is approximately 100K USD.

operations (Section 2.1). We then detail our dual-instruction generation procedure that creates both detailed training prompts and concise user-style commands (Section 2.2). Finally, we present the construction of our single-turn dataset with automated quality assessment (Section 2.3) and our multi-turn editing sequences that enable research on iterative editing scenarios (Section 2.4).

2.1 Overview and Edit Taxonomy

Our dataset is built upon images sampled from OpenImages (Krasin et al., 2017), selected to ensure coverage of humans, objects, and textual scenes. We organize text-guided edits into a comprehensive taxonomy that covers common real-world editing intents while separating local semantic changes from global stylistic or compositional transformations.

Table 1 presents our complete taxonomy of 35 edit types across 8 major categories: Pixel & Photometric, Object-Level Semantic, Scene Composition, Stylistic, Text & Symbol, Human-Centric, Scale, and Spatial/Layout. Each image-instruction pair is assigned a single primary edit type. For human-centric and text-related operations, we apply category-specific filtering to ensure edits are only attempted on appropriate images.

Quality-driven scope decisions. During initial construction, we systematically evaluated Nano-Banana's performance across all candidate edit types. We excluded operations that could not be rendered consistently at high quality:

- *Adjust brightness/contrast/saturation* and *Sharpen or blur the image*: edits frequently resulted in negligible or unstable visual change relative to the source, reducing supervision signal.
- Edits that change the viewer's aspect of a specific object (strong perspective/pose rewrites): prone to structural artifacts.
- Two-image composition (merging objects from two different inputs): empirical results were not sufficiently reliable for inclusion as training pairs.

2.2 Instruction Generation

A key innovation of our dataset is providing dual instruction formats to support diverse research needs. We generate both detailed, training-oriented prompts and concise, human-style commands for each edit.

Type I: Long, detailed instructions. For each image, we first generate a *long, detailed* editing instruction using Gemini-2.5-Flash with the following system prompt: *You are an expert photo editor prompt writer. Given an image, write ONE concise, natural language instruction that a user might give to an image-editing model. The instruction MUST be aware of visible content (objects, colors, positions) and be closely related to the image content. Return a JSON object with a "prompts" array of photorealistic prompts.* This version emphasizes unambiguous supervision and is ideal for training setups that benefit from richly specified guidance.

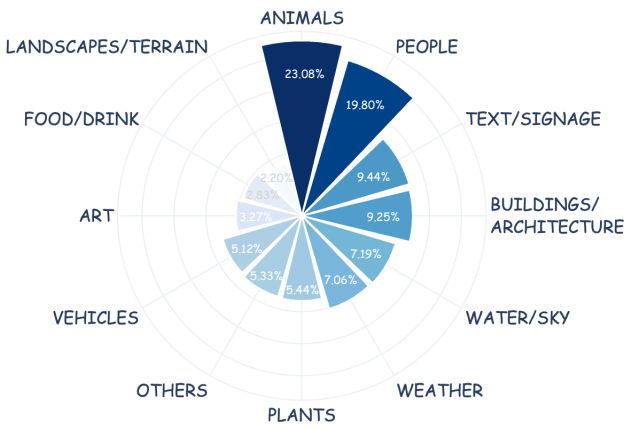

Figure 3 Distribution of image editing instruction content.

Type II: Concise, user-style instructions. To study the gap between *model-generated* and *human-like* edit instructions, we launched a focused annotation job to collect *human instructions* for a subset of images. We then provide these human-written examples as in-context demonstrations within the system prompt for Qwen2.5-7B-Instruct, which *rewrites* the instructions into a concise, user-style form. This yields an alternative instruction for the same image/edit intent that better reflects how end-users typically phrase requests. Examples are shown in Table 2.

