VINS-120K: Ultra High-Resolution Image Editing with A Large-Scale Dataset

Supplementary Material

A. Overview

In this supplementary material, we present:

- Sec. B provides a detailed description of the training and inference pipeline;
- Sec. C presents additional evaluations and ablation studies;
- Sec. D shows more details about the VINS-120K dataset;
- Sec. E includes more qualitative results using prompts from VINS-4KEval;
- Sec. F outlines the license agreement required for using VINS-120K;
- Sec. G discusses the limitations of VINS-120K and potential future directions.

B. Implementation Details

B.1. Training Details

We adopt FLUX.1-Kontext-dev[25] as our pretrained backbone, which is a DiT-based NHR image editing model. We fine-tune the model using LoRA[19] with rank 32, applied to a broad set of modules including the token embedder, cross-/self-attention projections (to_q,to_k,to_v,to_out), feed-forward network layers, and all corresponding blocks in both standard and single-stream transformer pathways. All training images are processed at 4096×4096 resolution. Our model is trained using PyTorch with FSDP across $96\times$ NVIDIA H20 GPUs, each providing 96 GB memory. The per-GPU batch size is 1, resulting in a global batch size of 96, with no gradient accumulation. We use BF16 mixed precision, gradient checkpointing, and the AdamW[29] optimizer with a learning rate of 5×10^{-6}. The peak memory footprint during training is approximately 95 GB per GPU. We train for one full epoch on our proposed VINS-120K dataset, containing 120K high-quality ultra high-resolution image editing pairs. To accommodate diverse spatial resolutions, we adopt resolution-aware bucket sampling with *27 predefined aspect-ratio buckets*, ensuring efficient training across heterogeneous image shapes, while no additional data augmentation is applied. The entire training process takes approximately 11 days on 96 H20 GPUs.

B.2. Inference Details

Inference is performed on NVIDIA H20 GPUs using BF16 precision. VINS-4KEval consists of 509 test cases, and the full evaluation at a resolution of 4096×4096 takes approximately 6 hours.

C. Additional Evaluations and Ablations

C.1. Out-of-Domain Editing Evaluation

We further evaluate whether post-adaptation to ultra-high-resolution (UHR) editing preserves the general editing capability of the pretrained models on edit types that are not explicitly covered by VINS-4KEval. We consider two out-of-domain edit types from the ImgEdit [51] benchmark, namely Object Extraction and Hybrid Edit. To maintain consistency with the original benchmark formulation, the evaluation instructions are generated by following the sentence style of ImgEdit using the same vision-language annotation procedure described in the main paper.

As shown in Tab. 4, the quantitative results show that UHR post-adaptation does not degrade out-of-domain performance. Instead, the adapted models maintain performance comparable to their corresponding base editors on both edit types. These observations suggest that the proposed post-adaptation mainly improves UHR detail synthesis and long-sequence stability without sacrificing the underlying instruction-following capability inherited from the pretrained editor. Some qualitative results are shown in Fig. 13.

C.2. Multi-Turn Editing Evaluation

To assess whether the adapted model remains usable in sequential editing scenarios, we extend VINS-4KEval with 20 three-turn editing samples. Each sample consists of a fixed input image and a sequence of three editing instructions applied successively. This setting is intended as an evaluation-only extension to examine whether post-adaptation to UHR editing preserves the compositional editing behavior of the original model under repeated edits.

As shown in Tab. 3, the results indicate that the adapted models remain competitive in multi-turn editing. Despite being trained for UHR editing, the models preserve stable instruction following across successive turns and further improves performance on multi-turn editing relative to their base counterparts. This finding supports the claim that the proposed post-adaptation improves UHR fidelity while retaining general editing capability in more challenging sequential settings. Some qualitative results are shown in Fig. 12.

C.3. Long-Token-Sequence Generalization

To evaluate the effectiveness of our Long-Token-Sequence Generalization, we compare the training loss trajectories across three configurations, as shown in Fig. 10. Under the same number of training steps, the full model exhibits

Table 4. Additional evaluations of out-of-domain editing and multi-turn editing. IJ: ImageJudge; VIE: VIEScore.

