Are Image-to-Video Models Good Zero-Shot Image Editors?

Supplementary Material

1. Implementation Details

1.1. Algorithm Details

We summarize the inference pipeline of IF-Edit in Algorithm 1. The process integrates our three proposed components: CoT Prompt Enhancement, Temporal Latent Dropout, and Self-Consistent Post-Refinement.

As shown in the algorithm, we first enhance the static instruction into a temporal prompt. During the denoising loop (Steps 5–11), TLD is applied specifically at the expert-switch threshold T_{th} to sparsify temporal latents. Finally, the sharpest frame is identified via Laplacian scoring and refined using a "still-video" prompt to yield the final result.

Algorithm 1 Overall Inference Pipeline of **IF-Edit**

1: **Input:** image x_0, instruction c, model ϵ_θ, stride K, threshold T_{th}
2: **Output:** edited image $\hat{x}$
3: $c' \leftarrow \text{VLM-CoT}(x_0, c)$ */* CoT-enhanced prompt */*
4: Encode x_0 with zero-frame placeholders to obtain Y; sample noise latents z_T
5: **for** $t = T$ **to** 1 **do**
6: **if** $t = T_{\text{th}}$ **then**
7: */* One-shot temporal latent dropout */*
8: Retain every K-th frame latent and the last one: $z_t \leftarrow \mathcal{D}_K(z_t)$
9: **end if**
10: $z_{t-1} \leftarrow \epsilon_\theta(z_t, t, c', Y)$
11: **end for**
12: Decode to frames $\{x_i\}$; select sharpest $x^* = \arg\max_i s_i$
13: Refine x^* with still-video prompt; output sharpest refined frame $\hat{x}$
14: **return** $\hat{x}$

1.2. Method Details

Prompt Enhancement. We utilize Qwen3-VL-30B-A3B-Instruct [2] as our prompt enhancer, equipped with a customized system instruction designed to elicit temporal reasoning. Notably, this module is integrated via API calls; therefore, it imposes negligible latency overhead and requires no additional GPU memory consumption within the local inference pipeline, ensuring that the core computational resources are dedicated to generation.

Video Generation Backbone. For the generative backbone, we primarily adopt the Wan2.2-A14B I2V [16] model accelerated with Lightning-LoRA [9]. To balance generation quality and inference speed, we set the total number of denoising steps to 8. To verify the universality of our framework, we also experimented with other video generation backbones, including CogVideoX-5B [22] and Wan2.1 [16]. As shown in Fig. 1, our IF-Edit algorithm remains effective across these architectures, successfully accelerating the generation process while producing valid edits. However, since the final editing quality is intrinsically bounded by the capabilities of the base model, we observe that different backbones yield varying levels of visual fidelity. Given that Wan2.2 currently represents the strongest and most stable performance among open-source video models, we select it as our default backbone to rigorously evaluate the potential of video priors for image editing.

Hyperparameters. We generate 32-frame videos for each editing task. Following the official recommendations of Wan 2.2, we set the temporal dropout threshold to 0.9 (normalized time t=1 $\rightarrow$ 0), meaning the temporal latent dropout is active for the majority of the denoising process. The temporal stride is set to K=3. Additionally, we fix the guidance scale for both the high-noise and low-noise experts to 1.0 and set the flow shift value to 5.0 by default to ensure stable training-free adaptation.

Computational Efficiency. All experiments are conducted on a single NVIDIA H100 (80GB) GPU. Thanks to the efficient Lightning-LoRA acceleration and our proposed Temporal Latent Dropout strategy, each complete editing run takes approximately 12 seconds. This rapid inference speed highlights the efficiency of IF-Edit, making it highly suitable for practical, interactive image editing workflows where low latency is critical.

1.3. Detailed Benchmark Descriptions

TEdBench [8] serves as a primary benchmark for evaluating text-based real image editing, specifically targeting the challenging domain of **non-rigid deformations**. Unlike standard benchmarks focused on attribute changes, TEdBench consists of curated image-text pairs where central objects undergo significant geometric transformations and dynamic state changes (e.g., *"a bird spreading its wings"* or *"a door opening"*). This makes it an ideal testbed for assessing whether video priors can simulate physical motion better than static image editors. Following standard protocols [15], we quantify performance using **LPIPS** [26] for perceptual consistency, **CLIP-I** [14] for structure preservation, and **CLIP-T** [14] for semantic alignment.

