# PROXYUP: TRAINING-FREE GENERATION OF COHERENT VIDEO INTERACTIONS FROM MOTION PROXIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Motion proxies can specify an action without depicting the objects, contacts, and scene needed to realize it naturally. We study how to turn such incomplete guidance into visually natural videos with coherent interactions while retaining the intended actions and key event structure. We propose **ProxyUp**, a training-free framework built on a pretrained video generator. Its inference design combines region-wise proxy initialization with full-latent refinement using predict-and-perturb updates, followed by ODE sampling. The mask selects initial guidance but does not restrict subsequent updates: the guided object and new content can adapt together at contacts and occlusions. The proxy thus guides generation without becoming a framewise target. On simulation and real-world proxies, ProxyUp achieves higher event-based motion and mechanics scores, imaging quality, and mean human ratings than the evaluated baselines.

## 1 INTRODUCTION

Video generative models (Hong et al., 2023; Yang et al., 2024; Wan et al., 2025; Kong et al., 2024; Blattmann et al., 2023; Brooks et al., 2024) can produce detailed scenes from text, yet realizing a particular action with coherent object interactions remains difficult. Motion proxies offer a way to specify that action, but do not necessarily depict the contacts, interacting objects, or visual detail needed for the target scene (Fig. 1).

We study *proxy-conditioned video generation*: turning incomplete motion guidance into visually natural videos with coherent interactions. The proxy and prompt provide compatible, partly overlapping guidance: the prompt describes the intended action and scene, while the proxy provides a concrete visual reference for its spatial and temporal realization. In bread cutting, the proxy provides cutting and slice-separation events, while the model must generate the hand, knife, contact, and occlusion that make the action visually coherent. Success requires preserving the action while constructing the missing interaction.

Three goals must be considered together: retaining the intended action and key event structure, generating coherent interactions, and producing natural imagery. A coarse proxy is not a framewise target. Local trajectories, contact motion, deformation, and rhythm may adapt to newly generated content, while the core action, direction, and event relationships should remain recognizable. Strong source preservation can retain crude appearance or restrict new interactions; excessive modification can lose the action. Sec. 3 illustrates this tension. Paired training data could address it, but aligned proxies and complete target scenes are difficult to collect.

We propose **ProxyUp**, combining region-wise latent noising (RLN) with full-latent refinement. RLN inverts the masked proxy latent and mixes it with fresh noise outside the guidance region. Our refinement stage, Stochastic Flow Relaxation (SFR), uses the basic Predict-and-Perturb (P&P) update of Self-Refining Video Sampling (Jang et al., 2026). The design uses the mask only to select initial guidance, then updates both regions before ODE sampling. This lets the guided object change together with new content at contacts and occlusions.

We evaluate action consistency, interactions, and visual quality using automatic metrics, human ratings, and ablations on simulation and real-world proxies. Human motion ratings jointly assess



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

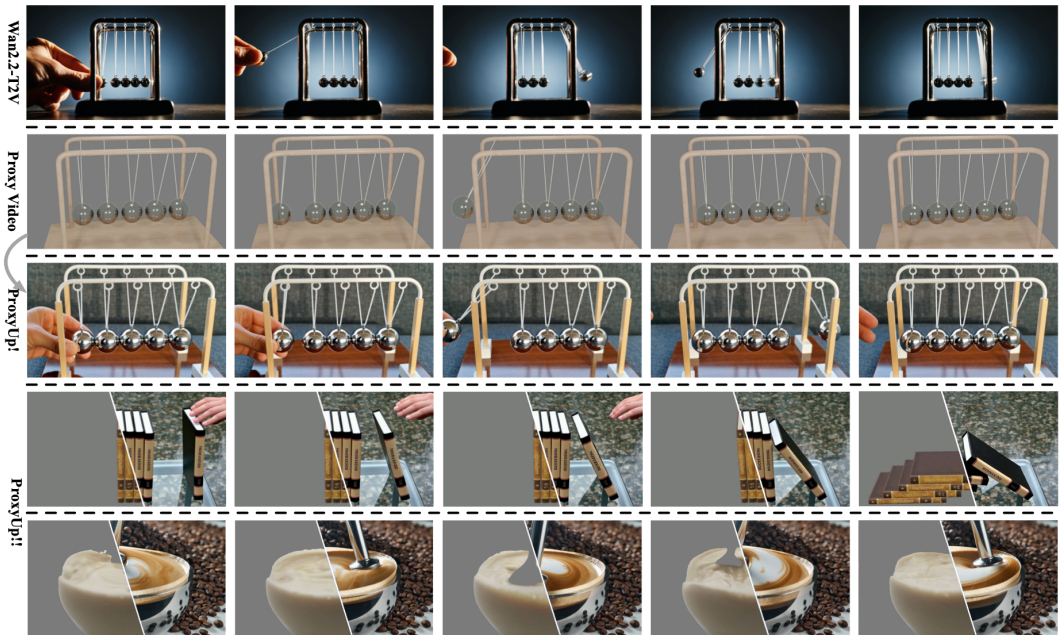



Figure 1: From motion proxies to complete interactions. ProxyUp uses selected proxy regions as visual guidance for generating new objects, contacts, and scene content. The text-only Wan2.2 comparison illustrates the difficulty of realizing the desired interaction from text alone.

coherence and proxy agreement; prompt-specific questions assess actions and interaction properties. The results support the combined design, while parameter sweeps expose trade-offs among these goals.

Our contributions are:

- We formulate generation from incomplete motion proxies as a joint task of action retention, coherent interaction, and visual synthesis.
- We combine local proxy initialization with full-latent P&P refinement, using the mask only at initialization so guided objects and new content can adapt together.
- We evaluate these goals using proxy-referenced human ratings, event and quality metrics, and ablations of the combined design.

## 2 Related Work

### 2.1 Controllable Video Generation

Prior work on controllable video generation introduces various conditioning signals beyond text and images, including human motion cues (2D/3D skeletons, SMPL parameters) (Ma et al., 2024; Zhong et al., 2024; Zhou et al., 2026), explicit motion representations (optical flow, point trajectories) (Wang et al., 2023; Nam et al., 2025), structural priors (edges, depth, normals) (Chen et al., 2024; Hu & Xu, 2023; Wang et al., 2023), semantic layouts (segmentation maps, mask tubes, bounding boxes) (Li et al., 2025b; Kim & Joo, 2026; Xi et al., 2026), and camera controls (extrinsics, intrinsics, view trajectories) (He et al., 2024; Bai et al., 2025a;b; Cao et al., 2025).

### 2.2 Motion Transfer

Motion transfer aims to synthesize a target video whose motion follows a driving source while its appearance or identity is changed. Early systems often rely on explicit intermediate representations, such as human pose or learned keypoints, to retarget motion from a source performer or object to a target subject (Chan et al., 2019; Siarohin et al., 2019). Recent diffusion-based methods

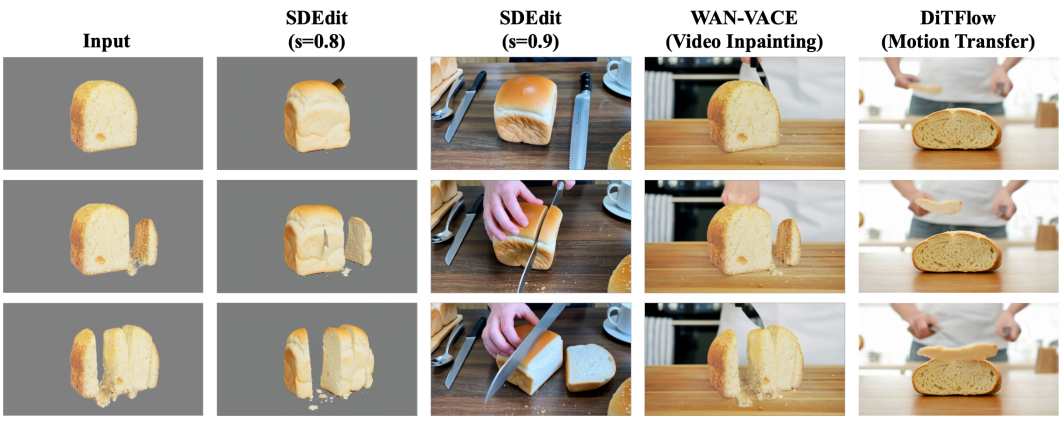



Figure 2: Generating a complete cutting interaction from a simplified motion proxy. Editing, inpainting, and motion transfer baselines must preserve the action while generating the person, knife, and contact specified by the prompt.

extend this idea to text-conditioned video synthesis: MotionDirector (Zhao et al., 2024) learns reusable motion concepts with decoupled spatial and temporal LoRAs, MotionFlow (Meral et al., 2026) transfers motion through attention-derived masks in pretrained video diffusion models, and DiTFlow (Pondaven et al., 2025) extracts attention motion flow from diffusion transformers for training-free motion transfer. Our setting additionally requires coherent interactions with newly generated content, allowing local motion to adapt while retaining the proxy's intended action.

### 2.3 Video Editing and Inpainting

Video editing methods adapt image diffusion models using inversion, feature injection, and temporal regularization (Ceylan et al., 2023; Qi et al., 2023; Geyer et al., 2023; Cong et al., 2023; Kara et al., 2024), or use native video models for joint spatiotemporal editing (Gu et al., 2023; Wang et al., 2025; Ouyang et al., 2024; Chen et al., 2025; Kim et al., 2025). Inpainting methods complete masked regions using visible context (Li et al., 2022; Liu et al., 2021; Zhou et al., 2023; Jiang et al., 2025). These methods offer useful starting points for proxy conditioning. Our setting permits changes to the proxy-guided object itself, including its appearance and occlusion by newly generated objects. We build on the inversion-and-noising approach used in diffusion-based editing (Meng et al., 2022).