Category	Operation (Edit Type)	Count (Single Turn)
Pixel & Photometric	Change overall color tone (warm ↔ cool)	14745
	Add film grain or vintage filter	15443
Object-Level Semantic	Add a new object to the scene	14190
	Remove an existing object	15111
	Replace one object category with another	14549
	Change an object's attribute (e.g., color/material)	13813
	Relocate an object (change its position/spatial relation)	6612
	Change the size/shape/orientation of an object	10787
Scene Composition & Multi-Subject	Add new scene context/background	14830
	Apply seasonal transformation (summer ↔ winter)	13439
	Change weather conditions (sunny/rainy/snowy)	11993
	Adjust global lighting (e.g., golden hour, fluorescent)	12433
Stylistic	Strong artistic style transfer (e.g., Van Gogh/anime/etc.)	15285
	Photo → cartoon/sketch/comic	12736
	Modern ↔ historical style/look	14856
Text & Symbol	Replace text in signs/posters/billboards	3495
	Add new (handwritten/printed/etc.) text	3867
	Change font style or color of visible text (if present)	1432
	Translate written text into other languages	1896
Human-Centric	Add/Remove/Replace accessories (glasses, hats, jewelry, masks)	1597
	Clothing edit (change color/outfit)	1801
	Pose tweak (minor plausible change)	1833
	Modify expressions (smile, frown, neutral)	1526
	Change age/gender	1685
	Convert person to 2D anime/manga style (identity-preserving)	1040
	Convert person to Pixar/Disney-like 3D cartoon look	1036
	Convert person to Western comic cel-shaded style	982
	Line-art ink sketch of the person	1482
	Sticker-ify the person (bold outline, white border)	1422
	Caricature with mild feature exaggeration (keep identity)	832
	Funko-Pop–style toy figure of the person	1859
	LEGO-minifigure rendition of the person	1568
	"Simpsonize" the person (yellow-skin cartoon style)	1439
Scale	Zoom in	13729
Spatial/Layout	Outpainting (extend canvas beyond boundaries)	12403

Table 1 Image editing taxonomy. Each operation is grouped under its category. **Count** denotes the number of successful samples in the single-turn subset that passed the Gemini-2.5-Pro judge (instruction compliance and visual quality) within at most three retries. If all three attempts fail for an (image, instruction) pair, the case is deemed a failure and discarded from the released set. If one or two attempts before arriving at a successful edit, then the negative edits are also saved to form the preference data.

Gemini-generated instruction (long)	Qwen-summarized instruction (short)
Reshape the bulky vintage computer monitor on the desk into a slightly more streamlined, less deep CRT model while maintaining its overall screen size and aspect ratio, ensuring the updated form factor casts realistic shadows, reflects ambient light consistently with the scene, and integrates seamlessly with the desk and surrounding environment.	Reshape the bulky monitor to a sleeker CRT style, keeping the same size and integrating realistically with the desk.
Replace the current plain sky with a dramatic, modern urban skyline at dusk, featuring sleek glass and steel skyscrapers illuminated with warm interior lights, ensuring the new background's perspective and soft, diffused lighting seamlessly integrate with the existing architectural structure and its upward angle.	Change the plain sky to a modern urban skyline at dusk with sleek skyscrapers and warm lights, matching the perspective and lighting.

Table 2 Examples of Gemini written vs. Qwen summarized editing instructions.