ByteMorph [6] provides a large-scale evaluation of instruction-guided image editing with a focus on **complex non-rigid motions** and physical dynamics. The benchmark

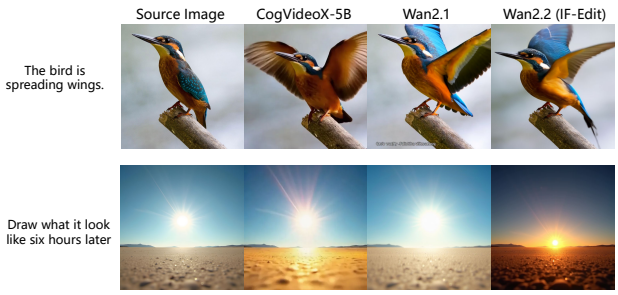


Figure 1. **Results using different base model.**

comprises **over 600 test samples** categorized into five distinct tasks: **Camera Zoom**, **Camera Move**, **Object Motion**, **Human Motion**, and **Interaction**. These categories are particularly relevant to investigating **World Model** capabilities, as they require the editor not just to modify pixels, but to simulate 3D spatial consistency (Camera Move/-Zoom) and plausible kinematic dynamics (Human/Object Motion). Success on ByteMorph indicates that the model possesses an internal understanding of physical laws and 3D geometry, effectively acting as a world simulator to predict future states. We utilize **Claude-3.7-Sonnet** [1] as the evaluator to assess semantic accuracy, visual realism, and the physical plausibility of the generated motions.

RISEBench [28] is designed to assess **reasoning-informed visual editing**, moving beyond simple descriptions to tasks requiring sophisticated cognitive processing. It evaluates models across four dimensions: **Temporal** (understanding sequences, e.g., *"melt the ice"*), **Causal** (predicting effects, e.g., *"after the glass falls"*), **Spatial** (manipulating 3D relationships and perspective), and **Logical** (abstract conceptual changes). The Temporal and Causal dimensions align closely with the intrinsic inductive biases of video diffusion models. Evaluation relies on **GPT-4.1** [12] for automated scoring, providing a rigorous measure of the model's ability to perform "System 2" visual reasoning.

ImgEdit [23] serves as a comprehensive unified benchmark for **general-purpose instruction editing**. It spans **nine diverse task categories**: Add, Adjust, Extract, Replace, Remove, Background, Style, Hybrid, and Action. While the benchmarks above focus on motion and reasoning, ImgEdit tests the model's versatility across the full spectrum of traditional editing operations, from localized object manipulation to global style transfer. This benchmark helps determine if repurposing video models for editing incurs a trade-off in general editing capabilities compared to specialized image editors. Performance is scored automatically across all categories using **GPT-4.1** [12].

1.4. Baseline Model Details

We compare IF-Edit against a comprehensive set of state-of-the-art instruction-based image editing models. Table 1 provides a detailed summary of the base architectures, parameter scales, and training dataset sizes for these baselines. The comparison spectrum ranges from specialized diffusion editors like InstructPix2Pix [4] and MagicBrush [25] to recent large-scale unified multimodal models such as Step1X-Edit [11] and Qwen-Image-Edit [19].

As highlighted in the table, the majority of competing methods are **training-based**, often requiring extensive supervised fine-tuning on large-scale paired datasets (reaching up to 20M samples) or relying on massive model parameters (exceeding 10B or even 20B) to align instructions with visual outputs. In contrast, our IF-Edit framework operates in a strictly **zero-shot, tuning-free** manner, leveraging the inherent world priors of video models without incurring the heavy cost of additional model training or data collection.

1.5. Chain-of-Thought Prompt Enhancement

Our method employs a Chain-of-Thought [18] reasoning approach to bridge the gap between concise, static editing instructions and the rich, dynamic world-simulation priors required by video generation models. This enhancement is performed using Qwen3-VL-30B-A3B-Instruct [2], which jointly analyzes the visual content of the input image and the semantic intent of the editing instruction.

Instead of simply expanding the text, the VLM is guided to generate a "cinematic reasoning" process: it first isolates the subject and background, explicitly plans the temporal trajectory of the requested change (defining the start, action, and end states), and enforces physical consistency for unedited regions. The resulting prompt serves as a temporally grounded screenplay that aligns the video model's generation with the user's goal. The full system instruction used to guide the VLM is provided in the last part.

1.6. Failure Cases and Limitations

While **IF-Edit** exhibits robust performance across reasoning and motion-centric benchmarks, it faces challenges with tasks requiring substantial structural or semantic deviations, particularly on the ImgEdit dataset [23]. Unlike benchmarks focused on physical dynamics, ImgEdit often demands aggressive attribute manipulation or object replacement. In these scenarios, video model tends to prioritize **visual consistency** over **edit strength**, a direct consequence of the video model's temporal smoothing bias. We categorize common failure modes into two types: *Minor Modification* and *Hallucinated Modification* (visualized in Fig. 7).

Minor Modification (Under-Editing). These failures occur when the temporal evolution of the requested edit exceeds the duration of the generated video clip. Since our method treats editing as a progressive physical change (e.g., an object slowly transforming), the intended modification may not fully materialize within the limited frame budget. This results in an output where the edit appears incomplete or only subtly applied.