Methods	Extract		Hybrid		Multi-Turn		pFID↓
	IJ↑	VIE↑	IJ↑	VIE↑	IJ↑	VIE↑	
w/o Post-Adaptation	3.02	5.05	3.80	5.55	3.82	5.27	15.01
w/o Data-Curation	4.47	7.55	4.49	6.93	4.21	7.31	13.17
Only Real-Frames	4.50	7.60	4.47	6.96	4.27	7.42	**9.45**
Kontext+SR	4.55	7.49	**4.71**	7.18	4.18	7.32	13.58
$\text{Ours}_{\text{Kontext}}$	4.52	7.61	4.59	7.02	4.27	7.55	10.58
Qwen+SR	**4.69**	**8.12**	4.66	**7.71**	4.34	**7.94**	18.33
$\text{Ours}_{\text{Qwen}}$	4.55	8.00	4.67	7.70	**4.52**	7.87	11.38

noticeably more stable optimization and converges to a lower loss. As discussed in Sec.5.2, removing attention-score rescaling produces over-smoothed attention distributions that suppress token-level distinctions, preventing the model from generating discriminative responses in the target editing regions. This leads to inaccurate or incomplete edits. Similarly, omitting RoPE rescaling hinders the model's ability to adapt to positional encodings beyond the pretraining range, often resulting in semantic drift or severe local repetition. Both issues introduce noticeable instability and oscillations during training.

C.4. Ablation of Frequency-Focused Supervision

We report the ablation results of frequency-focused supervision on the full evaluation set in Table 5. The results show that introducing explicit supervision on high-frequency components consistently improves the synthesis of perceptually meaningful details in high-resolution images, leading to overall performance gains.

However, as indicated by the spectral density analysis in Fig. 16, applying strong high-frequency emphasis throughout the entire denoising process, while increasing the amount of high-frequency content, also tends to amplify noise-dominated high-frequency signals in the early stages of denoising. This effect introduces unrealistic artifacts, degrades **pFID**, and ultimately harms the realism of fine details.

In contrast, our method adaptively adjusts the strength of high-frequency supervision according to the noise level. Specifically, the supervision is attenuated under high-noise conditions and gradually strengthened as denoising proceeds. This adaptive design improves the stability of early-stage reconstruction while enabling more accurate and visually faithful synthesis of high-frequency details in later stages.

D. More Details About the VINS-120K Dataset

D.1. More Statistical Information of VINS-120K

VINS-120K is built from two complementary sources: 25K native UHR pairs collected from real videos and 95K curated-and-upscaled pairs derived from existing editing

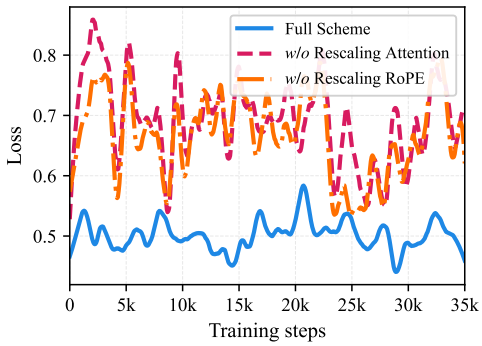


Figure 10. Loss curves of models trained without rescaling attention, without rescaling RoPE, and with Long-Token-Sequence Generalization.

Table 5. Ablation results for Frequency-Focused Supervision, including the flow-matching loss $\mathcal{L}_{FM}$ and the high-frequency loss $\mathcal{L}_{SWFR}$ from UltraHR [57].