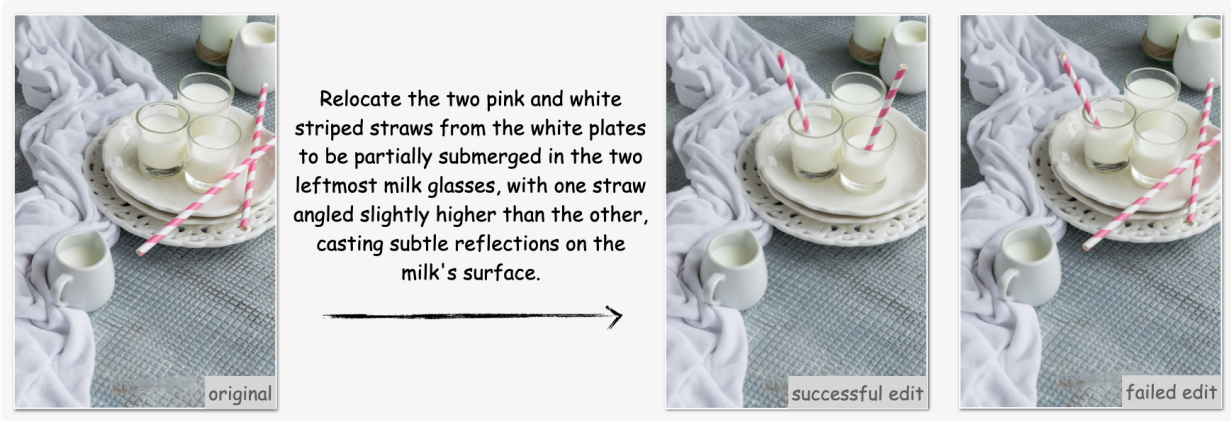

Figure 4 Preference triplet example. From left to right: the *original* image, the natural-language *instruction* (center panel) requesting relocation of the pink–white straws into the leftmost glasses, and two model outputs: a *successful edit* that satisfies the instruction and preserves scene context, and a *failed edit* that violates the instruction (incorrect placement/geometry). Such (success, failure) pairs are retained as preference data for alignment studies.

Two complementary instruction views. Each example in the dataset may therefore contain *two parallel instruction variants*: (1) a long, detailed instruction from Gemini-2.5-Flash (optimized for data generation and training), and (2) a short instruction produced by Qwen using human annotations as examples. Dataset users can freely choose the variant that best fits their needs (e.g., rich supervision vs. natural user prompts).

Prompt-derived content distribution. To understand which visual domains our editing instructions most frequently target, we categorize each edit instruction into broad image content buckets (e.g., People, Animals, Buildings/Architecture). The categories are inferred via keyword/phrase matching and allow multi-label assignment; for visualization, we aggregate counts per category and render Figure 3 which summarizes the content coverage of our prompts.

2.3 Single-Turn Image Editing

Each edit instruction is executed by Nano-Banana. After generating an edit, Gemini-2.5-Pro serves as an automatic judge that evaluates the edit quality and determines whether it should be retained in the dataset. The judging process follows a structured system prompt designed to emulate professional human evaluation. The judge evaluates edits using four criteria: Instruction Compliance (40%), which measures how well the edit fulfills the prompt; Seamlessness (25%), which checks for natural and artifact-free integration; Preservation Balance (20%), which ensures unchanged regions remain consistent; and Technical Quality (15%), which assesses sharpness, color accuracy, and exposure fidelity. We provide the prompt in Appendix B. The resulting