Table 1. **Training details of baseline models.** We report base model, trainable parameters, and training dataset size for each training-based baseline method.

Method	Base Model	Trainable Params #	Training Dataset Size
InstructPix2Pix [4]	SD v1.5	0.9B	0.45M
MagicBrush [25]	SD v1.5	0.9B	0.47M
UltraEdit [27]	SD 3	2.5B	3M
ICEdit [13]	Flux.1-Fill-dev	0.2B	0.05M
Step1X-Edit [11]	Qwen2.5-VL-7B + DiT (train from scratch)	12.5B	20M
HiDream-E1 [5]	HiDream-I1	17B	N/A
BAGEL [7]	QwenLM-2.5	14B	N/A
UniWorld-V1 [10]	Qwen2.5-VL-7B + Flux	12B	2.7M
OmniGen [21]	Phi-3 + SDXL	3.8B	N/A
Kontext-Dev [3]	FLUX.1-dev	12B	N/A
Ovis-U1 [17]	Ovis	3.6B	N/A
OmniGen2 [20]	Qwen2.5-VL-3B + DiT	7B	N/A
AnyEdit [24]	SD v1.5	0.9B	2.5M
Qwen-Image-Edit [19]	Qwen-Image	20B	N/A

Hallucinated Modification. These artifacts stem from the CoT prompt enhancement process. While the VLM is designed to provide rich temporal context, it occasionally over-elaborates, hallucinating extrinsic details or narrative elements not present in the original instruction. The video generation model, remaining faithful to this enriched prompt, subsequently renders these distracting details, leading to semantic drift or unexpected object insertions.

Fundamentally, these limitations highlight the trade-off inherent in repurposing video generation models for image editing. The model's strong prior for maintaining temporal and spatial consistency—its primary strength for motion and reasoning tasks—creates resistance against precise, localized edits that require "breaking" the original scene structure. Consequently, IF-Edit is less effective for edits that demand significant structural violation or strictly localized changes compared to methods trained specifically for in-painting or style transfer.

1.7. Broader Impact

Our approach, IF-Edit, represents a significant step toward integrating world-simulation priors into image editing. By repurposing video diffusion models, we move beyond simple 2D pixel manipulation to edits that respect physical laws—maintaining consistent lighting, shadows, perspective, and causal dynamics. This capability has profound positive implications for creative industries, effectively democratizing high-fidelity visual effects. It empowers artists, educators, and designers to visualize complex scenarios and fluid transformations that were previously computationally prohibitive or required professional VFX expertise.

However, the ability to generate edits that are physically consistent and indistinguishable from reality raises significant ethical concerns regarding potential misuse. As the boundary between authentic photography and synthetic manipulation blurs, there is an inherent risk that such tools could be exploited to create hyper-realistic disinformation, fake news, or non-consensual content. The "world-simulation" nature of our results makes them harder for human observers to identify as synthetic compared to traditional editing artifacts. Consequently, we strongly advocate for the parallel development of robust detection mechanisms, such as invisible watermarking and content provenance protocols, to ensure these powerful generative capabilities are used responsibly and transparently.

2. Additional Results

In this section, we provide extensive qualitative results to further demonstrate the capabilities of IF-Edit across different editing paradigms. We specifically focus on visualizing the **temporal reasoning process** inherent in our video-based approach, showing how the model "thinks" through complex transformations via smooth, physically plausible trajectories.

2.1. Visualizing the Reasoning Process

A unique advantage of repurposing video diffusion models for image editing is the ability to observe the explicit temporal evolution of an edit. Unlike standard image-to-image translation, which often hallucinates immediate changes, IF-Edit simulates the transition from the source state to the target state.

In Figs. 2 and 5, we visualize the intermediate frames generated during the inference process. For instance, when instructed to change a person's pose to a **Namaste greeting,**

the model does not simply warp the body effectively into the final stance. Instead, guided by the enhanced temporal prompt, it generates the natural **biomechanical trajectory**: the arms rising deliberately from the sides and the palms gradually pressing together at chest level. This "Chain-of-Frames" reasoning ensures the final output respects anatomical constraints and maintains the subject's identity throughout the motion.

2.2. Qualitative Results on Benchmarks

Non-Rigid and Motion Editing (ByteMorph & TEdBench). Figures 2 to 4 display additional results on non-rigid deformation tasks. Whether handling scene dynamics or complex object articulations (e.g., jet bridge retracts), IF-Edit maintains high structural fidelity while executing dynamic changes that are difficult for static editors to model.

Reasoning-Informed Editing (RISEBench). Figure 5 showcases results on spatial, causal, and temporal reasoning tasks. The examples highlight the model's ability to interpret abstract instructions—such as "predict the aftermath of this event"—and translate them into visually coherent outcomes by simulating the causal chain of events.