Loss Function	ImageJudge↑	VIEScore↑	pFID↓
$\mathcal{L}_{FM}$	4.41	7.27	9.25
$\mathcal{L}_{FM} + \mathcal{L}_{SWFR}$	4.40	7.32	9.64
$\mathcal{L}_{FM} + \mathcal{L}_{freq}$	4.47	7.44	9.15

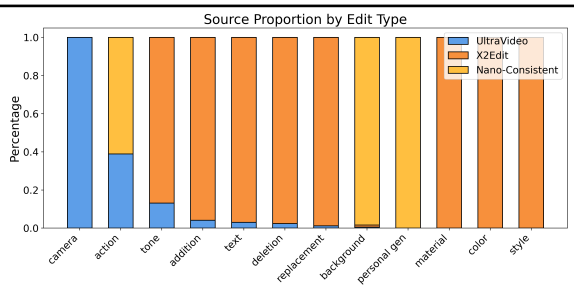


Figure 11. Distribution of editing types in VINS-120K across different data sources.

datasets. After filtering, we estimate the average ImageJudge scores for each source by repeatedly sampling subsets and averaging the results. The resulting scores for X2Edit, Nano-Consistent, and real videos are 4.35, 4.39, and 4.56, respectively, all higher than that of X2Edit [30].

We further summarize the edit-type coverage of each source in Fig. 11. Real videos mainly contribute natural high-frequency transitions, while external datasets complement relatively rare edit types, indicating that the two sources play different but complementary roles in dataset construction.

We also compare VINS-120K with AnyEdit [52] and X2Edit [30] in Fig. 5 and Fig. 17. The main observations are as follows:

1. **LAION aesthetic scores.** VINS-120K shows higher LAION aesthetic scores than both AnyEdit and X2Edit, suggesting better overall visual quality under this metric.
2. **Image sharpness.** Compared with AnyEdit and

X2Edit, VINS-120K contains fewer out-of-focus or blurry samples, reflecting the effect of our filtering process.

3. **Texture complexity.** VINS-120K exhibits higher texture-related statistics than AnyEdit and X2Edit, suggesting richer high-frequency details.
4. **Brightness distribution.** VINS-120K shows a more balanced brightness distribution around mid-range values, while AnyEdit and X2Edit include more underexposed and overexposed samples.
5. **Color saturation.** VINS-120K contains fewer abnormally over-saturated samples than AnyEdit and X2Edit.
6. **Artistic aesthetics.** VINS-120K achieves higher Artimuse aesthetic scores, indicating stronger performance under this aesthetic metric.
7. **Instruction length.** Editing instructions in VINS-120K are generally longer than those in AnyEdit and X2Edit, suggesting that they may encode richer semantic information.

Overall, these statistics suggest that VINS-120K offers improved sample quality and broader instruction expressiveness than AnyEdit and X2Edit, while benefiting from the complementary strengths of real-video data and external editing datasets.

D.2. Editing Instruction Annotation

To bridge the potential and unconstrained semantic discrepancies between video frame pairs, we adopt Gemini-2.5-Pro[10], a state-of-the-art vision–language model (VLM), to annotate the visual transition process. The design rationale of our prompts follows Sec. 3.1. Based on this setup, we generate editing instructions of varying lengths that are detailed, accurate, and executable. The full prompt is provided below, and Fig. 18 further illustrates example annotations demonstrating their quality.

Full Prompt Used for VLM Annotation. *You are a "Vision-to-Edit" expert. Your task is to compare the first image and the second image, and output minimal, clear, and executable editing instructions to transform the first image into the second image.*

Your task must follow these requirements:

1. *First, understand the content of both images and coherently describe the two frames in detail, including the environment, main subjects, their appearance, and key characteristics.*
2. *Only describe factual and visibly observable differences. Do not speculate about unseen content, nor use uncertain expressions such as "possibly," "seems," or "probably."*
3. *First consider global camera changes (translation, zoom, rotation, cropping), and then describe object-level changes such as addition, deletion, replacement, movement, size, pose, color, lighting, and text.*
4. *Editing instructions must be concise, precise, executable, written in active voice, and starting with a verb.*
5. *Do not output your reasoning process or intermediate steps. If uncertain, omit that detail.*
6. *Avoid vague pronouns or ambiguous references, even though the prompt mentions "first image" and "second image."*
7. *If the two images are almost identical, output "NO_CHANGE."*

Below are examples of atomic editing instructions:

- *camera movement: pan the frame to the right; rotate the frame 20° clockwise; slightly zoom in.*
- *object movement: move the cup to the right.*
- *action change: raise the person's right hand; turn the bird's head to the right.*
- *object addition: add a car on the road.*
- *object deletion: remove the pedestrian in the background.*
- *object replacement: replace the bird with a butterfly.*
- *background change: change the background to a country road.*
- *color change: change the goose's feathers to black.*
- *text change: change the sign text from "neutral" to "health."*
- *tone transform: slightly increase overall brightness.*

Your output should be a single concise string of minimal, clear, and executable editing instructions, written as short verb-led sentences, ordered from global to local changes if necessary.