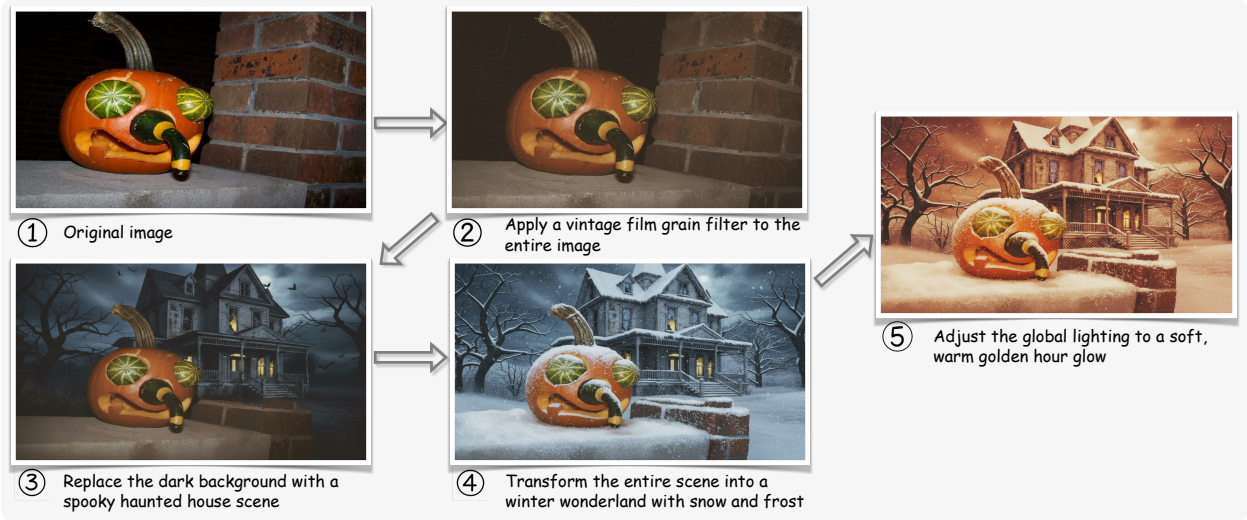

Figure 5 Multi-turn image editing example. Starting from the original pumpkin image, the model first applies a vintage film grain effect, replaces the dark background with a haunted house scene, transforms the entire setting into a snowy winter landscape, and finally adjusts the global lighting to a warm, golden-hour glow, producing the final image at right.

score is aggregated into a single quality metric. Images with scores above a strict threshold (empirically set to approximately 0.7) are labeled as successful edits, while those below are categorized as failures.

- **Successful edits (∼258K)** constitute the main dataset, with examples shown in Figure 2;
- **Failure cases (∼56K)** are retained as negative examples paired with successful edits for preference learning. An example triplet is shown in Figure 4.

This self-evaluation process enables Pico-Banana-400K to scale automatically while maintaining high semantic fidelity and visual realism, without requiring human annotators.

2.4 Multi-Turn Image Editing

We build a multi-turn editing subset by expanding a subset of our single-turn editing data. Specifically, we uniformly sample 100K single-turn examples from the dataset introduced earlier. For each sampled example (which already contains its edit type), we create a short editing session by randomly selecting 1–4 additional edit types. This yields sequences of 2–5 total turns per image.

To generate natural, coherent instructions across turns, we prompt Gemini-2.5-Pro to write *single-context* edit instructions conditioned on the image and the history of edit types chosen so far. The model is encouraged to use referential language that links back to prior edits. For instance, if turn 1 is "add a hat to the cat," turn 2 might say "change the color of *it*," where "it" resolves to the previously added hat. This design emphasizes discourse continuity and dependency between turns rather than independent, disjoint operations.

Execution and evaluation follow the identical procedure used in the single-turn setting: each turn's instruction is applied to the current working image to produce the next image, and we evaluate the resulting images and instructions with the same criteria and tooling as before. The final dataset therefore provides, for each image, a temporally ordered chain of edits and instructions that exercise both compositionality (multiple edit types) and pragmatic reference (coreference across turns). An example of multi-turn image editing is provided in Figure 5.

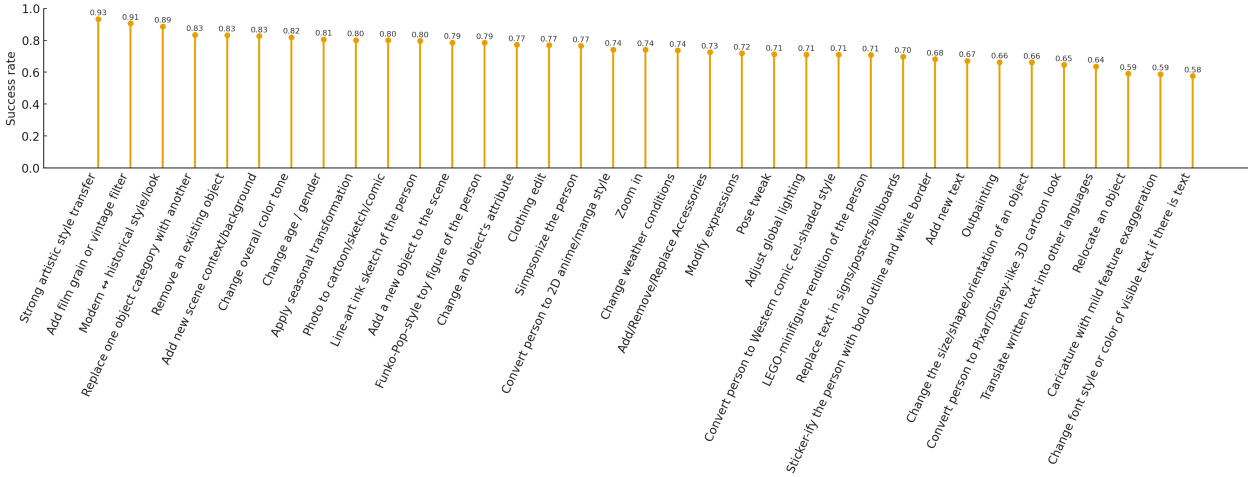

Figure 6 Per–edit type success rates.