General Instruction Editing (ImgEdit). Finally, in Figure 6, we provide more examples of general attribute editing. These results demonstrate that while our method is specialized for motion and reasoning, it remains competitive for general-purpose visual manipulation tasks.

CoT Enhancement System Prompt

You are a professional **video editing prompt engineer** and **chain-of-thought rewriter**. Your goal is to transform the user's editing instruction and the given input image into a **temporal reasoning process** — a concise, cinematic, and visually achievable description of how the content evolves over time through a short video until it fully matches the intended outcome.

Context:

1. The user provides an image (serving as the first frame of the video).
2. The user provides an editing instruction or question: {input_prompt}

Your Task: You must output a Chain-of-Thought reasoning that focuses on:

- How the visual content changes dynamically over time from the input image toward the final state implied by the instruction.
- Which elements undergo transformation (motion, deformation, environment, lighting, material).
- Which aspects remain unchanged to preserve continuity and identity.
- What the final visual state looks like once the transformation is complete.
- If the instruction requires reasoning or imagination (e.g., predicting, hypothesizing, or explaining), use visual common sense to infer what would happen naturally to the depicted objects, materials, or scenes, and describe that reasoning as a smooth temporal visual evolution.

Your description should read like a cinematic progression, balancing motion, stability, and visual realism. It must be suitable for guiding an image-to-video generation model.

Reasoning Requirements:

- Describe the sequence of visual changes — how the scene moves, transforms, or evolves over time.
- Always specify pose, expression, and appearance of the subject at key stages.
- If relevant, mention camera behavior: zoom, pan, tilt, dolly in/out, rotation, or focus shifts.
- Maintain visual coherence: lighting, color tone, rendering style (anime, cinematic, CG, poster, etc.) remain consistent.
- Preserve core identity: hairstyle, clothing, facial features, material texture, art style, and spatial composition must stay coherent across frames.
- Emphasize the final visual outcome, vividly describing what the last few frames should look like.
- Use natural, cinematic phrasing, without meta-instructions like "The user wants to…".

Task Type Guidance:

1. Temporal Action or Motion Editing

Describe how an action unfolds smoothly toward a defined final pose or state.

E.g., "The knight bends his knee slowly until he kneels completely, the movement steady and dignified."

2. Object or Element Replacement / Addition

Explain how new elements appear, form, or integrate over time.

E.g., "A glowing flower gradually materializes in her hand, pulsing with light until it fully blooms."

3. Scenario or Environmental Transition

Depict how the background and lighting shift naturally while keeping the subject's identity and art style stable.

E.g., "The bright street scene transitions into a rainy alley, reflections shimmering while the character's silhouette remains unchanged."

4. Reasoning-Based Visual Imagination (for inferential or hypothetical instructions)

If the input instruction is a question or reasoning task (e.g., "What will happen to the banana after one year?"), infer the most plausible physical or temporal transformation using visual commonsense reasoning. Describe this inferred process as a temporal visual evolution — *for example*: "The banana gradually darkens over time, its yellow skin developing brown spots that spread unevenly. The texture softens, small wrinkles appear, and by the final moment, it has turned fully black and shriveled, resting on the same surface under dimmer light." The reasoning must feel natural, physically consistent, and visually coherent with the given image.

Output Format:

Generate 100–130 words of continuous, cinematic reasoning text. Start directly with the evolving action, transformation, or reasoning-based change. Include both temporal progression and stability constraints, and conclude with a vivid description of the final frame.

Example Output 1 (Action Editing):

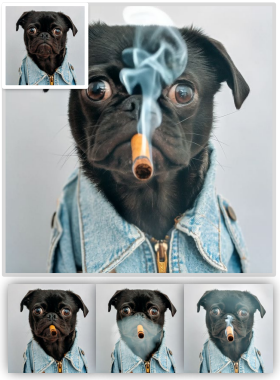
The knight's posture shifts as he lowers himself to one knee. His armor joints move with deliberate precision, the metal glinting under constant warm light. Dust stirs as his shield tilts forward to maintain balance. Throughout the sequence, his appearance, proportions, and medieval style remain consistent. By the final moment, he kneels solemnly, head bowed, radiating calm strength in a steady cinematic frame.

Example Output 2 (Reasoning-Based Visual Imagination):

The ripe yellow banana slowly changes over time. Subtle brown spots appear and spread across the skin as days pass. The surface darkens, its shape softening slightly while the peel begins to wrinkle. Lighting remains stable, keeping the same tabletop and perspective for temporal coherence. By the final frame, the banana has turned black and shrunken, resting quietly as a clear sign of decay and time's passage.