### 2.4 Sampling Refinement

Self-Refining Video Sampling (Jang et al., 2026) introduces Predict-and-Perturb (P&P) updates at fixed noise levels, studies refinement early in sampling, and provides an uncertainty-aware variant. We adopt its basic update within a pipeline for incomplete motion proxies. Sec. 4.5 discusses this combination; Appendix B covers other refinement methods.

## 3 From Motion Proxies to Coherent Interactions

Editing, inpainting, and motion transfer methods (Meng et al., 2022; Kulikov et al., 2025; Li et al., 2025a; Jiang et al., 2025) provide starting points for this task. Preserving source motion alone, however, does not ensure that newly generated objects make coherent contact or that the resulting scene looks natural.

Fig. 2 illustrates the trade-off on bread cutting. With editing strength[1] $s = 0.8$ ($t = 0.922$), SDEdit retains the gray proxy background. At $s = 0.9$ ($t = 0.964$), it changes the appearance more but loses much of the cutting motion. VACE retains the foreground yet does not generate a coherent knife–bread interaction. DiTFlow also struggles to add the human–object interaction absent from the

[1]Strength $s$ is the fraction of sampling steps remaining at the chosen initialization time under the shifted scheduler.

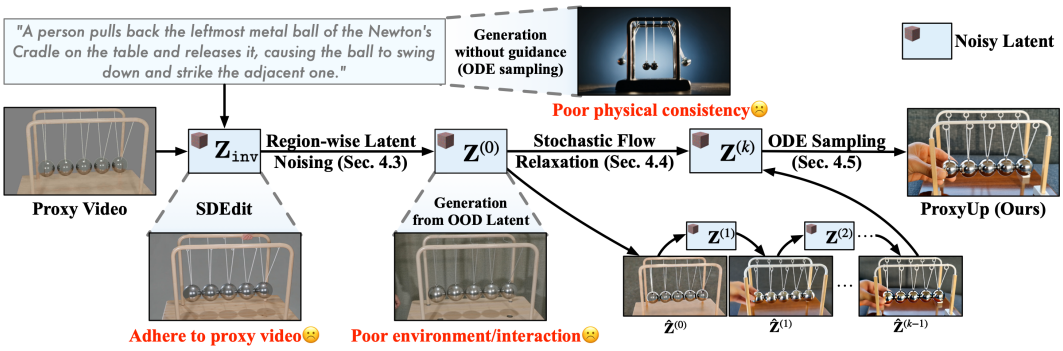



Figure 3: Overview of ProxyUp. Region-wise initialization supplies spatial and temporal guidance from the proxy. SFR jointly refines the guided object and new content, allowing contacts and occlusions across the initial mask boundary. ODE sampling and decoding produce the video.

proxy. These examples motivate an initialization that separates the guidance region from the rest of the scene, followed by joint refinement of both regions.

# 4 Method

Fig. 3 shows the three stages of ProxyUp: region-wise initialization, joint refinement, and ODE sampling.

## 4.1 Preliminaries: Rectified Flow

Rectified Flow (Liu et al., 2023) learns a probability flow ordinary differential equation (ODE) that transforms a noise distribution into a data distribution by modeling a time-dependent velocity field. Let $\mathbf{Z}_0 \sim p_{\text{data}}$ denote a clean sample and $\mathbf{Z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote Gaussian noise. Rectified Flow induces a linear interpolation path:

$$\mathbf{Z}_t = (1-t)\mathbf{Z}_0 + t\mathbf{Z}_1, \quad t \in [0,1]. \tag{1}$$

A neural network $\mathbf{v}_\theta(\mathbf{Z}_t, t, c)$ is optimized to predict the constant velocity of this trajectory, conditioned on a signal $c$. The training objective simplifies to a regression of the target displacement:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{\mathbf{Z}_0, \mathbf{Z}_1, t}\left[\|\mathbf{v}_\theta(\mathbf{Z}_t, t, c) - (\mathbf{Z}_1 - \mathbf{Z}_0)\|_2^2\right]. \tag{2}$$

At inference, generation is performed by solving the learned ODE from $t = 1$ to $t = 0$. Using a first-order Euler solver, the latent is updated as:

$$\mathbf{Z}_{t-\Delta t} = \mathbf{Z}_t - \Delta t \cdot \mathbf{v}_\theta(\mathbf{Z}_t, t, c), \tag{3}$$

where $\Delta t > 0$ is the step size.

## 4.2 Problem Formulation and Pipeline Overview

Given a proxy video $\mathcal{V}_p$, region mask $\mathbf{M}$, and prompt $c$, we seek a natural video $\mathcal{V}_g$ with coherent interactions and the intended action structure. Action information may appear in both inputs. Cutting must retain the specified cuts and slice separation; curtains must move in the specified opening or closing direction. Local trajectories, contact, and rhythm may adapt to new objects without changing these events. These are task objectives, not hard sampler constraints. The mask selects initial guidance, not pixels that must be copied.

RLN initializes from the masked video representation, carrying appearance, layout, and temporal cues while making room for new content. SFR and ODE sampling then update both regions under the prompt, allowing the guided object and newly generated objects to adapt together. Initialization strength and refinement count balance guidance retention with this freedom to change.

### 4.3 Region-wise Latent Noising for Motion Guidance

Let $\mathbf{Z}_0$ be the encoded proxy and $\mathbf{M}$ its binary guidance mask at latent resolution. We apply ODE inversion to $\mathbf{M} \odot \mathbf{Z}_0$ from $t = 0$ to $t_{\text{init}}$, obtaining $\mathbf{Z}_{inv}$. Masking before inversion reduces direct dependence on the proxy background, though VAE encoding and subsequent model updates can still mix information across regions.

Outside the guidance region, we sample Gaussian noise:

$$\epsilon_{bg} \sim \mathcal{N}\big(\mathbf{0}, ((1 - t_{\text{init}})^2 + t_{\text{init}}^2)\mathbf{I}\big). \tag{4}$$

This variance matches the second moment of the linear interpolation when clean latents are zero-mean with identity covariance and independent of the noise. It is an approximation to the noise scale, not a match to the full latent distribution. We form the initial latent as

$$\mathbf{Z}^{(0)} = \mathbf{M} \odot \mathbf{Z}_{inv} + (1 - \mathbf{M}) \odot \epsilon_{bg}. \tag{5}$$

The inverted region carries the proxy's joint spatiotemporal representation, including appearance and layout; it is not a motion-only encoding. Fresh noise allows new scene content, while initialization strength controls how much of the proxy representation is retained.

### 4.4 Stochastic Flow Relaxation (SFR)

RLN supplies regional guidance but does not construct the missing contacts or interaction partners. SFR is our full-latent refinement stage: it applies the basic P&P update (Jang et al., 2026) for $K$ iterations at $t_{\text{init}}$, starting from the composed latent:

$$\hat{\mathbf{Z}}_0^{(k)} = \mathbf{Z}^{(k-1)} - t_{\text{init}}\, \mathbf{v}_\theta(\mathbf{Z}^{(k-1)}, t_{\text{init}}, c), \tag{6}$$

$$\mathbf{Z}^{(k)} = (1 - t_{\text{init}})\hat{\mathbf{Z}}_0^{(k)} + t_{\text{init}}\epsilon^{(k)}, \qquad \epsilon^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

The first step estimates a clean latent; the second adds fresh noise at the same interpolation time. Neither region is clamped or remasked, allowing the guided object to participate in new contacts and occlusions. Finite $K$ limits refinement that can also weaken proxy cues.

**Proposition 1 (Conditional KL non-increase)** *At fixed $t \in (0, 1)$ and text condition $c$, let $q_k$ be the latent distribution after $k$ SFR updates. Assume the transition kernel admits a stationary distribution $\pi_*$ and $D_{KL}(q_0 \parallel \pi_*) < \infty$. Then, for $k \geq 1$,*

$$D_{KL}(q_k \parallel \pi_*) \leq D_{KL}(q_{k-1} \parallel \pi_*). \tag{8}$$

This general kernel property neither identifies $\pi_*$ with the model's noisy marginal nor guarantees motion preservation or a benefit of our composition; the proof is in Appendix A.

### 4.5 ODE Solver for Deterministic Generation

Starting from $\mathbf{Z}^{(K)}$ at $t_{\text{init}}$, we integrate toward $t = 0$. For a decreasing schedule $t_i > t_{i+1}$, let $\Delta t_i = t_i - t_{i+1} > 0$. The Euler update is

$$\mathbf{Z}_{t_{i+1}} = \mathbf{Z}_{t_i} - \Delta t_i\, \mathbf{v}_\theta(\mathbf{Z}_{t_i}, t_i, c). \tag{9}$$

The final latent is decoded to obtain $\mathcal{V}_g$. Sampling is deterministic given the refined latent and prompt; the RLN and SFR noise draws provide stochasticity. Algorithm 1 summarizes the procedure.

**Relation to prior methods.** Our contribution is a proxy-conditioned inference design with selective initialization and full-latent refinement. The P&P update and early refinement are inherited (Jang et al., 2026); the single-time block after RLN is an implementation choice, not an established improvement over other schedules. Like SDEdit (Meng et al., 2022), ProxyUp generates from a noisy input. RLN limits direct source initialization to selected regions, while full-latent P&P allows those regions to change with new content. The RLN-augmented SDEdit controls in Appendix C.4 differ from original SDEdit. Inpainting completes a specified region using visible context (Jiang et al., 2025). Here, the proxy-guided object may change appearance or be occluded by a new hand or tool. The mask selects initial guidance without imposing a subsequent preservation boundary. This combination introduces proxy cues while leaving interacting content free to adapt.

