# Beyond the Visible: Disocclusion-Aware Editing via Proxy Dynamic Graphs

## Supplementary Material

In this supplementary material, we provide additional evaluation results, user study details, implementation details, and extended comparative results. The supplementary website (`Supplementary_Material/resulting_videos/web.html`) contains all generated videos.

## A. Extended Evaluation

### A.1. PDG vs. Dragging-based Manipulation

We provide in Figure A1 additional visual comparisons with dragging-based video generation methods, including Puppet-Master [29] and DragAnything [57], as well as Veo3+$\mathbf{I}$+$\mathbf{T}_m$. The output video resolutions are 256 × 256 for Puppet-Master and 576 × 320 for DragAnything. For consistency, we downscale our image frames to the larger dimension and apply zero-padding to match the target resolution. Our PDG-based manipulation method reliably follows the target motion while preserving object identity, whereas other approaches drift, distort, or produce inconsistent and uncontrolled movements. See the resulting video for our superiority.

### A.2. Full Ablation Study Statistics

We provide in Table A1 the statistics on all the evaluation metrics of varying the replacement step $M$ to 25, 30, 40, and 50.

### A.3. Details of the User Study Results

In Figure A2, we present detailed quantitative results collected from 32 participants. Our method is strongly preferred over all competitors across all three evaluation criteria (Q1, Q2 and Q3).

The questionnaire used in our user study is provided in (`Supplementary_Material/user_study/Questionnaire.pdf`), and the detailed results for each question are available in (`Supplementary_Material/user_study/Study_Analysis.pdf`). Among the 36 responses we received, the first three were collected during the pilot study and are therefore excluded. The last response was submitted after the paper submission deadline and is also not included in the analysis.

## B. Implementation Details

All experiments ran on a Linux server with 2× Intel Xeon 6710E CPUs (64 cores per socket, 128 total cores, 3.2 GHz max) and 1× NVIDIA H100 NVL (94 GB) GPU. It takes approximately 3 minutes to generate a 49-frame video at a resolution of 720 × 480.

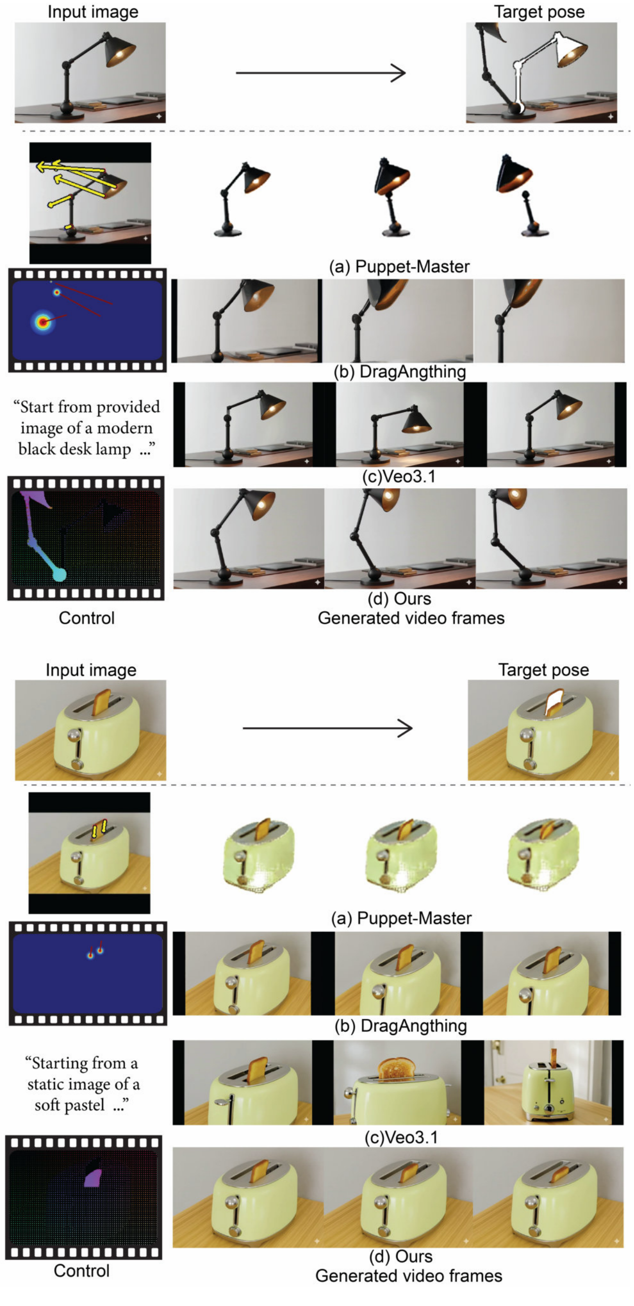


Figure A1. Given an input image and a target manipulation (top, target pose), each method produces a short video. Puppet-Masterr [29] frequently drifts from the target pose and introduces large distortions, while DragAnything [57] fails to move the articulated parts consistently. Veo3.1 exhibits random, uncontrolled motion. In contrast, our method accurately follows the specified motion and preserves object identity.