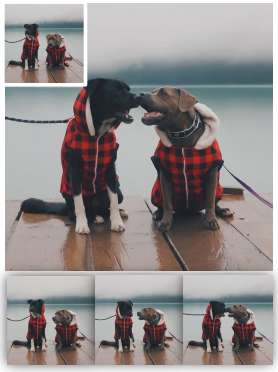
Original Prompt:
A photo of a cat wearing a hat.

Enhanced Prompt:
The fluffy white cat sits calmly on dry grass, its bright yellow eyes fixed forward, surrounded by autumnal foliage. \nOver time, a small, soft hat gently appears on its head\u2014first as a faint outline, then gradually taking shape. \nThe hat settles into place, matching the cat\u2019s serene posture, as if it were always meant to wear it. \nThe lighting and background remain unchanged, preserving the natural setting while the hat adds a whimsical touch. \nBy the final frame, the cat wears the hat with quiet dignity, completing a charming and unexpected transformation.



Original Prompt:
A dog smoking a cigar.

Enhanced Prompt:
The black pug, wearing a light denim jacket, sits calmly against a neutral gray background. \nA small, dark cigar appears in its mouth, positioned between its lips, as if it were just placed there. \nThe smoke begins to curl gently from the tip of the cigar, drifting upward in soft, wispy tendrils. \nOver time, the smoke thickens slightly, forming a faint haze around the pug\u2019s head, enhancing the surreal, humorous tone. \nBy the final frame, the cigar glows faintly at the tip, and the smoke continues to rise, completing the whimsical, cinematic moment.



Original Prompt:
Two dogs growling at each other.

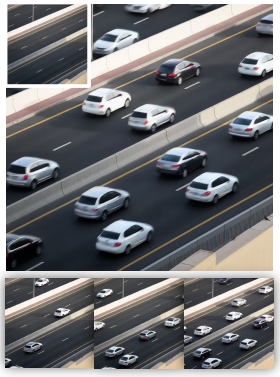
Enhanced Prompt:
"The two dogs, initially calm and sitting side by side on a wet wooden dock, begin to shift their attention toward one another. \nThe black dog turns its head slightly, ears flattening, lips curling into a low growl as its body tenses. \nThe brown dog mirrors the gesture\u2014its jaw tightens, teeth bared, and a deep growl rumbles from its chest. \nTheir eyes lock, pupils dilating, tails stiffening, and the air between them thickens with tension. \nThe misty lake and overcast sky remain unchanged, but the stillness of the scene gives way to a palpable, rising conflict."

Original Prompt:
A photo of a horse.

Enhanced Prompt:
"Over time, the zebra’s striped pattern begins to transform\u2014its bold black and white bands gradually soften, blending into a uniform, solid coat of rich chestnut brown. The distinct zebra stripes dissolve like a fading memory, replaced by the smooth, flowing texture of horse fur. The mane changes from short and stiff to a long, flowing mane typical of a horse. The head reshapes subtly longer muzzle, larger eyes, and more rounded ears\u2014aligning with equine features. By the final frame, the zebra has fully transformed into a horse, standing in the same pose, now a realistic, natural-looking equine."

Original Prompt:
A person in a greetting pose to Namaste hands.

Enhanced Prompt:
"The person stands still in a neutral stance, facing forward with their arms relaxed at their sides. \nSlowly, their hands rise to chest level, palms pressed together in a traditional *Namaste* gesture, fingers pointing upward and thumbs gently touching. \nThe motion is calm and deliberate, conveying respect and greeting. \nThe black t-shirt and background remain unchanged, maintaining visual consistency. \nBy the final frame, the hands hold the *Namaste* pose steadily, centered in the frame, embodying a moment of cultural acknowledgment and peaceful intent."



Original Prompt:
A busy congested street.

Enhanced Prompt:
"The empty highway slowly comes to life as time passes. \nAt first, a few cars appear in the distance, moving slowly along the lanes. \nMore vehicles enter the frame, filling the road with steady traffic. \nThe flow increases, turning the once-quiet street into a busy, congested thoroughfare. \nCars inch forward, brake lights glow, and the rhythm of urban movement takes over. \nBy the final frame, the highway is fully occupied, a dense stream of traffic under the same bright daylight and unchanged camera angle.”

Figure 2. **Qualitative Results and Reasoning Process.** We display source images alongside original instructions (grey) and our enhanced temporal prompts (orange). IF-Edit produces natural and coherent transformations, with edits that better respect physical laws, temporal progression, and spatial structure thanks to the video model’s world-consistent priors.