Table 1: Quantitative comparison on the real-video and physics-simulation subsets. Best values under the benchmark convention are bold; DD measures motion extent rather than motion fidelity. SC, BC, MS, DD, and AQ are rounded to four decimals.

| Subset | Method | SC | BC | MS | DD | AQ | IQ | MR | Mech. | Mat. |
|---|---|---|---|---|---|---|---|---|---|---|
| Real | Wan2.2 | 0.9334 | 0.9645 | 0.9886 | **0.6750** | 0.5896 | 66.39 | 0.70 | 0.91 | 0.80 |
| | VACE | 0.9280 | 0.9600 | 0.9933 | 0.4500 | 0.5402 | 63.08 | 0.65 | 0.87 | 0.88 |
| | DiTFlow | 0.9155 | 0.9461 | 0.9866 | 0.4500 | 0.5361 | 64.11 | 0.57 | 0.72 | 0.60 |
| | SDEdit | 0.8497 | 0.9119 | 0.9906 | 0.5750 | 0.4937 | 61.49 | 0.48 | 0.88 | 0.86 |
| | FlowDirector | 0.9388 | 0.9518 | 0.9881 | 0.5750 | 0.5865 | 66.40 | 0.65 | 0.87 | 0.73 |
| | ProxyUp | **0.9474** | **0.9699** | **0.9944** | **0.6750** | **0.6124** | **68.67** | **0.73** | **1.00** | **0.89** |
| Sim. | Wan2.2 | 0.9142 | 0.9562 | 0.9914 | 0.6579 | 0.5467 | 64.42 | 0.31 | 0.86 | **1.00** |
| | VACE | 0.8808 | 0.9426 | 0.9931 | 0.6316 | 0.4738 | 62.89 | 0.41 | 0.92 | **1.00** |
| | DiTFlow | 0.9215 | 0.9475 | 0.9873 | 0.5789 | 0.5066 | 57.18 | 0.27 | 0.75 | 0.50 |
| | SDEdit | 0.8201 | 0.9116 | 0.9895 | 0.5789 | 0.4803 | 67.99 | 0.28 | 0.96 | **1.00** |
| | FlowDirector | 0.9184 | 0.9509 | 0.9914 | 0.6316 | 0.5870 | 64.12 | 0.45 | 0.90 | 0.89 |
| | ProxyUp | **0.9223** | **0.9660** | **0.9940** | **0.7368** | **0.6114** | **68.88** | **0.52** | **1.00** | **1.00** |

# 5 Experiments

## 5.1 Experiment Settings

We build ProxyUp on Wan2.2 (Wan et al., 2025) and use a task-specific set of 76 proxy videos, evenly split between simulation and real-world recordings. SC, BC, MS, DD, AQ, and IQ measure visual and temporal quality following VBench (Huang et al., 2024; 2025). Automatic MR, Mech., and Mat. use prompt-specific VBench-2.0-style questions (Zheng et al., 2025) to assess specified actions and interaction properties. Human motion rationality (H-MR) jointly rates coherence and consistency with the reference proxy. Automatic MR measures event completion; H-MR is a composite judgment of motion quality and reference agreement. All outputs use $832 \times 480$ resolution. The default configuration uses 32 inversion steps, 15 SFR iterations, and 32 sampling steps, with $s = 0.8$ and $t_{\text{init}} = 0.922$. Appendix C gives the scoring protocol and guidance settings.

## 5.2 Quantitative Results

Table 1 compares ProxyUp on the real-video and physics-simulation subsets with Wan2.2 (Wan et al., 2025) for text-to-video generation, VACE (Jiang et al., 2025) for video inpainting, DiTFlow (Pondaven et al., 2025) for training-free motion transfer, FlowDirector (Li et al., 2025a) for training-free video editing, and SDEdit (Meng et al., 2022) adapted to video diffusion models. Baseline implementation details are provided in Appendix C.4.

Except for text-only Wan2.2, all methods receive the proxy and prompt. Backbones and inference schedules differ across systems. ProxyUp has the highest reported MR, Mech., and IQ on both subsets. Simulation MR is 0.52 versus the next-highest 0.45; real-video MR is 0.73 versus Wan2.2's 0.70. The latter scores measure prompt-specified event completion, not how closely motion matches the proxy. Mech.=1.00 reflects the finite question set and reported precision, not complete physical correctness.

In the 20-participant study, H-MR is 4.00 for ProxyUp, 2.63 for text-only Wan2.2, and 3.42 for VACE, the next-highest method (Appendix C.5). This favors ProxyUp on the combined coherence-and-reference question; separate gains in the two aspects cannot be recovered from this rating. ProxyUp also leads human VQ, Mech., and Mat., supporting the complete pipeline across the evaluated dimensions.

## 5.3 Qualitative Results

Fig. 4 compares an MPM bread-cutting proxy and a Blender curtain proxy. ProxyUp depicts the three specified cuts with knife–bread contact and deformation. It also generates a person pulling the curtain, an interaction missing from the shown DiTFlow, FlowDirector, and VACE outputs. These cases illustrate retaining the action while generating the objects and contacts needed to realize it.




Figure 4: Qualitative comparison on bread cutting and curtain pulling. Methods below the proxy row receive the same proxy video. The examples show differences in the generated objects, motion events, and interactions. *For more demo results, please refer to our supplementary material.*




Figure 5: Component ablations. Without region-wise latent noising (RLN), outputs retain the proxy appearance; without stochastic flow relaxation (SFR), visual quality and content richness degrade. The prompt matches Fig. 4. See Appendix D for quantitative results.

Table 2: Component ablations. Higher scores are better. Table 9 reports all metrics.

| | Real | | Simulation | |
|---|---|---|---|---|
| Variant | IQ | MR | IQ | MR |
| Without SFR | 64.73 | 0.52 | 65.30 | 0.31 |
| Without RLN | 66.93 | 0.70 | 67.16 | 0.50 |
| Full method | 68.67 | 0.73 | 68.88 | 0.52 |

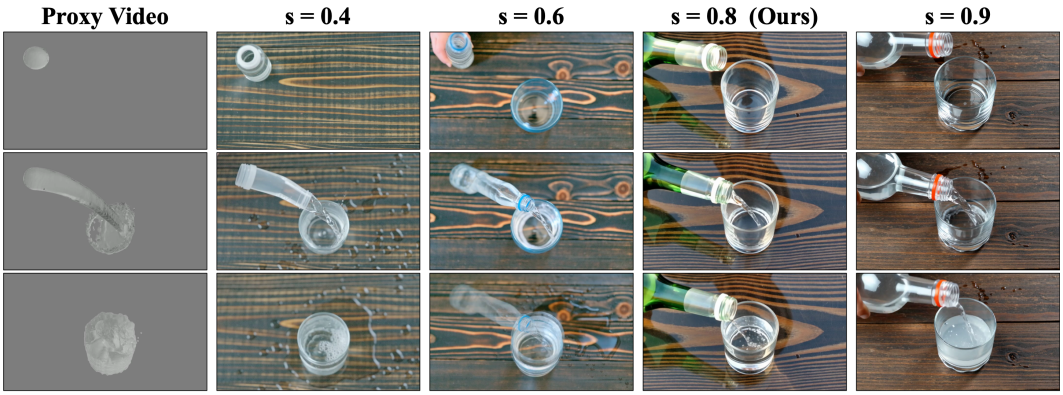
### 5.4 Ablations and Hyperparameters

Fig. 5 and Table 2 examine the two stages. Without RLN, the shown output retains the proxy's gray appearance; without SFR, it contains noise and blurred details. Adding RLN raises real-video IQ from 66.93 to 68.67 and MR from 0.70 to 0.73; simulation scores rise from 67.16 to 68.88 and from 0.50 to 0.52. SFR has the larger observed effect: removing it lowers MR by 0.21 on both subsets. RLN thus adds smaller event-score gains alongside better IQ and less proxy appearance in the example. These results support using both stages for the joint task, but do not establish an advantage over alternative P&P pipelines at matched cost. Appendix D gives all metrics.

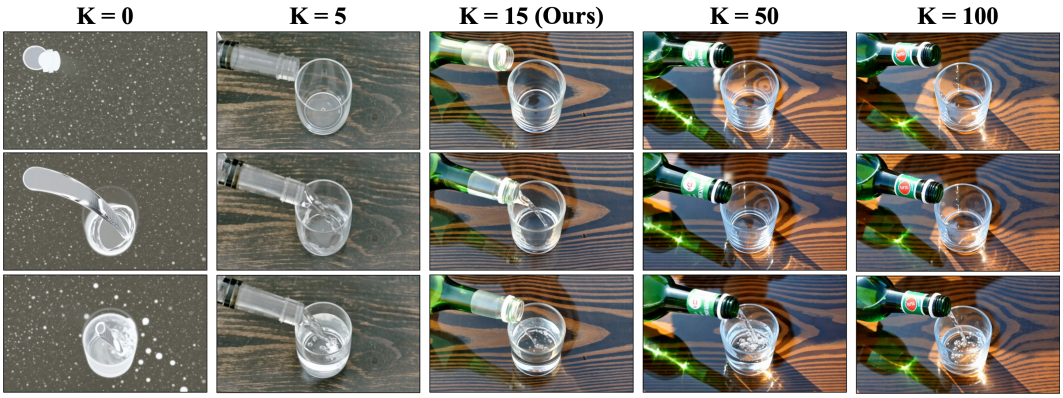
**Refinement iterations.** In Fig. 6 (bottom), refinement reduces visible noise but also changes object appearance. In the quantitative sweep (Table 7), IQ increases while MR and Mech. fluctuate. Real-video MR falls from 0.74 at $K = 5$ to 0.65 at $K = 50$, whereas simulation MR rises from 0.52 at $K = 15$ to 0.59 at $K = 50$. We use $K = 15$ as an empirical compromise, not a uniform optimum.