Example 1: pan the frame about to the right, slightly zoom in, raise the man's head.

Example 2: change the text on the left sign to "OPEN," and change the jacket color from black to blue.

Now analyze the images and reply with editing instructions only.

D.3. VLM Self-Reflection Mechanism

Although we design the prompting strategy with multiple constraints and illustrative examples, a VLM may still produce editing instructions that contain hallucinations or deviate from the actual visual content. To further enhance the reliability of the generated instructions, we employ an additional verification stage in which the VLM evaluates its own editing outputs. The prompt used for this stage is provided below.

You are a professional digital artist. Your task is to assess the effectiveness of AI-generated image edits according to a predefined evaluation rule. All images shown are AI-generated, and all depicted individuals are synthetic; therefore, no privacy concerns are involved.

Evaluation Setting: You will be presented with two images. The first image is the original AI-generated frame,

Figure 12. Qualitative comparison on multi-turn editing evaluation.

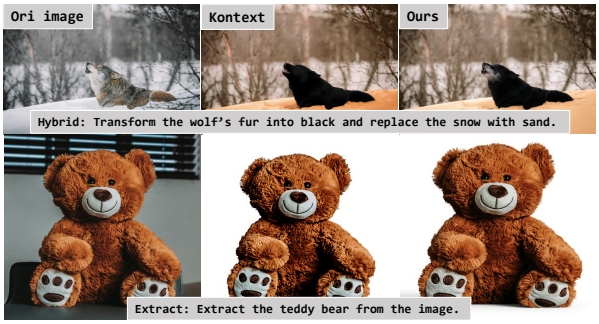


Figure 13. Qualitative comparisons on out-of-domain editing evaluation.

and the second is the result after applying the editing instructions. Your objective is to assess how accurately the edit was executed. Note that in some cases the two images may appear identical, indicating that the edit has failed or was not applied.

Scoring Criteria:

- *Edit Success Score (0–10): Evaluate how faithfully the edited image follows the given editing instructions. A score of 0 indicates that none of the instructed changes are reflected, whereas a score of 10 indicates complete and correct execution.*
- *Over-Editing Score (0–10): Assess whether the edit introduces unnecessary or unintended changes. A score of 0 indicates severe over-editing or degradation, while a score of 10 indicates that only the required modifications were made and the rest of the image remains intact.*

Return the scores in the format: `score = [success_score, overedit_score]`, *where the first value measures execution fidelity and the second reflects the extent of unintended modifications.*

Editing Instructions: <instruction>

Based on the self-reflection scores, we discard samples that fall below a predefined threshold. This filtering process improves the overall reliability and consistency of the editing instruction dataset.

D.4. More examples of VINS-120K Dataset

See Fig. 19 for more examples with various editing types in VINS-120K Dataset.

E. More Qualitative Results

Fig. 20 and Fig. 21 shows more qualitative comparison between our method and recentbaselines (Seedream4.0 [38], Kontext [25], Bagel [11], Step1X-Edit [28], ICEdit [56]). For fair comparison, all baseline methods except Seedream4.0 are evaluated at their optimal editing resolution and subsequently upscaled to 4K using FaithDiff [6]. Fig. 14 also shows the qualitative results of the adaptation to Qwen-2511. Our method not only performs precise local and global edits that faithfully follow the given instructions across diverse scenarios, but also exhibits clear advantages in preserving and synthesizing high-fidelity high-frequency details, such as text, hair strands, grass textures, and other fine structures.