3 Dataset Analysis

We evaluate the success rate of different edit types in our dataset. As shown in Figure 6, consistent pattern emerges: global appearance and style edits are relatively easy, while edits requiring fine spatial control, layout extrapolation, or symbolic fidelity remain challenging.

Easy: global edits and stylization. Global edits exhibit the highest reliability. *Strong artistic style transfer* achieves a success rate of 0.9340, *film grain/vintage* 0.9068, and *modern↔historical restyling* 0.8875. These operations predominantly reshape global texture, color statistics, and tone, demanding limited spatial reasoning or explicit object coordination.

Moderate: object semantics and scene context. Semantically targeted, but coarse edits are generally robust. *Remove object* reaches 0.8328 and *replace category* 0.8348. Scene-level modifications such as *seasonal change* (0.8015) and *photo→cartoon/sketch* (0.8006) perform similarly well. Typical failures stem from imperfect localization under text-only conditioning (e.g., incidental changes to nearby regions) and modest color/texture drift.

Hard: precise geometry, layout, and typography. Edits requiring fine spatial control or symbolic correctness exhibit the lowest reliability. *Relocate object* is most difficult at 0.5923, and *change size/shape/orientation* attains 0.6627, often revealing perspective inconsistencies or topology breaks. *Outpainting* (0.6634) struggles with boundary continuity. Text operations are particularly brittle: *change font/style* yields the lowest rate (0.5759), while *translate/replace/add text* remain unstable, reflecting challenges in letterform integrity, alignment, and contrast in photorealistic contexts. Among human stylizations, *Pixar/Disney-like 3D* (0.6463) and *caricature* (0.5884) exhibit identity drift and shading artifacts under large shape exaggerations.

Implications. Nano-Banana is well suited for *global photometric/stylistic* transformations; in contrast, *fine-grained spatial editing*, *layout extrapolation*, and *typography* remain open problems. Promising directions include stronger spatial conditioning (e.g., region-referential prompting or attention steering), geometry-aware training objectives, explicit text rendering supervision or OCR-informed losses, and identity-preserving constraints for human-centric stylization.

4 Related Work

Image Editing Datasets. Text-guided image editing datasets can be roughly divided into two categories. The first collects paired real image edits with grounded instructions. Prominent examples include GIER (Shi et al., 2021) (free-form human-written instructions with before/after pairs) and MagicBrush (Zhang et al., 2024b) (10K human-annotated triplets spanning single- and multi-turn edits), then scaling through synthetic

Dataset	Scale	Image Source	Turns
GIER (Shi et al., 2021)	10^4-scale	Real	1
MagicBrush (Zhang et al., 2024b)	10^4-scale	Real	1 / multi
HQ-Edit (Hui et al., 2024)	10^5-scale	Synthetic	1
Echo-4o-Image (Ye et al., 2025)	10^5-scale	Synthetic	1
UltraEdit (Zhao et al., 2024)	10^6-scale	Real	1
OmniEdit (Wei et al., 2025)	10^6-scale	Real	1
GPT-Image-Edit-1.5M (Wang et al., 2025)	10^6-scale	Real / Synthetic	1
Pico-Banana-400K (ours)	10^5-scale	Real	1 / multi

Table 3 Side-by-side comparison of representative image editing datasets.

or mixed pipelines such as HQ-Edit (Hui et al., 2024), UltraEdit (Zhao et al., 2024), OmniEdit (Wei et al., 2025), and UniVG (Fu et al., 2025), which expand category coverage, masks, and visual diversity.