Table A1. Ablation study results. We evaluate several choices of the replacement step $M$ and report their effects on motion accuracy, last-frame similarity, and the overall video quality.

| | Motion Accuracy | Last-frame Similarity | | | Video Quality | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | OptFlow (↑) | Idiff (↓) | Idiff$_m$ (↓) | FID (↓) | FVDS (↓) | FVDC (↓) | SSIM (↑) | PSNR (↑) | LPIPS (↓) | CLIP-S (↑) |
| M=25 | 0.64 | 23.86 | 7.81 | 60.69 | **1629.01** | **1633.33** | 0.71 | 16.63 | 0.32 | 0.22 |
| M=30 | **0.65** | 23.85 | 7.45 | 59.48 | 1633.73 | 1637.90 | 0.71 | 16.63 | 0.32 | 0.22 |
| (default) M=35 | **0.65** | 23.82 | 6.91 | 57.14 | 1639.47 | 1643.88 | 0.71 | 16.63 | 0.32 | 0.22 |
| M=40 | **0.65** | 23.36 | 6.71 | **56.95** | 1641.26 | 1645.58 | 0.71 | **16.64** | 0.32 | 0.22 |
| M=50 | **0.65** | **23.28** | **6.46** | 59.13 | 1640.33 | 1644.83 | 0.71 | **16.64** | 0.32 | 0.22 |

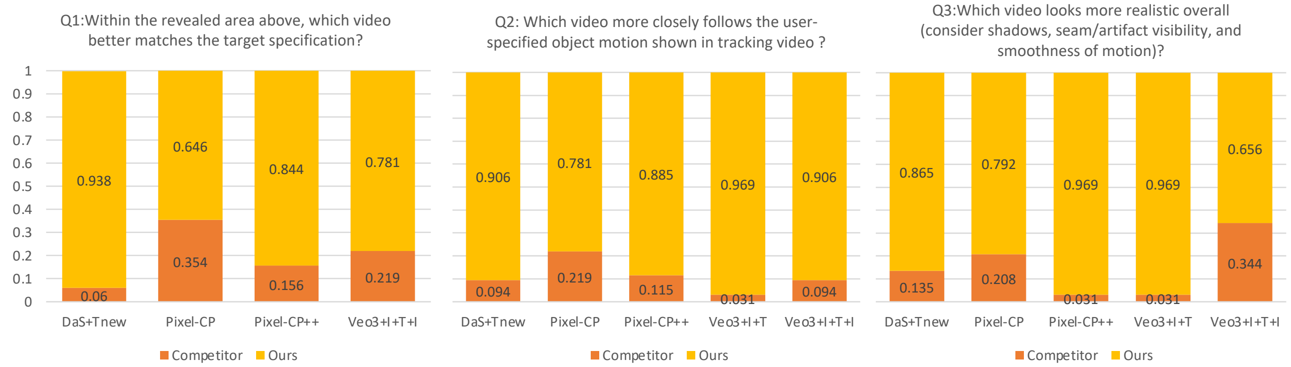


Figure A2. **User evaluation results.**

We follow the implementation of DaS, which overwrites non–first-frame features in the VAE latent space with zero. Accordingly, we also zero out the non–first-frame features within the non-disocclusion regions.