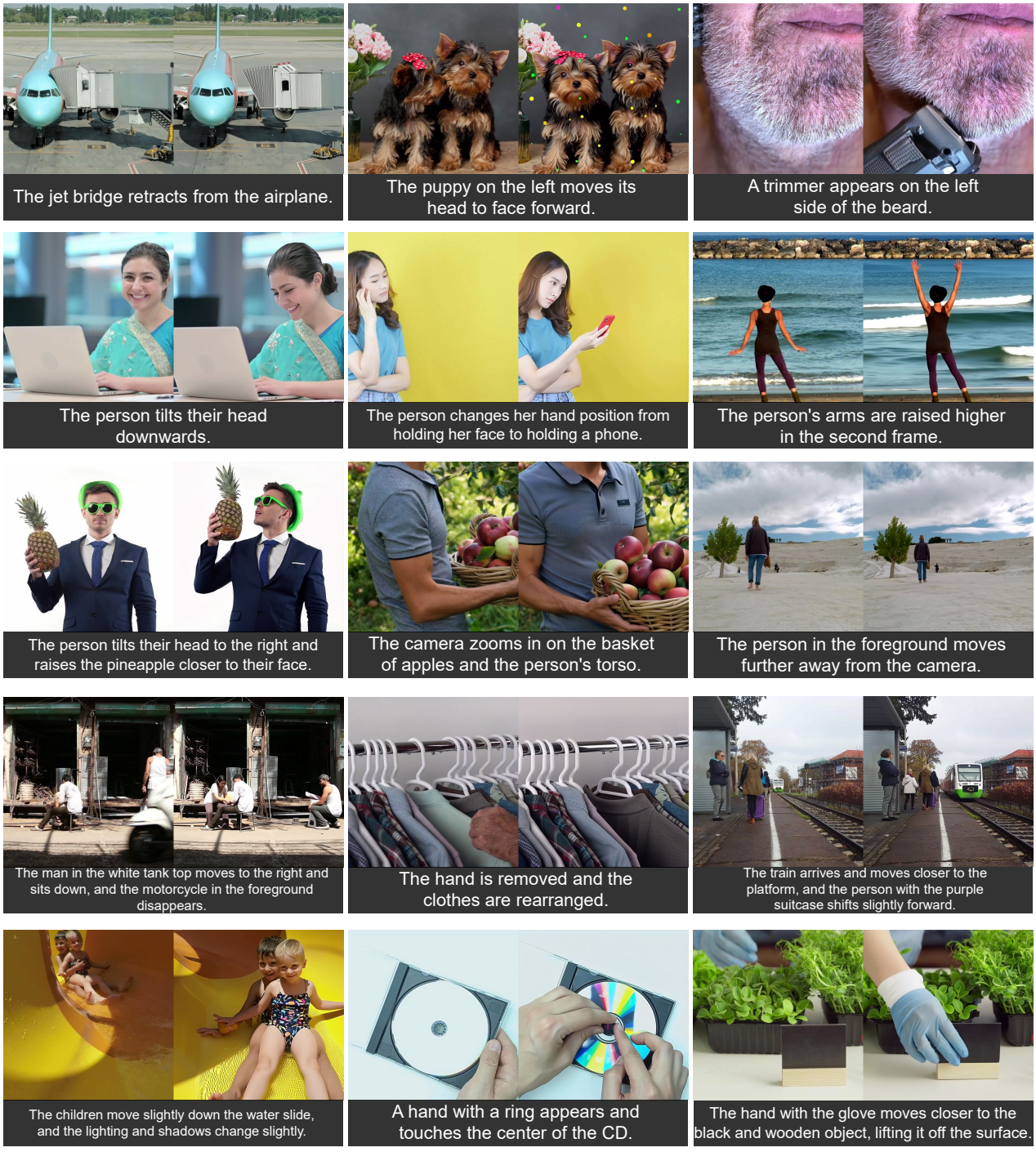


Figure 3. **Additional Qualitative Results on ByteMorph.** We present a diverse collection of non-rigid editing results, covering **Camera Move**, **Object Motion**, **Human Motion**, and **Interaction**. IF-Edit produces natural and coherent transformations, effectively handling complex kinematic changes (e.g., head tilts, arm raising) and perspective shifts (e.g., camera zoom, depth changes) by leveraging the video model's world-consistent priors.