**Initialization strength.** Fig. 6 (top) shows increasing appearance change with strength. Raising $s$ from 0.4 to the default 0.8 lowers DPS from 0.615 to 0.443 and raises appearance difference (ADS) from 0.189 to 0.230 (Table 11), while giving the highest IQ and Mech. among tested strengths (Table 8). The default balances motion retention with quality and interaction scores. Lower DPS does not establish more natural motion, and higher ADS indicates more change, not better quality; both compare ProxyUp settings only.

**Mask sensitivity.** Appendix D tests dilation, erosion, pixel flips, missing mask segments, and temporal jitter. IQ varies by less than one point on each subset, and MR changes by at most 0.03 on real proxies and 0.04 on simulation proxies. Mech. and Mat. are more sensitive: real-video Mech. falls from 1.00 to 0.90 under pixel flips, while simulation Mat. falls from 1.00 to 0.88 when a mask segment is removed. Thus stable image quality does not imply that interactions and material behavior remain equally reliable. These results apply to the tested perturbation types and severities.




**(a) Hyperparameter analysis on the denoising strength.**




**(b) Hyperparameter analysis on the number of SFR steps K.**

Figure 6: Hyperparameter ablations. Top: increasing strength $s$ changes the proxy appearance and the generated environment. Bottom: increasing the number of SFR iterations $K$ reduces visible noise and adds detail, but can also change the guided object. Each column shows three frames from one output. The qualitative sweep includes $K = 100$, which is not evaluated in Table 7.

## 6 Limitations

ProxyUp relies on proxy cues and the pretrained generator; ambiguous motions can produce incorrect contacts (Appendix F). Local adaptation has no guarantee of physical correctness or proxy-motion correction. We report point estimates without significance tests. Our comparisons evaluate complete pipelines: they do not separate temporal proxy cues from appearance, layout, text, or refinement. Holding text and compute fixed while using static, shuffled, or alternative-motion proxies remains untested. H-MR combines coherence and reference agreement. Baseline DPS and matched-budget P&P comparisons also remain to be evaluated.

## 7 Conclusion

ProxyUp combines selective proxy initialization with existing P&P refinement to generate natural videos of complete interactions. Using the mask only at initialization lets guided objects and new content adapt together while aiming to retain intended actions. Event scores, quality metrics, and proxy-referenced human ratings support this pipeline; parameter sweeps expose trade-offs among these goals. Local adaptation is permitted, but proxy-motion correction remains unverified.

## AI USE STATEMENT

In this work, we used generative AI tools for manuscript restructuring and editing, interpretation of the reported results, review and revision of the mathematical exposition, and LaTeX preparation. We have not used generative AI tools for the creation of original research content that are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

ProxyUp can generate plausible videos of events that did not occur and may inherit biases from its pretrained generator. Generated outputs should be identified as synthetic and should not be treated as evidence of physical correctness. Appendix C describes the sources of the proxy videos; their reuse and redistribution remain subject to the source licenses. The human study reports aggregate ratings without participant identities. Its protocol and questions are described in Appendix C.5.

## REFERENCES


Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *ICLR*, 2025b.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *NeurIPS*, 2013. URL https://arxiv.org/abs/1305.6663.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 2024.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5933–5942, 2019.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. In *ICLR*, 2024.

Yiyang Chen, Xuanhua He, Xiujun Ma, and Yue Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *arXiv preprint arXiv:2509.17818*, 2025.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 2022. URL https://arxiv.org/abs/2112.05146.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *CVPR*, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.

Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. In *ICML*, 2026. URL https://arxiv.org/abs/2601.18577.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *CVPR*, 2024.

Min-Jung Kim, Dongjin Kim, Seokju Yun, and Jaegul Choo. Tv-live: Training-free, text-guided video editing via layer informed vitality exploitation. *arXiv preprint arXiv:2506.07205*, 2025.

Taeksoo Kim and Hanbyul Joo. Target-aware video diffusion models. In *ICLR*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *ICCV*, 2025.

Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv:2506.05046*, 2025a.

Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. In *WACV*, 2025b.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022.

Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14040–14049, 2021.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, 2024.

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

Tuna Han Salih Meral, Hidir Yesiltepe, Connor Dunlop, and Pinar Yanardag. Motionflow: Attention-driven motion transfer in video diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8043–8051, 2026.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025. URL https://arxiv.org/abs/2501.09038.

Hyelin Nam, Jaemin Kim, Dohun Lee, and Jong Chul Ye. Optical-flow guided prompt optimization for coherent video generation. In *CVPR*, 2025.

Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia*, 2024.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *CVPR*, 2025.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, 2023.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2023.

Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. In *CVPR*, 2025.

Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. Omnivdiff: Omni controllable video diffusion for generation and understanding. In *AAAI*, 2026.

Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2306.14878.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Yong Zhong, Min Zhao, Zebin You, Xiaofeng Yu, Changwang Zhang, and Chongxuan Li. Posecrafter: One-shot personalized video synthesis following flexible pose control. In *ECCV*, 2024.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

Zanwei Zhou, Taoran Yi, Jiemin Fang, Chen Yang, Lingxi Xie, Xinggang Wang, Wei Shen, and Qi Tian. Few-step flow for 3d generation via marginal-data transport distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13853–13861, 2026.

## A Proof of Proposition 1

We prove Proposition 1 under its stated stationarity and finite-KL assumptions. The text condition is fixed and omitted below. This is a general data-processing result for a stationary Markov kernel, not a property specific to RLN or evidence for the chosen refinement schedule.

**Step 1: Formulating the Markov Transition Kernel.** Define the flow-induced denoiser $D_\theta(\mathbf{Z}, t) = \mathbf{Z} - t\mathbf{v}_\theta(\mathbf{Z}, t)$, using the same guidance as in SFR. The one-step denoising prediction is:

$$\hat{\mathbf{Z}}_0^{(k)} = \mathbf{Z}^{(k-1)} - t \cdot \mathbf{v}_\theta(\mathbf{Z}^{(k-1)}, t) = D_\theta(\mathbf{Z}^{(k-1)}, t) \tag{10}$$

Following the re-noising step $\mathbf{Z}^{(k)} = (1-t) \cdot \hat{\mathbf{Z}}_0^{(k)} + t \cdot \epsilon^{(k)}$, the overall SFR loop forms a first-order time-homogeneous Markov Chain. The transition kernel $K(\mathbf{Z}' \mid \mathbf{Z})$ from step $(k-1)$ to step $k$ is an additive Gaussian kernel:

$$K(\mathbf{Z}' \mid \mathbf{Z}) = \mathcal{N}\left(\mathbf{Z}'; (1-t)D_\theta(\mathbf{Z}, t), t^2\mathbf{I}\right) \tag{11}$$

**Step 2: Non-increase of the KL Divergence.** At iteration $k$, the distribution $q_k(\mathbf{Z}')$ is updated via the transition kernel:

$$q_k(\mathbf{Z}') = \int K(\mathbf{Z}' \mid \mathbf{Z}) q_{k-1}(\mathbf{Z}) d\mathbf{Z} \tag{12}$$

According to the Data Processing Inequality (DPI) for Markov chains, applying the same transition kernel to two distributions does not increase their KL divergence. By the assumed stationarity of $\pi_*$ (i.e., $\pi_* = \int K\pi_*$), we have:

$$D_{KL}(q_k \parallel \pi_*) \leq D_{KL}(q_{k-1} \parallel \pi_*) \tag{13}$$

Since the sequence of KL divergence is non-negative and non-increasing, it must converge to a limit.

The existence of $\pi_*$ is an assumption, not a consequence of the positivity of the Gaussian transition density. The non-increase above implies

$$\lim_{K\to\infty} D_{KL}(q_K \parallel \pi_*) = L, \qquad L \geq 0, \tag{14}$$

The argument does not establish $L = 0$.

## B Algorithm

We summarize the inference procedure in Algorithm 1.

**Other refinement methods.** More broadly, predictor–corrector sampling (Song et al., 2021) combines reverse-time integration with score-based correction, and denoising-autoencoder sampling (Bengio et al., 2013) alternates reconstruction and corruption. CCDF (Chung et al., 2022) uses initialization and data-consistency steps for inverse problems; Restart (Xu et al., 2023) alternates noise injection and ODE integration over noise intervals.

**Algorithm 1** ProxyUp: Proxy-Guided Video Generation

**Require:** Flow matching model $\mathbf{v}_\theta$, proxy latent $\mathbf{Z}_0$, mask $\mathbf{M}$, condition $c$, noise level $t_{\text{init}}$, SFR iterations $K$, decreasing schedule $t_0 = t_{\text{init}} > \cdots > t_N = 0$.