# C. More Results

## C.1. Additional Comparative Results

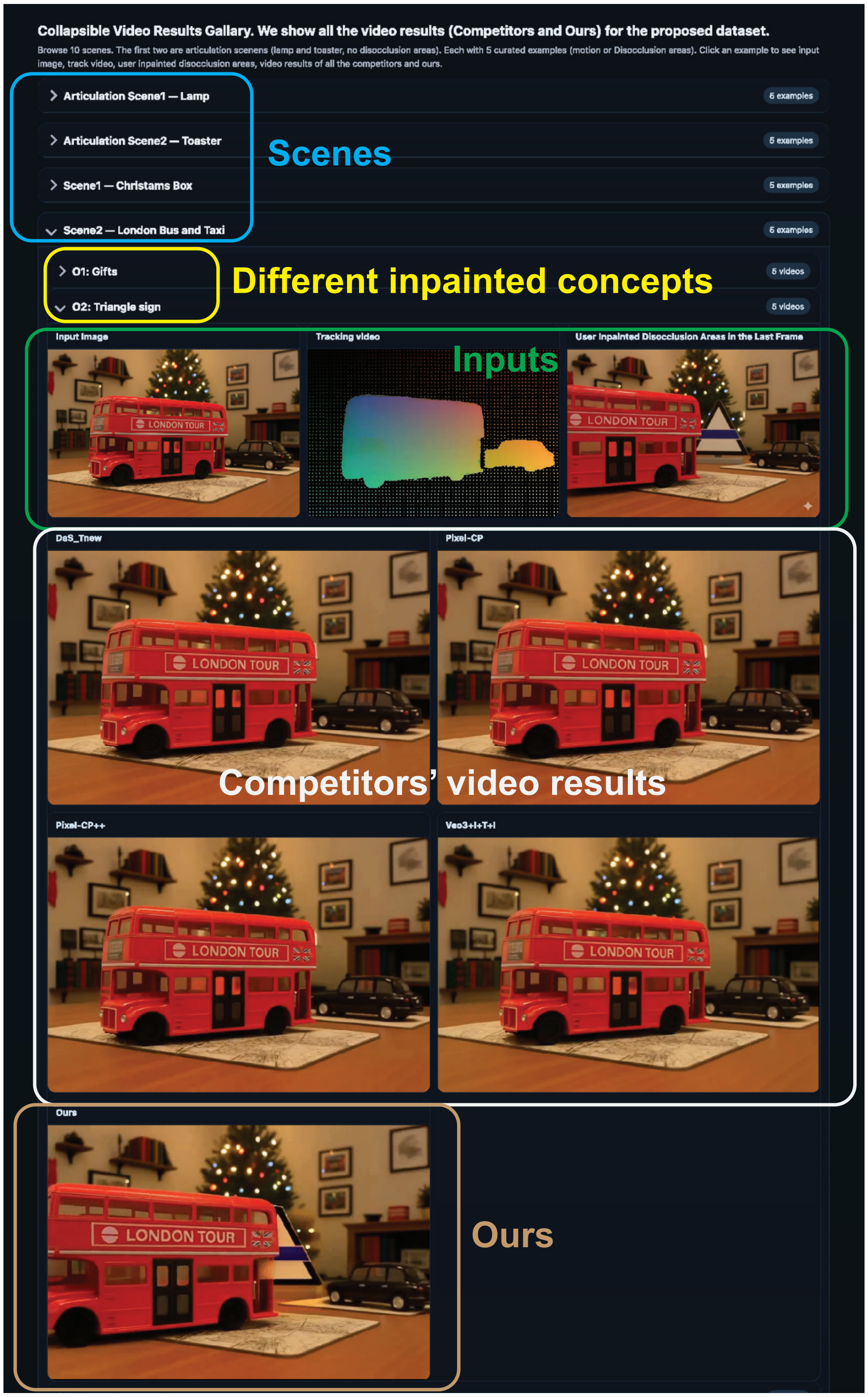
We include all video results—ours and competitors—in the uploaded supplementary material in the format of a webpage (`Supplementary_Material/resulting_videos/web.html`). It includes all 10 benchmark scenes. The first two (desk lamp and toaster) have richer articulation, simple backgrounds, and no disocclusions; each provides five curated manipulation examples. The remaining eight feature simpler motions but challenging disocclusions, each with five user-inpainted variants. Clicking any example reveals the input image, tracking video, user-edited disocclusion masks, and generated results from all methods. Figure A3 shows the results layout of the webpage.

## C.2. Expanded Text Prompts

We provide the detailed prompts (*i.e.*, $\mathbf{T}$, $\mathbf{T}_{new}$, and $\mathbf{T}_m$) used for the **Drawer** example in the second row in Figure 5 (Scene3-Drawer, O4: White fabric on the webpage). They are:

- $\mathbf{T}$: “The drawer is closing.”
- $\mathbf{T}_{new}$: “There is a white fabric in the drawer below.”
- $\mathbf{T}_m$: “A highly realistic static-camera shot of a wooden wardrobe with two drawers, viewed from a fixed angle at chest height. The upper drawer is made of light oak wood with a metallic knob, and the lower drawer is darker walnut with a matching brass knob. The wardrobe is softly lit by warm ambient light, with neatly folded clothes on the shelf above the drawers. Starting state: both drawers are slightly open, showing empty interiors faintly. Motion: the upper light-oak drawer slides smoothly backward along its straight rails until it closes completely, maintaining level alignment with no shake. Simultaneously, the darker walnut drawer glides forward along its track, moving outward in a straight, even path as its dark interior gradually becomes visible. The camera remains fixed, without any pan, tilt, or zoom; the lighting remains steady. End state: the lower drawer is fully open, revealing a white, curved fabric resting at its center. The upper drawer stays fully closed, and the scene remains calm and stable under the same warm light”.

We prompt ChatGPT-5.1[37] using the input image, the tracking video, and the final frame containing user-inpainted new concepts to obtain the $\mathbf{T}_m$ description. Note that this prompt is ONLY used in competitors.



**Collapsible Video Results Gallary. We show all the video results (Competitors and Ours) for the proposed dataset.**

Browse 10 scenes. The first two are articulation scenens (lamp and toaster, no disocclusion areas). Each with 5 curated examples (motion or Disocclusion areas). Click an example to see input image, track video, user inpainted disocclusion areas, video results of all the competitors and ours.

Articulation Scene1 — Lamp
5 examples
Articulation Scene2 — Toaster
5 examples
Scenes
Scene1 — Christams Box
5 examples
Scene2 — London Bus and Taxi
5 examples
01: Gifts
5 videos
Different inpainted concepts
02: Triangle sign
5 videos
Input Image
Tracking video
Inputs
User Inpainted Disocclusion Areas in the Last Frame
DaS_Tnew
Pixel-CP
Competitors' video results
Pixel-CP++
Veo3+I+T+I
Ours
Ours

Figure A3. A screenshot of the webpage.