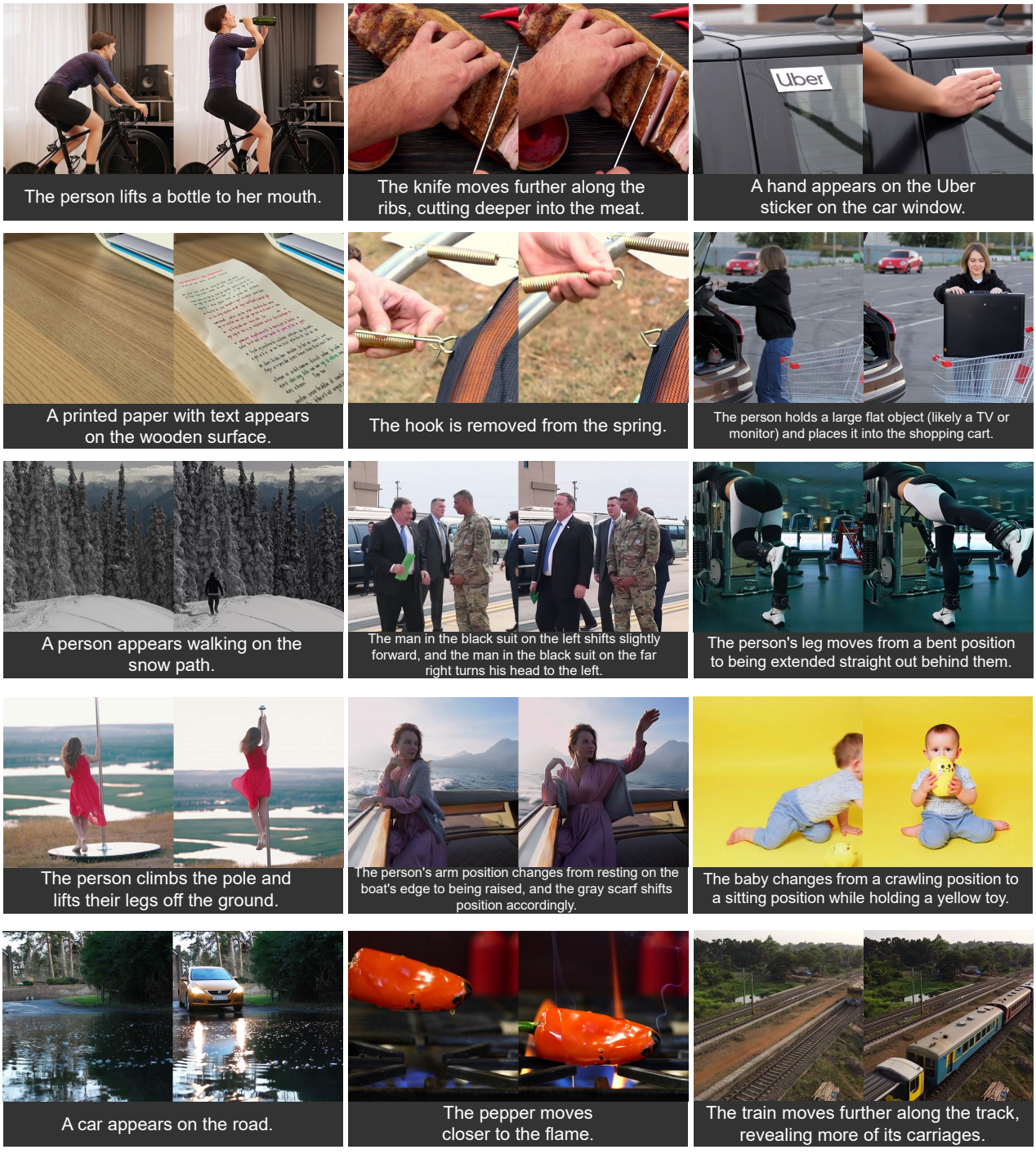


Figure 4. **Extended Qualitative Results on Complex Dynamics.** We showcase IF-Edit's capability in handling diverse physical interactions and object transformations. From fine-grained manipulations (e.g., cutting meat, removing a hook) to large-scale scene updates (e.g., a car appearing on a wet road with accurate reflections), our method maintains high visual fidelity and logical consistency, confirming the benefit of utilizing video priors for realistic image editing.

Figure 5. **Additional Qualitative Results on RISEBench.** We illustrate IF-Edit's capabilities across diverse reasoning dimensions, including **Temporal** (e.g., dough rising, water filling), **Causal** (e.g., paper shredding, X-ray scanning), and **Spatial** (e.g., assembling a snowman or hammer). The displayed **Enhanced Prompts** reveal how our CoT module translates abstract commands into detailed physical trajectories. IF-Edit produces natural and coherent transformations, with edits that better respect physical laws, temporal progression, and spatial structure thanks to the video model's world-consistent priors.