*// Stage 1: Region-wise Latent Noising*

$\mathbf{Z}_{fg} \leftarrow \mathbf{M} \odot \mathbf{Z}_0$ ▷ Extract foreground

$\mathbf{Z}_{inv} \leftarrow \text{ODE_Invert}(\mathbf{Z}_{fg}, 0 \to t_{\text{init}})$ ▷ Invert to $t_{\text{init}}$

$\sigma_{bg}^2 \leftarrow (1 - t_{\text{init}})^2 + t_{\text{init}}^2$

$\epsilon_{bg} \sim \mathcal{N}(\mathbf{0}, \sigma_{bg}^2 \mathbf{I})$ ▷ Sample background noise

$\mathbf{Z}^{(0)} \leftarrow \mathbf{M} \odot \mathbf{Z}_{inv} + (1 - \mathbf{M}) \odot \epsilon_{bg}$ ▷ Construct mixed latent

*// Stage 2: Stochastic Flow Relaxation (SFR)*

**for** $k = 1, 2, \ldots, K$ **do**

  $\hat{\mathbf{Z}}_0^{(k)} \leftarrow \mathbf{Z}^{(k-1)} - t_{\text{init}} \cdot \mathbf{v}_\theta(\mathbf{Z}^{(k-1)}, t_{\text{init}}, c)$ ▷ Flow-induced denoising

  $\epsilon^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Sample noise

  $\mathbf{Z}^{(k)} \leftarrow (1 - t_{\text{init}}) \cdot \hat{\mathbf{Z}}_0^{(k)} + t_{\text{init}} \cdot \epsilon^{(k)}$ ▷ Re-noising

**end for**

*// Stage 3: Guided ODE Sampling*

$\mathbf{Z}_{t_{\text{init}}} \leftarrow \mathbf{Z}^{(K)}$ ▷ Initialize from relaxed latent

**for** $i = 0, 1, \ldots, N - 1$ **do**

  $\Delta t_i \leftarrow t_i - t_{i+1}$

  $\mathbf{Z}_{t_{i+1}} \leftarrow \mathbf{Z}_{t_i} - \Delta t_i \cdot \mathbf{v}_\theta(\mathbf{Z}_{t_i}, t_i, c)$ ▷ Deterministic Euler update

**end for**

$\mathcal{V}_g \leftarrow \text{Decode}(\mathbf{Z}_0)$ ▷ Decode to pixel space

**return** $\mathcal{V}_g$

# C Additional Experimental Details

## C.1 Implementation Details

We build ProxyUp primarily on top of Wan2.2 (Wan et al., 2025). Unless otherwise specified, all videos are generated at a resolution of $832 \times 480$. In all experiments, we use a shifted sampling scheduler with a shift value of 3.0. We use strength $s$ to denote the ratio of the remaining sampling steps at a specific time step $t$ to the total number of inference steps. We set the total number of inference steps to 40, the intermediate noise level $t_{\texttt{init}}$ to 0.922 ($s = 0.8$ in our sampling scheduler), and the number of stochastic flow relaxation steps $K$ to 15. Therefore, the numbers of inverse steps and sampling steps are both $40 \times 0.8 = 32$. For classifier-free guidance (CFG) (Ho & Salimans, 2022), we set it to 1 during inversion and 3 during stochastic flow relaxation, and follow Wan2.2 during ODE sampling by using guidance scales of 4 and 3 for the high-noise and low-noise stages, respectively. These operations add inference cost despite requiring no training. Because guidance differs across stages, step counts alone do not specify the number of model forward passes. All experiments are conducted on 8 NVIDIA A100 GPUs using Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023) with Ulysses sequence parallelism (Jacobs et al., 2023), both configured with size 8.

## C.2 Evaluation Datasets

Existing video generation, editing, and motion-transfer benchmarks are not directly aligned with our proxy-guided setting, since they typically provide only text prompts or source videos, but do not jointly provide proxy videos, foreground masks, and target prompts for dynamics-preserving regeneration. Therefore, following the common practice in controllable video generation and video editing, we construct a task-specific evaluation set for proxy-guided dynamic generation. The set contains 76 proxy videos, including 38 simulation-based videos and 38 real-world videos. The simulation subset is derived from 17 distinct scenarios, mainly generated in Blender, where foreground masks are rendered together with RGB frames. The real-world subset is collected from diverse daily scenes, and we use SAM3 (Carion et al., 2025) to extract foreground masks. These proxies cover a range of dynamic patterns, including rigid-body motion, soft-body deformation, object interaction, and fluid-like motion. We emphasize that this evaluation set is intended to diagnose whether a method

can follow externally provided proxy dynamics while synthesizing prompt-aligned content, rather than to serve as a large-scale general physics benchmark.

**Comparison with related evaluation datasets.** Motion-transfer evaluations such as DiTFlow (Pondaven et al., 2025) and MotionDirector (Zhao et al., 2024) assess reference motion under new prompts, while Physics-IQ (Motamed et al., 2025) assesses physical prediction through video continuation. Our evaluation pairs a proxy with a region mask, a target prompt, and scenario-specific criteria. It focuses on content generation from partial motion guidance, including interactions absent from the proxy. The set is a diagnostic collection rather than a comprehensive physics benchmark.

**Scenario coverage and provenance.** The 76 proxy videos cover 55 semantically distinct scenarios: 17 simulation scenarios and 38 real-world scenarios. Table 3 groups these scenarios by their dominant dynamics; the counts refer to scenarios, not clips. Several scenarios also involve human–object interaction or coupled phenomena, such as cutting with deformation and impact followed by splashing. Each evaluation sample combines a proxy video, a foreground mask, a target prompt, and scenario-specific evaluation criteria. The real-world proxies were collected from Pixabay, with foreground masks extracted using SAM3. Most simulation proxies were created using Blender; the bread-cutting example in Fig. 4 uses an MPM simulator. In addition to the main-paper examples, the supplementary video includes a book toppling onto a baseball, a five-domino chain, five books falling sequentially, and a cue ball scattering three colored balls, covering small objects and multiple interacting objects.

Table 3: Scenario-level taxonomy of the evaluation set. The 55 scenarios provide 76 proxy clips (38 simulation-based and 38 real-world). Each scenario is assigned to one dominant dynamics category.

| Dominant dynamics | Sim. | Real | Total | Examples |
|---|---|---|---|---|
| Rigid/contact/collision/gravity | 8 | 8 | 16 | Cradle, collisions, toppling |
| Deformation/cutting/fracture | 4 | 17 | 21 | Curtains, slicing, shattering |
| Fluid/fluid–object interaction | 4 | 11 | 15 | Pouring, splashing, buoyancy |
| Elastic/oscillatory | 1 | 0 | 1 | Spring bouncing |
| Granular/phase change | 0 | 2 | 2 | Sand, melting snow |
| Total | 17 | 38 | 55 | |

### C.3 Evaluation Metrics

We assess action retention, interaction coherence, and visual quality with complementary measures. H-MR combines motion coherence and agreement with the proxy in one human rating; automatic MR, Mech., and Mat. assess actions and interaction properties; IQ, AQ, and human VQ assess visual quality. These measures overlap rather than isolate three independent factors. For automatic evaluation, target prompts are written from the proxy videos to describe their motion events together with the desired scene and interactions. Prompt-specific VLM questions then assess whether generated videos realize these actions (MR), interactions (Mech.), and material behavior (Mat.). These questions measure task completion under the supplied conditions, rather than directly comparing the generated video with the proxy. H-MR uses the input proxy as its reference but does not separately score coherence and reference agreement (Appendix C.5). Neither measure quantifies framewise trajectory, velocity, or phase error. The DPS strength sweep provides a separate within-method analysis of motion preservation.

**Visual and temporal quality metrics from VBench.** The six metrics SC, BC, MS, DD, AQ, and IQ in Table 1 are all adopted from VBench (Huang et al., 2024). Subject Consistency (SC) measures whether the subject's identity and appearance remain consistent across frames. Background Consistency (BC) measures the temporal consistency of the background scene. Motion Smoothness (MS) assesses whether motion evolves smoothly across frames, without abrupt or unnatural transitions. Dynamic Degree (DD) measures the extent of motion in the generated video. Aesthetic Quality (AQ) evaluates the visual appeal of the frames, including composition and color. Imaging Quality (IQ) evaluates technical image quality, including sharpness and the absence of blur, noise, and exposure artifacts. The table follows the benchmark convention of reporting higher values as better. DD measures motion extent, however, and a larger value need not indicate closer agreement with a

particular proxy. MR, Mech., and Mat. complement these metrics with event- and property-based questions.

**Scoring model and criterion construction.** For MR, Mech., and Mat., we follow the VBench-2.0 scoring protocol using Qwen2.5-VL-7B-Instruct. Gemini 3.0 Pro generated candidate criteria, which were manually inspected and corrected. Criteria were fixed independently of the generated outputs and shared across methods. The supplementary `evaluation_protocols` directory provides the simulation and real-video JSON files, with `prompt_en`, `dimension`, and `auxiliary_info` specifying each prompt and its questions. The proxy enters this automatic evaluation through the prompt-derived criteria; the scores are not pairwise video-similarity measurements. The annotation collection contains 166 rubrics and 427 entries (Table 4), including motion-order and dynamic-spatial-relation annotations in addition to MR, Mech., and Mat. These totals describe the annotation collection, not the number of scored videos or the denominator for each metric.

Table 4: Size of the annotation collection, including MR, Mech., Mat., motion-order, and dynamic-spatial-relation annotations. Counts are not per-metric evaluation denominators.

| Evaluation material | Simulation | Real-world | Total |
|---|---|---|---|
| Prompt-specific rubrics | 61 | 105 | 166 |
| Annotation entries | 159 | 268 | 427 |

**Concrete evaluation examples.** Table 5 lists representative questions for four scenarios. The bread-cutting questions specify both the action and its temporal structure, including the initial state, removal of each slice, and number of repetitions. The other examples address contact and momentum transfer, overflow behavior, and impact followed by fracture and splashing. Thus, the questions test specific proxy-derived events rather than generic video plausibility alone. They also assess desired interactions with newly generated content, such as a knife contacting the bread. Because the same events are described in the prompt, these scores alone cannot attribute successful motion to proxy conditioning rather than text. High scores, including a rounded Mech. score of 1.00, do not imply complete physical correctness.