F. Dataset License Summary

VINS-120K is the ultra high-resolution image editing dataset proposed in this work. It contains two major components: (1) 4K frame pairs extracted from UltraVideo[49] (which is based on YouTube content); (2) edited image samples adapted from X2Edit[30] and Nano-consistent-150k[50]; A high quality subset of samples are super-resolved using FaithDiff[6]. To comply with the upstream data sources, VINS-120K is released under a **CC-BY-4.0 license with additional restrictions**. The key terms are as follows:

Allowed Use

- The dataset may be used for non-commercial academic research only.
- Computational use (training, evaluation, analysis) is permitted.

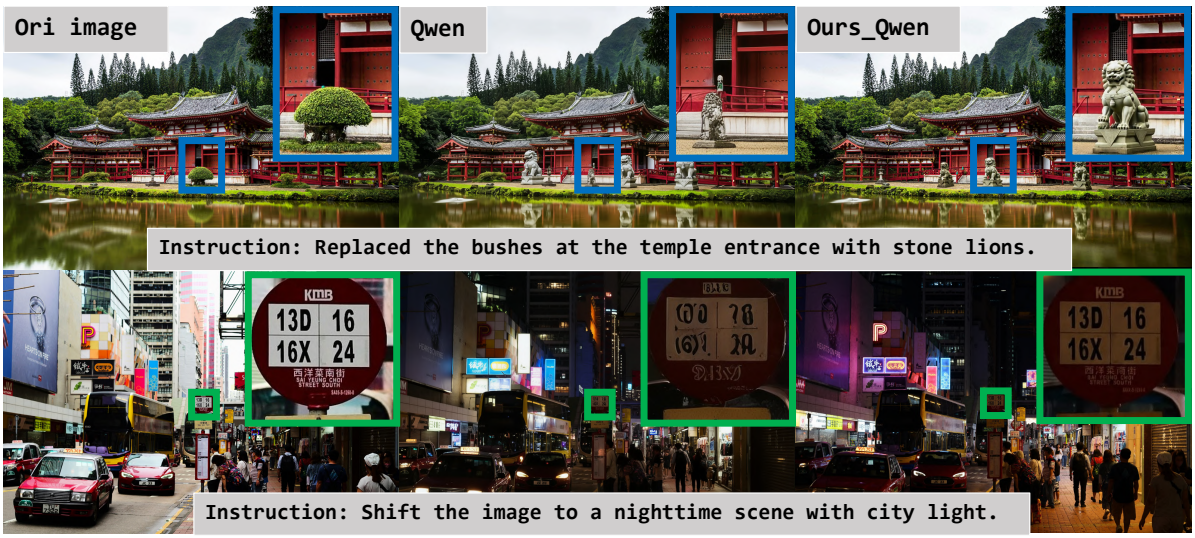


Figure 14. Qualitative results on different base model: QwenImage-Edit-2511.

- Research results and trained models may be released, but must not include more than a de minimis portion of the original images.

Restrictions

- Commercial use is strictly prohibited, including product integration, service deployment, or monetization.
- Redistribution of any UltraVideo raw content (e.g., original videos, audio tracks, subtitles, or any material traceable to YouTube) is not allowed.
- Since X2Edit and Nano-consistent-150k do not provide explicit license statements: users must not attempt to recover or trace back the original media; users must not claim ownership of the upstream content.
- The dataset must not be used for privacy violations, re-identification, misinformation generation, or any unethical/illegal purposes.

Redistribution Requirements

- Any redistribution must include this license summary and all original attributions.
- Downstream users must be bound by the same terms.
- Redistributions must clearly acknowledge the upstream sources: UltraVideo, X2Edit, and Nano-consistent-150k.

Disclaimer

- The dataset is provided "AS IS", and may contain noise, bias, or residual sensitive content.
- Users assume all risks associated with the use of the dataset.
- The dataset creators and upstream sources bear no liability for any damages arising from its use.

By using VINS-120K, you agree to these terms.

Figure 15. Editing failure examples.

G. Limitations and Feature Work

Our method still has certain limitations in text editing [12]. As shown in Fig. 15, when the resolution is scaled beyond the effective range of the pretrained positional encodings, the model may struggle to accurately localize the editing region or reproduce the correct glyphs, even after LoRA fine-tuning. Moreover, the exponential growth of sequence length at ultra-high resolutions imposes unprecedented challenges for downstream inference and deployment. Future research should therefore focus on more efficient attention mechanisms and positional encoding designs that remain effective under extremely long sequences, enabling more precise and efficient high-resolution image editing.