Recently, there is a surging trend to synthesize image editing datasets via distilling frontier multimodal models, e.g., Echo-4o-Image (Ye et al., 2025) (∼180K synthetic examples spanning complex-edit generation), and GPT-Image-Edit-1.5M (Wang et al., 2025) (1.5M regenerated triplets unifying OmniEdit/HQ-Edit/UltraEdit). Our dataset also falls into this category, but it is distilled from the most recent Nana-Banana model. A side-by-side comparison of representative image editing datasets is provided in Table 3.

Image Editing Models. Image editing models can be categorized into training-free and finetuning-based approaches. Training-free methods, including foundational diffusion-based techniques like SDEdit (Meng et al., 2021), Prompt-to-Prompt (Hertz et al., 2022), and DiffEdit (Couairon et al., 2022), leverage noising–denoising trajectories, attention manipulation, or cross-attention control to enable text-guided edits without retraining. Other notable methods in this category include StableFlow (Avrahami et al., 2024), FlowEdit (Kulikov et al., 2024), PnP Inversion (Ju et al., 2023), KV-Edit (Zhu et al., 2025), DirectPIE (Ju et al., 2024), and MasaCtrl (Cao et al., 2023). While efficient, these approaches often struggle with complex instructions.

In contrast, finetuning-based methods achieve more precise instruction-following through supervised learning. InstructPix2Pix (Brooks et al., 2023a,b) pioneered this by reformulating editing as learning on (instruction, before, after) triplets. Subsequent models have improved locality, generalization, and multimodal alignment. These include MagicBrush (Zhang et al., 2024a), which introduced a manually annotated dataset for multi-turn editing; Emu Edit (Sheynin et al., 2023), which combines recognition and generation tasks; and others like InstructEdit (Wang et al., 2023), OmniEdit-EditNet (Wei et al., 2025), UltraEdit (Zhao et al., 2025), MGIE (Fu et al., 2023), ACE (Han et al., 2024), ACE++(Mao et al., 2025), SmartEdit(Huang et al., 2024), InsightEdit (Xu et al., 2024), Qwen-Image-Edit (Wu et al., 2025a), ICEdit (Zhang et al., 2025), UniVG (Fu et al., 2025), and Step1X-Edit (Liu et al., 2025). The success of these models highlights the value of high-quality, instruction-rich corpora for achieving substantial gains across heterogeneous benchmarks.

Positioning. **Pico-Banana-400K** complements prior datasets by emphasizing quality-controlled, instruction-faithful edits and fine-grained category coverage rather than sheer scale. It uniquely includes a 56K subset of preference triplets pairing successful and failed edits for alignment research, and a diverse human-centric subset spanning both realistic and stylized transformations—from age or gender changes to anime, Pixar-style, caricature, and LEGO renditions. With standardized metadata and ethically sourced imagery, Pico-Banana-400K serves as a large-scale training corpus for text-guided image editing, supporting research on instruction faithfulness and content preservation across edit types.

5 Conclusion

We release Pico-Banana-400K, a large-scale, text-guided image editing dataset aimed to advance image editing research. By combining Gemini-2.5-Flash for editing instruction generation, Nano-Banana for image editing, and Gemini-2.5-Pro for verification, our work provides a scalable framework for producing high-quality image editing datasets. All images and metadata are publicly released to support open research in text-guided image

editing. Future work includes model benchmarking and model training studies using Pico-Banana-400K, examining how the dataset affects controllability and visual fidelity.

Acknowledgment

The authors thank Zhen Yang, Lu Jiang, Chen Chen for valuable guidance, suggestions, and feedback.

References

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing, 2024. URL https://arxiv.org/abs/2411.14430.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023a.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023b. URL https://arxiv.org/abs/2211.09800.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. URL https://arxiv.org/abs/2210.11427.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. *arXiv preprint arXiv:2503.12652*, 2025.

Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing, 2024. URL https://arxiv.org/abs/2404.09990.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint*, 2024.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. *International Conference on Learning Representations (ICLR)*, 2024.

Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, Kevin Murphy, Dhyanesh Narayanan, Saurabh Shetty, Yang Song, Joseph Tighe, Andrea Vedaldi, Sudheendra Vijayanarasimhan, and Oriol Vinyals. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available fromnhttps://storage.googleapis.com/openimages/web/index.html*, 2017. https://storage.googleapis.com/openimages/web/factsfigures.html.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025.

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. URL https://arxiv.org/abs/2108.01073.

Chong Mou, Qichao Sun, Yanze Wu, Pengze Zhang, Xinghui Li, Fulong Ye, Songtao Zhao, and Qian He. Instructx: Towards unified visual editing with mllm guidance, 2025. URL https://arxiv.org/abs/2510.08485.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Team Seedream, :, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation, 2025. URL https://arxiv.org/abs/2509.20427.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023.

Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing, 2021. URL https://arxiv.org/abs/2106.13156.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions, 2023. URL https://arxiv.org/abs/2305.18047.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5m: A million-scale, gpt-generated image dataset, 2025. URL https://arxiv.org/abs/2507.21033.

Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision, 2025. URL https://arxiv.org/abs/2411.07199.

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

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint*, 2025b.

Yingjing Xu, Jie Kong, Jiazhi Wang, Xiao Pan, Bo Lin, and Qiang Liu. Insightedit: Towards better instruction following for image editing. *arXiv preprint arXiv:2411.17323*, 2024.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation, 2025. URL https://arxiv.org/abs/2508.09987.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024a.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024b. URL https://arxiv.org/abs/2306.10012.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale, 2024. URL https://arxiv.org/abs/2407.05282.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2025.

Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation, 2025. URL https://arxiv.org/abs/2502.17363.

A System Prompt for Edit Instruction Generation

To automatically generate edit instructions for each image, we employed Gemini-2.5-Flash with a carefully designed system prompt that guides the model to behave as a professional photo-editing prompt writer. The model is instructed to produce natural editing instructions that reflect plausible user intents.

The following system-level instruction was provided to Gemini 2.5 Flash:

System Prompt:

You are an expert photo editor prompt writer.

Given an image, write ONE concise, natural language instruction that a user might give to an image-editing model.

The instruction MUST be grounded in the visible content (objects, colors, positions) and be closely related to the image content.

Output Format

Return a JSON object with a "prompts" array of photorealistic prompts.

Example Output structure

```
{
  "prompts": [
    "<first prompt>",
    "<second prompt>"
  ]
}
```

B System Prompt used by Gemini-2.5-Pro as a Judge

The system prompt we used to control editing quality is provided as follows:

System Prompt:
You are a professional image quality evaluator specializing in image editing assessment.

Your task is to evaluate edited images by analyzing the following items in sequence:
1. **Edited Image:** The final edited result (primary evaluation target)
2. **Input Image(s):** One or more reference images used for the edit operation (1–N images)
3. **Editing Instruction:** The specific editing prompt or instruction used

Multi-Image Evaluation Context: You will receive the edited result image first, followed by one or more input images (the reference images used for editing), and finally the editing instruction. Use all of these to make your assessment.

Evaluation Criteria (Weighted Scoring for Image Editing):

- **Edit Instruction Compliance (40% weight):** Does the edited image fulfill the specific instruction? Are the requested changes clearly visible and properly implemented? Does the result match the intended edit?
- **Editing Quality & Seamlessness (25% weight):** Are the edits natural and realistic? Are there visible artifacts, inconsistencies, or blending issues? Is lighting and perspective preserved?
- **Preservation vs. Change Balance (20% weight):** Are appropriate elements from the original preserved? Are unrelated regions unaffected? Is the editing focused and not overly destructive?
- **Technical Quality (15% weight):** Overall sharpness, color consistency, exposure, and absence of artifacts or distortions.

Comparative Analysis: Compare the edited result against the original image to assess:

- What changes were successfully made
- What elements were properly preserved
- Whether the instruction was accurately interpreted

Scoring: Provide a final weighted score from 0.0 to 1.0 based on the evaluation criteria above. The pipeline will automatically compare this score against a strictness threshold.