Table 5: Representative evaluation criteria used for different scenarios and dimensions. The same questions are applied across methods for the corresponding target prompt.

| Scenario | Dimension | Representative evaluation criteria |
|---|---|---|
| Cutting bread | Motion Rationality | Does the person perform a downward slicing motion? Is the bread initially intact? Is each slice moved away? Is the sequence repeated three times? |
| Newton's cradle | Mechanics | Is the leftmost ball pulled back? Does it strike the adjacent ball after release? Is momentum transferred through the row? |
| Water overflow | Material / Mechanics | Does the water level increase during pouring? Does it eventually overflow? Does the overflow flow down the cup as liquid? |
| Falling glass bottle | Mechanics / Material | Does the bottle hit the floor and shatter? Does the liquid splash out after impact? |

### C.4 Baseline Implementations

For all baselines, we use their official implementations and run inference with the default hyperparameter configurations provided by the corresponding codebases. Specifically, Wan2.2 uses the T2V-A14B model (Wan et al., 2025); VACE uses its released 14B model trained for video inpainting (Jiang et al., 2025); DiTFlow is implemented on CogVideoX-5B (Yang et al., 2024; Pondaven et al., 2025); and FlowDirector is implemented on Wan2.1-T2V-14B (Li et al., 2025a; Wan et al., 2025). SDEdit is adapted to the same Wan2.2 backbone as ProxyUp. Numbers of their inference steps are $40 \sim 50$.

The main comparison uses each method's default configuration and does not control for backbone or inference cost. We also provide qualitative SDEdit results with additional steps and repeated sampling (Fig. 8), ProxyUp on Wan2.1 (Fig. 9), and DiTFlow on Wan2.2 (Fig. 10). These examples illustrate behavior under selected settings; they do not establish quantitative gains under matched settings.

The component ablations test removals from ProxyUp, but do not compare alternative initialization and refinement strategies at matched inference cost. Distinguishing the combination from other uses of P&P would require fixing the backbone and model-evaluation budget while comparing text-only generation with P&P, full-video initialization with P&P, and RLN with extra ODE steps or the original P&P schedule. These comparisons are not reported here; differences in stage-wise guidance also prevent raw step counts from establishing equal inference cost.

### C.5 User Study

We conducted a user study with 20 participants. The generated videos were anonymized and randomized, and participants rated them on a five-point scale along the same four dimensions: Visual Quality (VQ), Motion Rationality (H-MR), Mechanics (Mech.), and Material (Mat.). Participants assessed each dimension using the following questions: (1) VQ: "Are the visuals clear and natural?"; (2) H-MR: "Is the motion coherent and consistent with the motion pattern in the reference video?"; (3) Mech.: "Do the object interactions obey physical laws?"; and (4) Mat.: "Is the material rendering realistic and consistent across space and time?" The reference video is the input proxy. We label this existing human rating H-MR to distinguish it from automatic MR. The question jointly assesses coherence and reference agreement; the responses do not provide separate scores for these aspects. Table 6 summarizes the human ratings, with values reported as means $\pm$ standard deviations (SD). The overall score is the average of the four dimensional means.

Table 6: User study with 20 participants and five-point ratings. H-MR jointly rates coherence and proxy agreement. Values are means $\pm$ standard deviations (SD); Overall averages the four dimensional means. The highest means are bold.

| Method | VQ | H-MR | Mech. | Mat. | Overall |
|---|---|---|---|---|---|
| ProxyUp | $\mathbf{4.43 \pm 0.69}$ | $\mathbf{4.00 \pm 1.04}$ | $\mathbf{3.98 \pm 1.06}$ | $\mathbf{4.12 \pm 0.96}$ | **4.13** |
| Wan2.2 | $3.79 \pm 0.84$ | $2.63 \pm 1.21$ | $2.73 \pm 1.20$ | $3.20 \pm 1.04$ | 3.09 |
| VACE | $3.42 \pm 1.05$ | $3.42 \pm 1.20$ | $3.27 \pm 1.24$ | $3.35 \pm 1.14$ | 3.36 |
| SDEdit | $2.58 \pm 1.23$ | $3.04 \pm 1.25$ | $3.01 \pm 1.30$ | $2.76 \pm 1.29$ | 2.85 |
| DiTFlow | $3.09 \pm 1.07$ | $2.05 \pm 1.03$ | $2.06 \pm 1.04$ | $2.66 \pm 1.04$ | 2.46 |
| FlowDirector | $3.56 \pm 0.92$ | $2.38 \pm 1.17$ | $2.70 \pm 1.19$ | $3.16 \pm 0.97$ | 2.95 |

ProxyUp has the highest mean rating on all four dimensions. Its mean human Mechanics score is 3.98/5, which also shows that a high automatic score need not imply perfect perceived physical plausibility. H-MR is 4.00 for ProxyUp, 2.63 for text-only Wan2.2, and 3.42 for VACE, the next-highest method. This compares complete pipelines on the composite question, rather than isolating the use of temporal proxy information. It does not measure trajectory or timing error. Agreement between human and automatic ratings has not been measured.

## D Additional Quantitative Ablations

We report quantitative ablations on both real-video and physics-simulation subsets. Unless varied in the corresponding experiment, the default settings are $K = 15$ and $s = 0.8$. IQ, MR, Mech., and Mat. follow the evaluation protocol described above.

### D.1 SFR Iterations

Table 7 varies the number of SFR iterations. All tested positive values improve IQ over $K = 0$, while MR and Mech. vary non-monotonically. The default $K = 15$ attains the highest or tied-highest Mech. on both splits, but larger $K$ improves IQ and some other settings yield higher MR. It is an empirical default, not a uniform optimum.

Table 7: Effect of the number of SFR iterations $K$. Higher scores are better.

| Split | Metric | $K = 0$ | $K = 5$ | $K = 10$ | $K = 15$ | $K = 30$ | $K = 50$ |
|---|---|---|---|---|---|---|---|
| Real | IQ | 64.73 | 67.10 | 67.47 | 68.67 | 69.23 | 69.67 |
| | MR | 0.52 | 0.74 | 0.52 | 0.73 | 0.74 | 0.65 |
| | Mech. | 0.93 | 0.93 | 0.89 | 1.00 | 1.00 | 0.91 |
| | Mat. | 0.89 | 0.80 | 0.90 | 0.89 | 0.89 | 0.80 |
| Sim. | IQ | 65.30 | 67.07 | 67.97 | 68.88 | 69.64 | 69.94 |
| | MR | 0.31 | 0.59 | 0.52 | 0.52 | 0.55 | 0.59 |
| | Mech. | 0.89 | 0.96 | 0.91 | 1.00 | 0.96 | 0.96 |
| | Mat. | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

### D.2 Noising Strength

Table 8 varies the noising strength $s$. Among the evaluated settings, $s = 0.8$ achieves the highest IQ and Mech. on both splits, together with the highest or tied-highest MR.

Table 8: Effect of noising strength $s$. Higher scores are better.

| Split | Metric | $s = 0.4$ | $s = 0.6$ | $s = 0.8$ | $s = 0.9$ |
|---|---|---|---|---|---|
| Real | IQ | 63.50 | 65.08 | 68.67 | 68.13 |
| | MR | 0.70 | 0.70 | 0.73 | 0.70 |
| | Mech. | 0.88 | 0.89 | 1.00 | 0.91 |
| | Mat. | 0.89 | 0.88 | 0.89 | 0.82 |
| Sim. | IQ | 65.42 | 66.51 | 68.88 | 67.42 |
| | MR | 0.52 | 0.48 | 0.52 | 0.50 |
| | Mech. | 0.95 | 0.95 | 1.00 | 0.92 |
| | Mat. | 0.92 | 1.00 | 1.00 | 1.00 |

### D.3 Component Ablations

Table 9 complements Fig. 5. Removing SFR has the larger effect on IQ and MR; RLN adds smaller gains in both metrics. The full method also improves simulation Mech. from 0.89 without SFR and 0.92 without RLN to 1.00, while simulation Mat. is unchanged. The stages therefore have different observed effects within this pipeline. The variant without SFR corresponds to $K = 0$ in Table 7.

Table 9: Ablation of SFR and RLN. Higher scores are better.

| Split | Metric | w/o SFR | w/o RLN | ProxyUp |
|---|---|---|---|---|
| Real | IQ | 64.73 | 66.93 | 68.67 |
| | MR | 0.52 | 0.70 | 0.73 |
| | Mech. | 0.93 | 1.00 | 1.00 |
| | Mat. | 0.89 | 0.88 | 0.89 |
| Sim. | IQ | 65.30 | 67.16 | 68.88 |
| | MR | 0.31 | 0.50 | 0.52 |
| | Mech. | 0.89 | 0.92 | 1.00 |
| | Mat. | 1.00 | 1.00 | 1.00 |

### D.4 Robustness to Foreground-Mask Perturbations

Using Wan2.2-T2V-A14B at $480 \times 832$ resolution, we binarize masks at a threshold of 128 and independently apply $31 \times 31$ dilation or erosion, 10% pixel flips, temporal jitter of up to $\pm 2$ frames, or removal of a contiguous 20% segment. All generation settings are fixed, with deterministic paired seeds. Table 10 reports the results. IQ and MR remain relatively stable, while Mech. and Mat. decrease under some perturbations. These results describe sensitivity to the tested perturbations and

severities. SFR updates both regions, which may help accommodate mask errors, but the experiment does not isolate that explanation. Resizing masks to latent resolution can also change their boundaries.