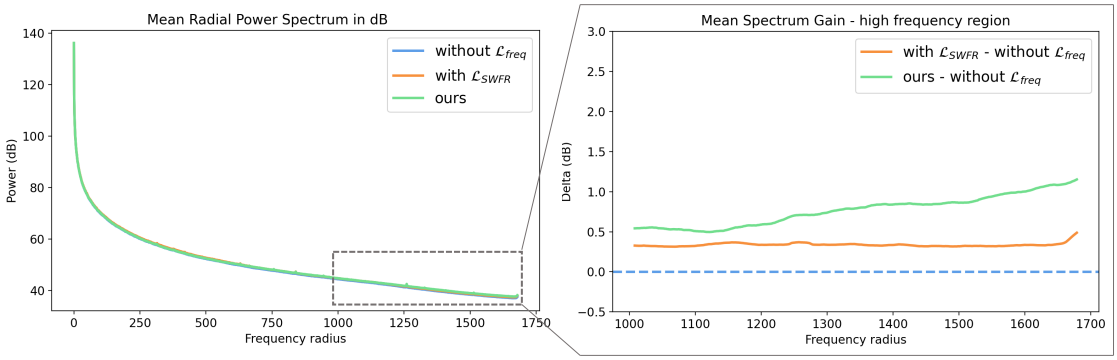


Figure 16. Spectral density analysis of the generated images.

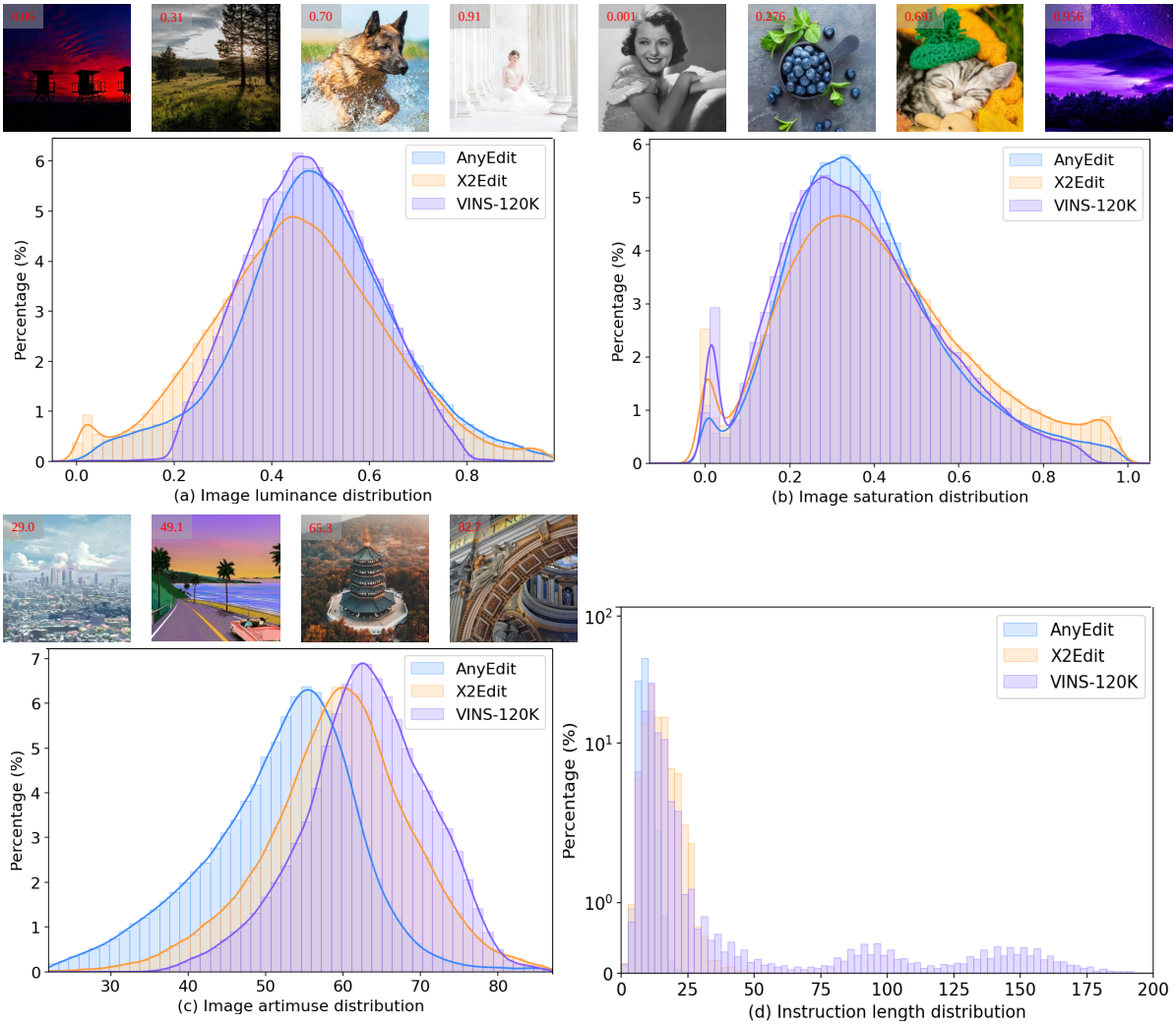
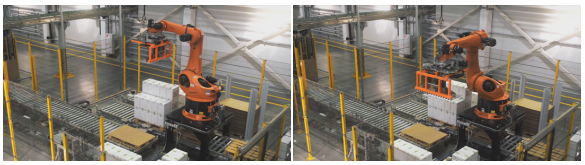