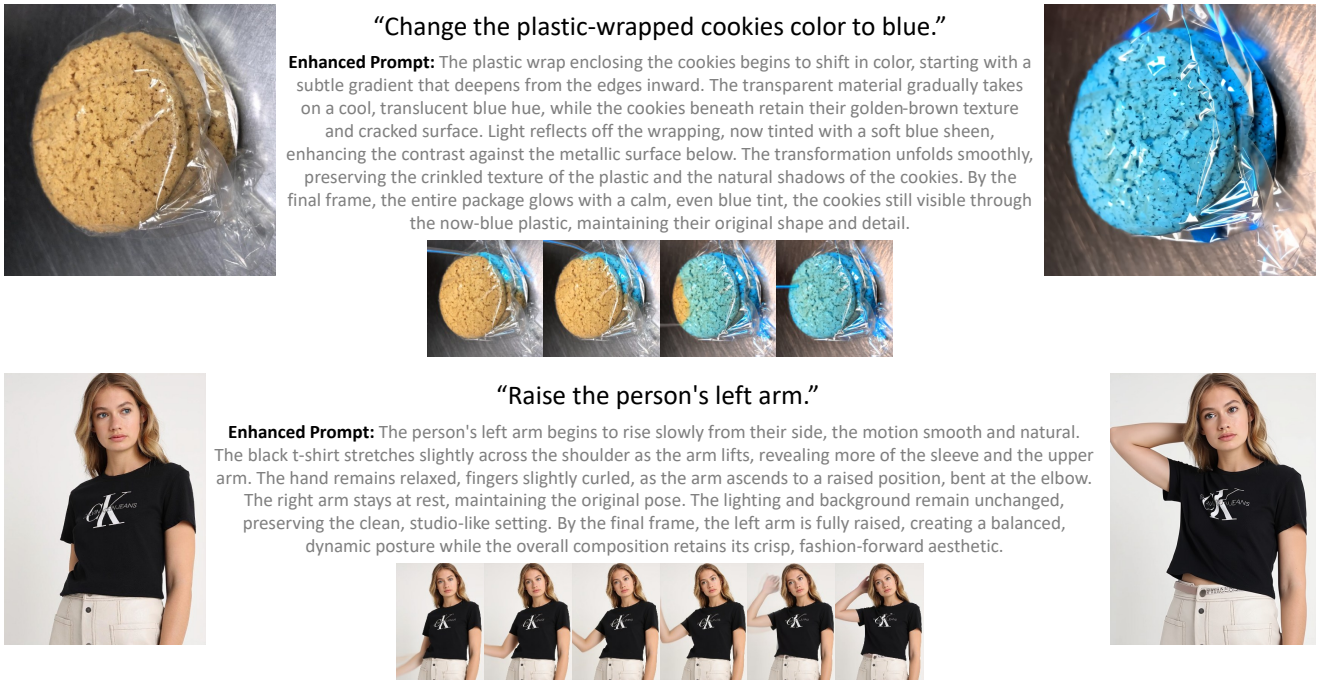


Figure 6. **Additional Qualitative Results on ImgEdit.** We demonstrate performance on general instruction editing, covering **Attribute Manipulation** (top) and **Action Editing** (bottom). Even for these standard tasks, IF-Edit leverages its temporal generation capability to ensure high-fidelity transitions—preserving the complex texture of the plastic wrap while shifting its color, and maintaining anatomical consistency and realistic cloth deformation during the arm's movement.

Figure 7. **Failed Cases on ImgEdit.** Compared to other benchmarks, ImgEdit often requires substantial structural changes or object replacements that conflict with the inherent temporal consistency priors of our zero-shot video model. Consequently, IF-Edit tends to prioritize visual consistency over edit strength, occasionally leading to reduced instruction adherence. We categorize these failures into *minor modification* (under-editing) and *hallucinated modification*.

References

[1] Anthropic. Introducing claude 3.7 sonnet: Our most intelligent model to date. *Blog Post*, 2025. 2

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv*, (arXiv:2511.21631), 2025. 1, 2

[3] Black Forest Labs. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*, 2025. 3

[4] 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, 2023. 2, 3

[5] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *arXiv preprint arXiv:2505.22705*, 2025. 3

[6] Di Chang, Mingdeng Cao, Yichun Shi, Bo Liu, Shengqu Cai, Shijie Zhou, Weilin Huang, Gordon Wetzstein, Mohammad Soleymani, and Peng Wang. Bytemorph: Benchmarking instruction-guided image editing with non-rigid motions. *arXiv preprint arXiv:2506.03107*, 2025. 1

[7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[8] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 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, 2023. 1

[9] LightX2V Contributors. Lightx2v: Light video generation inference framework. *GitHub repository*, 2025. https://huggingface.co/lightx2v/Wan2.2-Lightning/tree/main/Wan2.2-I2V-A14B-4steps-lora-rank64-Seko-V1. 1

[10] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation, 2025. 3

[11] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2, 3

[12] OpenAI. Introducing gpt-4.1 in the api, 2025. 2

[13] Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[15] Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaïd, and Ron Kimmel. Pathways on the image manifold: Image editing via video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2025. 1

[16] Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[17] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025. 3

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837, 2022. 2

[19] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 3

[20] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 3

[21] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 3

[22] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[23] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified

image editing dataset and benchmark. In *Advances in Neural Information Processing Systems*, 2025. 2

[24] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025. 3

[25] 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*, 36:31428–31449, 2023. 2, 3

[26] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1

[27] 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, 2024. 3

[28] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *Advances in Neural Information Processing Systems*, 2025. 2