Table 10: Sensitivity to independent mask perturbations. Noise denotes pixel flips; Missing denotes contiguous segment removal. Higher scores are better.

| Split | Metric | Clean | Dilate | Erode | Noise | Missing | Jitter |
|---|---|---|---|---|---|---|---|
| Real | IQ | 68.67 | 68.65 | 68.30 | 68.14 | 68.38 | 68.37 |
| | MR | 0.73 | 0.73 | 0.70 | 0.70 | 0.70 | 0.70 |
| | Mech. | 1.00 | 0.93 | 0.93 | 0.90 | 0.93 | 0.93 |
| | Mat. | 0.89 | 0.90 | 0.82 | 0.89 | 0.90 | 0.89 |
| Sim. | IQ | 68.88 | 68.67 | 68.26 | 68.10 | 68.46 | 68.75 |
| | MR | 0.52 | 0.48 | 0.50 | 0.48 | 0.50 | 0.52 |
| | Mech. | 1.00 | 1.00 | 1.00 | 0.95 | 1.00 | 0.95 |
| | Mat. | 1.00 | 1.00 | 1.00 | 0.89 | 0.88 | 0.91 |

### D.5 Dynamics Preservation and Appearance Change

We report a separate strength sweep over $s \in \{0.4, 0.6, 0.8, 0.9\}$, with $N = 76$ proxy videos. This analysis compares ProxyUp settings and does not include baseline DPS scores. Dynamics Preservation Score (DPS) uses CoTracker3-offline, sampling up to 128 points in the frame with the largest foreground area, and takes the geometric mean of trajectory, timing, direction/magnitude, and deformation similarities. Appearance Difference Score (ADS) uses CLIP ViT-L/14 embeddings of mask-cropped foregrounds sampled every eight frames:

$$\mathrm{ADS} = 1 - \mathrm{Median}\big(\cos\left(e^{\mathrm{fg}}_{\mathrm{proxy}}, e^{\mathrm{fg}}_{\mathrm{gen}}\right)\big)\,. \tag{15}$$

Table 11 reports means and 95% bootstrap confidence-interval half-widths. Higher DPS indicates better dynamics preservation; higher ADS indicates more appearance change, rather than necessarily better visual quality. DPS decreases and ADS increases across the tested strengths. Compared with $s = 0.4$, the default $s = 0.8$ lowers DPS from 0.615 to 0.443 while raising ADS from 0.189 to 0.230. This reflects a preference for content regeneration at the cost of motion retention.

Table 11: Dynamics preservation versus appearance change across noising strengths. Values are means $\pm$ 95% bootstrap confidence-interval half-widths.

| $s$ | DPS ↑ | CLIP-ADS ↑ |
|---|---|---|
| 0.4 | $0.615 \pm 0.062$ | $0.189 \pm 0.015$ |
| 0.6 | $0.562 \pm 0.064$ | $0.202 \pm 0.018$ |
| 0.8 | $0.443 \pm 0.060$ | $0.230 \pm 0.018$ |
| 0.9 | $0.412 \pm 0.053$ | $0.244 \pm 0.018$ |

## E Additional Qualitative Results

### E.1 Diverse Videos with Different Prompts

Fig. 7 shows outputs from a single proxy under different prompts. The examples illustrate changes in appearance and scene context while retaining recognizable motion cues.

### E.2 SDEdit with More Inference Steps

Fig. 8 shows SDEdit variants with additional sampling steps. "Denoising $\times 2$" doubles the denoising steps; "Loops $\times 2$" runs SDEdit twice, using the first output as the second input. "Inversion" applies the same inversion step as ProxyUp. All variants here also use RLN. The shown outputs do not recover the intended motion and interactions. These qualitative examples neither match model forward passes or wall-clock time nor compare alternative P&P schedules. They illustrate selected sampling failures without isolating the benefit of ProxyUp's refinement placement.

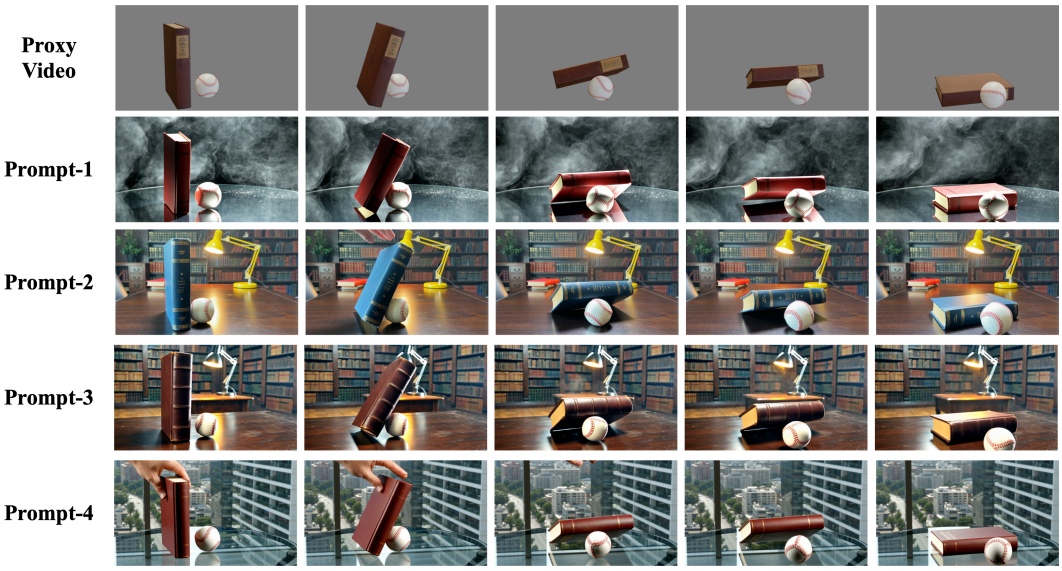



Figure 7: Outputs from the same proxy under different text prompts. The examples show changes in appearance and scene context. Detailed prompts are provided in the supplementary video.

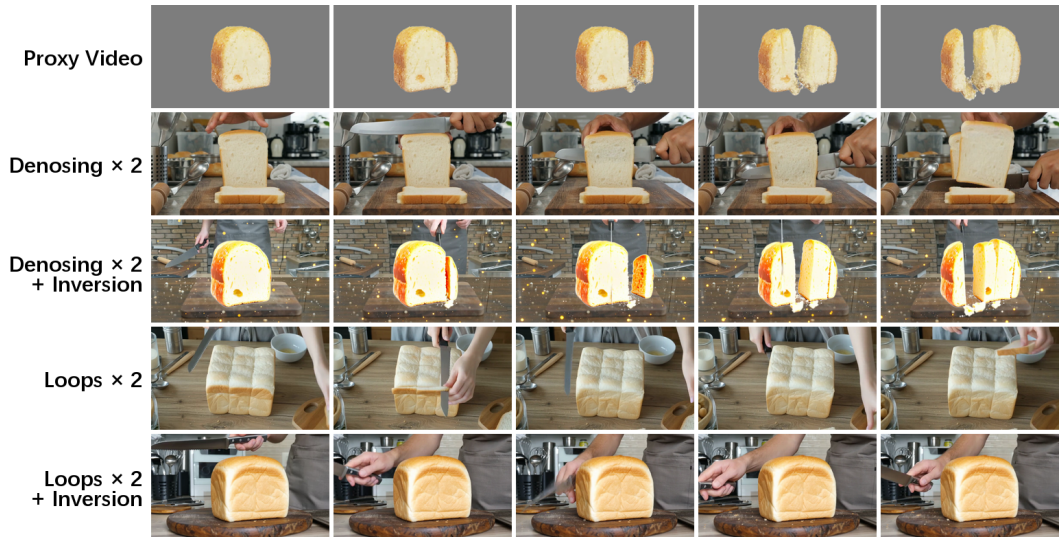



Figure 8: Qualitative results for increasing the inference budget of SDEdit. From top to bottom, the rows show the proxy video, SDEdit with doubled denoising steps, doubled denoising steps and inversion, two sampling loops, and two sampling loops with inversion.

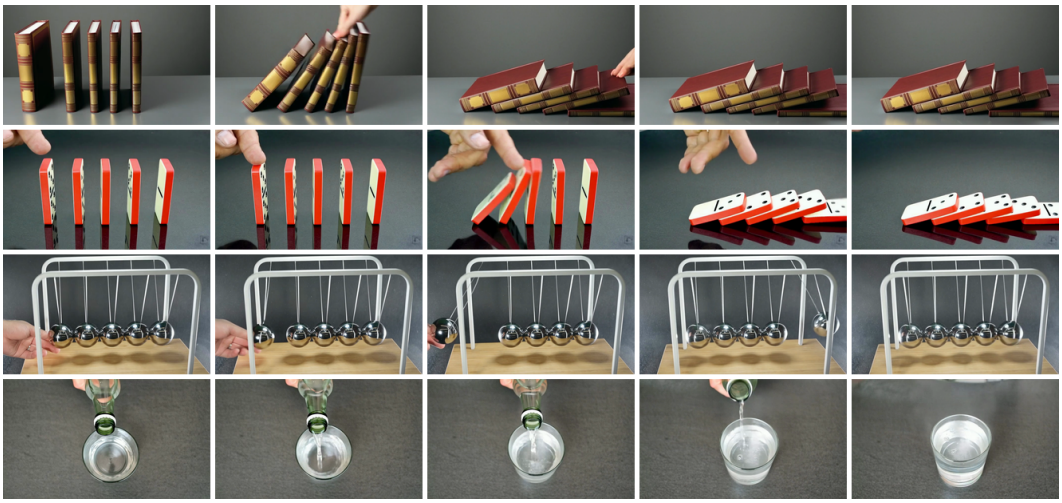


Figure 9: Cross-backbone qualitative results of ProxyUp on Wan2.1. The grid shows uniformly sampled frames from four representative generated videos.