Figure 17. Comparison of image statistics between AnyEdit [52] and X2Edit [30] on luminance, saturation, and artimuse distributions.

Turn the hyena's head to the left to show its profile

Slightly zoom in, remove the tomato from the upper left, and add a pizza cutter pressing down into the center of the pizza

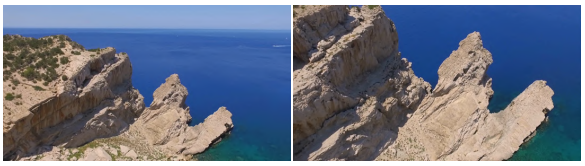


Slightly zoom out and lower the robotic arm to place the layer of boxes it is holding onto the stack on the pallet

Turn the bird's head out from behind the branch to reveal its face and beak

Remove the 'BROWN OWL' text box from the bottom left. Lift the owlet's head to face forward

Add a hand wearing a transparent glove from the top of the frame, placing a sesame seed bun onto the burger



Zoom in and tilt the frame down

Pan the frame to the left, positioning the plate on the right side of the image

Close the mouth of the man on the left and pucker his lips. Turn the head of the older man on the right to look down and close his eyes

Raise the groom's head to face the bride. Change the expressions of both the groom and the bride to smiles. Reposition the couple's hands so they are holding both of each other's hands.

Figure 18. Examples of Editing Instruction Annotation performed by our pipeline.

Change the woman's pose to putting on sunglasses

Replace background with a train station waiting hall, keep the person unchanged.

Replace dragons with a giant phoenix

Change the style of the picture to painting

Place the woman at a starry night camping site

Turn the wood furniture in the kitchen into a deep blue tone

Brighten the image and make it more vibrant

Remove the hair from the picture

Replacement of buildings with wooden structures

Turn the bird's head to the left

Add a cruise ship to the sea far away

Add a line to the right side of the picture: "It's the evening wind, it's the end of the water."

Figure 19. More high-quality examples from VINS-120K Dataset.

Figure 20. More qualitative comparison (1/2) between our method and recent baselines (Seedream4.0 [38], Kontext [25], Bagel [11], Step1X-Edit [28], ICEdit [56]).

Figure 21. More qualitative comparison (2/2) between our method and recent baselines (Seedream4.0 [38], Kontext [25], Bagel [11], Step1X-Edit [28], ICEdit [56]).