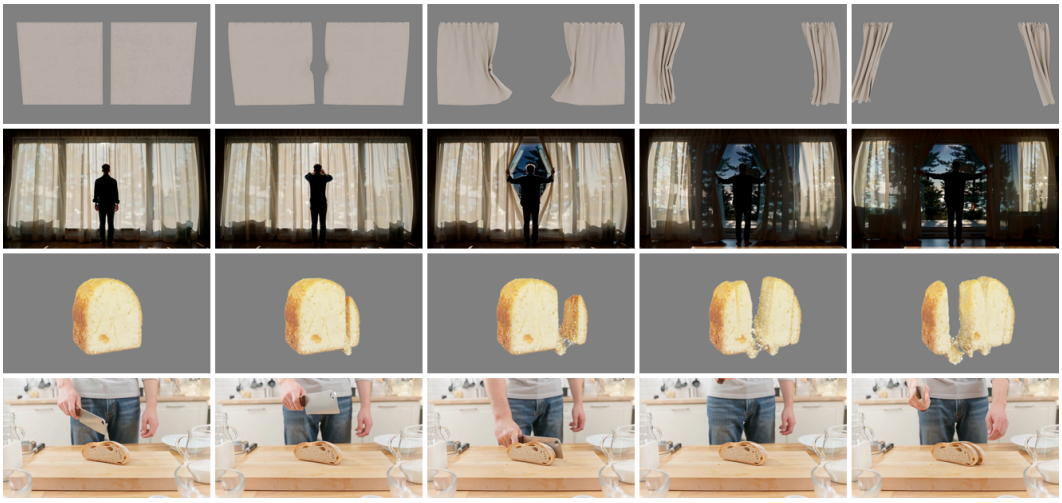


Figure 10: Cross-backbone qualitative results of DiTFlow on Wan2.2. Rows alternate between proxy videos and the corresponding DiTFlow generations on the Wan2.2 backbone.

### E.3 ProxyUp on Wan2.1

The four examples in Fig. 9 demonstrate the same pipeline on Wan2.1.

### E.4 DiTFlow on Wan2.2

Fig. 10 provides additional cross-backbone examples for DiTFlow on Wan2.2. These results complement the ProxyUp-on-Wan2.1 study and help compare the behavior of different proxy-guided mechanisms under varied backbone settings.

## F Failure Cases

Fig. 11 shows two failures. ProxyUp does not generate the intended action of lifting two balls in the Newton's cradle example and places the person above the curtain rod in the second example. The latter may reflect an ambiguous force location in the proxy.



(a) Newton's cradle (b) Pulling curtains

Figure 11: Failure cases of ProxyUp: an incorrect lifting action in Newton's cradle and an implausible person location above a curtain rod.

## G Additional Experimental Comparisons

We provide additional qualitative comparison results on six simulation examples in Figs. 12–14. Please refer to our supplementary video for more details.

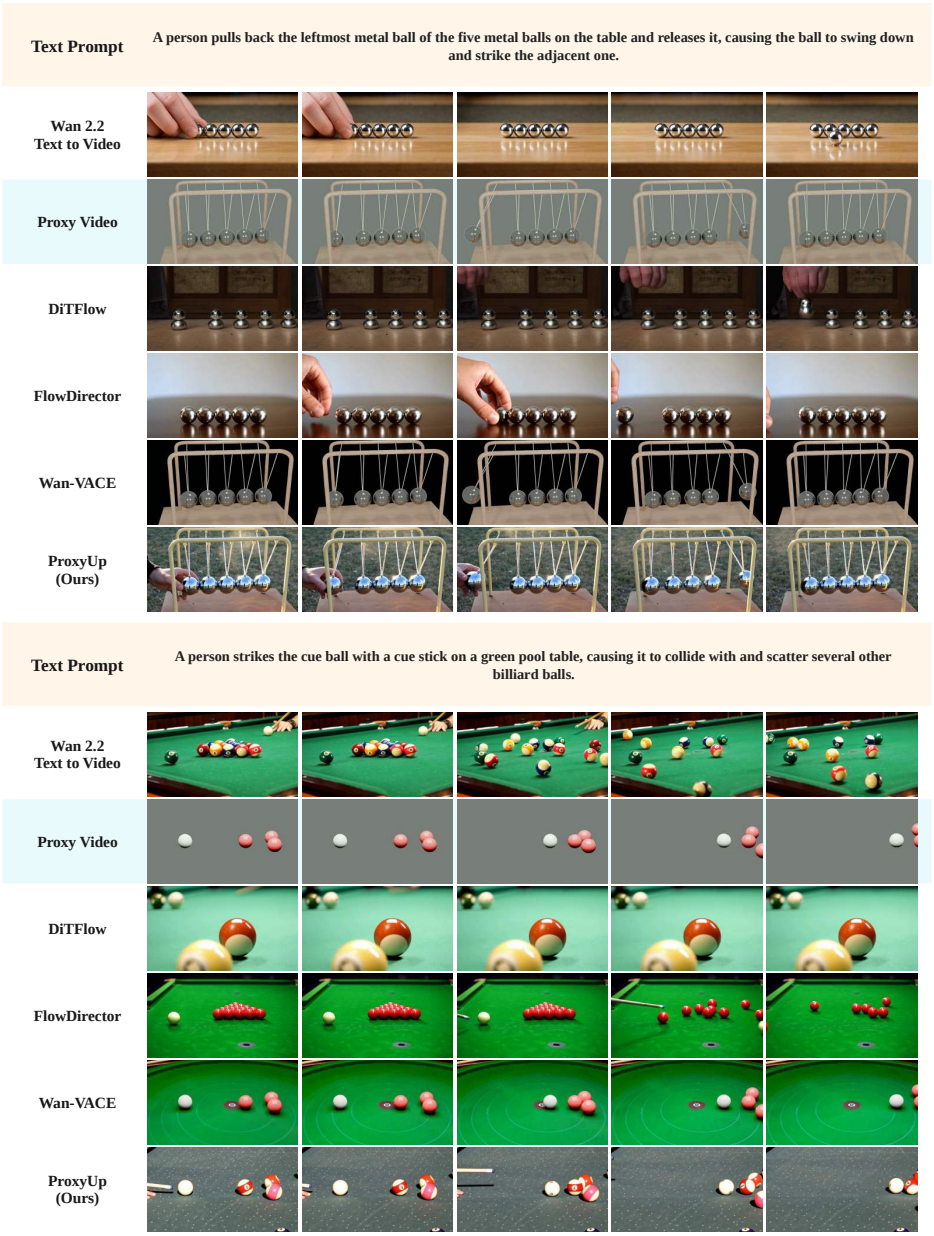



Figure 12: Additional qualitative comparison results for Newton's cradle and billiards.




Figure 13: Additional qualitative comparison results for book–baseball contact and book toppling.

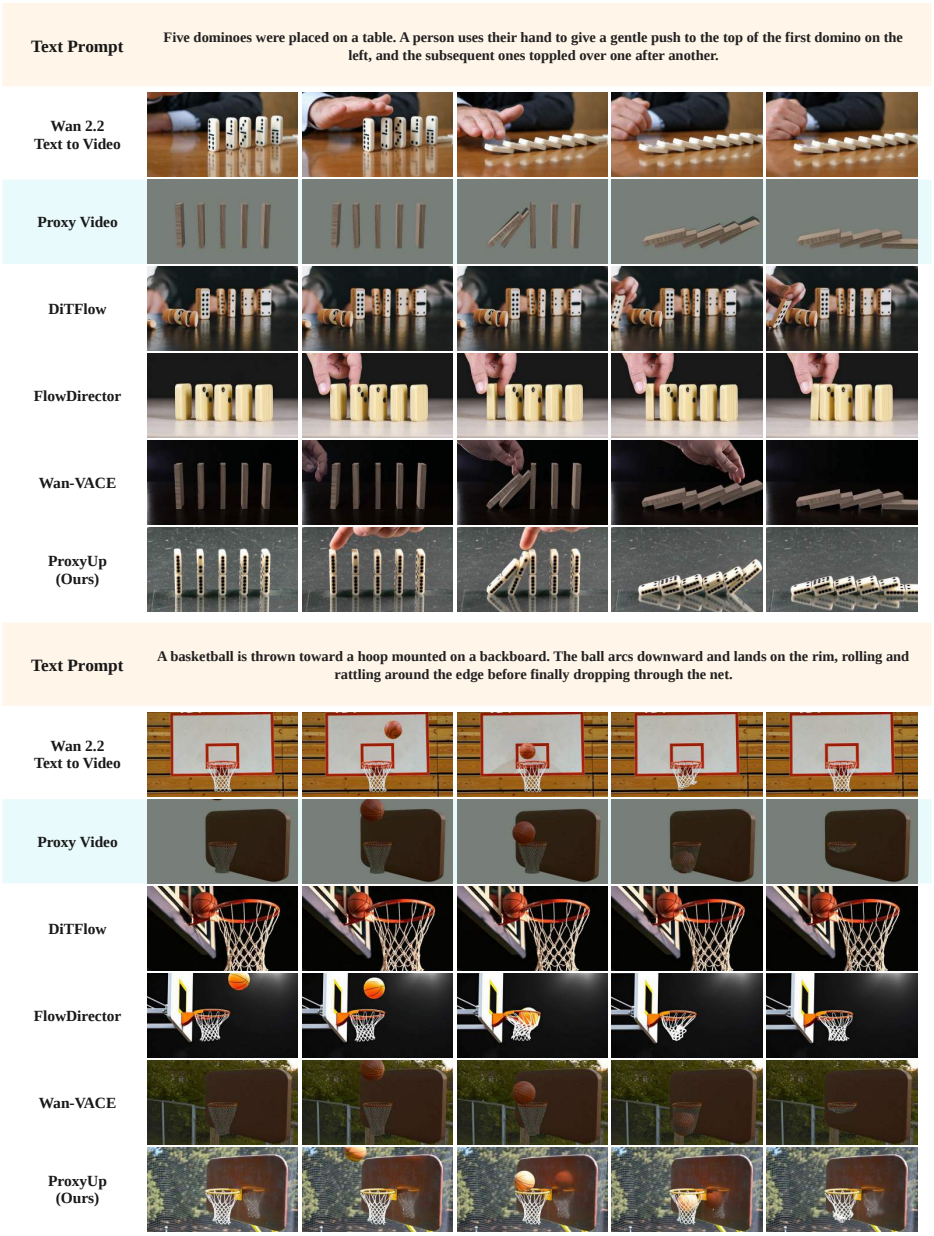



Figure 14: Additional qualitative comparison results for dominoes and basketball.