Scaling Instruction-Based Video Editing with a High-Quality Synthetic Dataset

Supplementary Material

A. Overview

In this appendix, we provide additional details and results to supplement our main paper. In Sec. B and Sec. C, we present comparisons of data pipelines and the analysis of the in-context data generator. Sec. D contains additional quantitative and qualitative results of our dataset and model, and a word cloud visualizing the distribution of instructions in our dataset. Finally, we showcase the interface used for our human evaluations in Sec. E. We strongly recommend reviewing the video samples in **"index.html"** in Supplementary Materials for a better understanding.

B. Comparison of Data Pipelines

We compare the data synthesis and filtering strategies of Ditto with InsViE and Señorita in Table S1. Regarding the synthesis pipeline, InsViE relies on I2V inversion which is inherently slow and costly, while Senorita necessitates the complex training and maintenance of 19 distinct expert models. In contrast, Ditto employs all-in-one models, offering a significantly more efficient and scalable generation process. In terms of data filtering, we address the limitations of Señorita's CLIP-based approach by introducing tracking + VLM protocol. Specifically, our tracking mechanism precisely filters out samples with insufficient motion, ensuring superior dynamic range compared to Senorita, while the VLM rigorously guarantees editing quality and safety. For concrete visual comparisons demonstrating these advantages, we strongly recommend readers refer to the samples provided on our **HTML page**.

Table S1. Comparison of Data Synthesis Pipelines. We contrast the synthesis models and filtering strategies.

Method	Synthesis Pipeline	Data Filtering
InsViE	I2V Inversion	Optical Flow + VLM
Señorita	Multiple Expert Models	CLIP
Ditto	All-in-one Models	Tracking + VLM

C. Justifying Data Generation Pipeline Design

We provide an analysis of the videos synthesized by the in-context video generator, VACE, to justify the design of our data pipeline. As in Fig. S1, we first observe that using only depth maps to guide the generator results in a significant loss of content from the source video, leading to poor fidelity. Conversely, conditioning the generator on a keyframe from the original source video alongside the editing instruction fails to produce the desired edit - the output remains almost identical to the source. These findings reveal that while the base generator excels at motion transfer, its inherent instruction-following capability for editing tasks is limited. Based on this analysis, we validate our proposed approach: using a keyframe modified by an advanced image editor, in conjunction with depth guidance as the context. This method achieves the optimal balance of instruction adherence, temporal consistency, and source fidelity for our data synthesis.

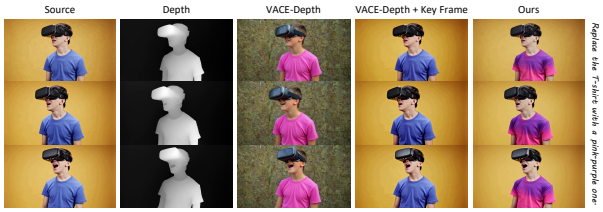

Figure S1. Results of various settings for data generation.

D. Additional Results of Dataset and Model

Table S2. Ablation on modality curriculum learning (MCL).

Configuration	CLIP-T ↑	CLIP-F ↑	VLM ↑
w/o MCL	24.79	98.91	7.60
w/ MCL	**25.54**	**99.03**	**8.10**

We first conduct ablation study on modality curriculum learning (MCL) in Tab. S2 to validate its effectiveness. We then include results of Vision-Language Model (VLM) data filtering in Fig. S2, as well as the local editing data on adding and removal in Fig. S3. We present additional qualitative results to further demonstrate our dataset and model's performance across a wide range of editing instructions in Fig. S4. We also include a word cloud in Fig. S5 that illustrates the diversity and distribution of the editing prompts within the dataset, highlighting its coverage.

E. Demonstration of User Study Interface

To collect human preference data, we designed a user-friendly evaluation interface, as shown in Fig. S6. For each source video and text prompt, we presented participants with the edited results from different methods in a randomized order. They were then asked to rank the videos from best (1) to worst (4) based on three criteria: Instruction Following, Temporal Consistency, and Overall preference. The final scores are calculated based on the ranking.

Figure S2. Data filtering results from the Vision-Language Model (VLM). We design a set of rules for the VLM to guide its data filtering process. Experiments demonstrate that Large VLMs can reliably detect semantic and other types of failures.

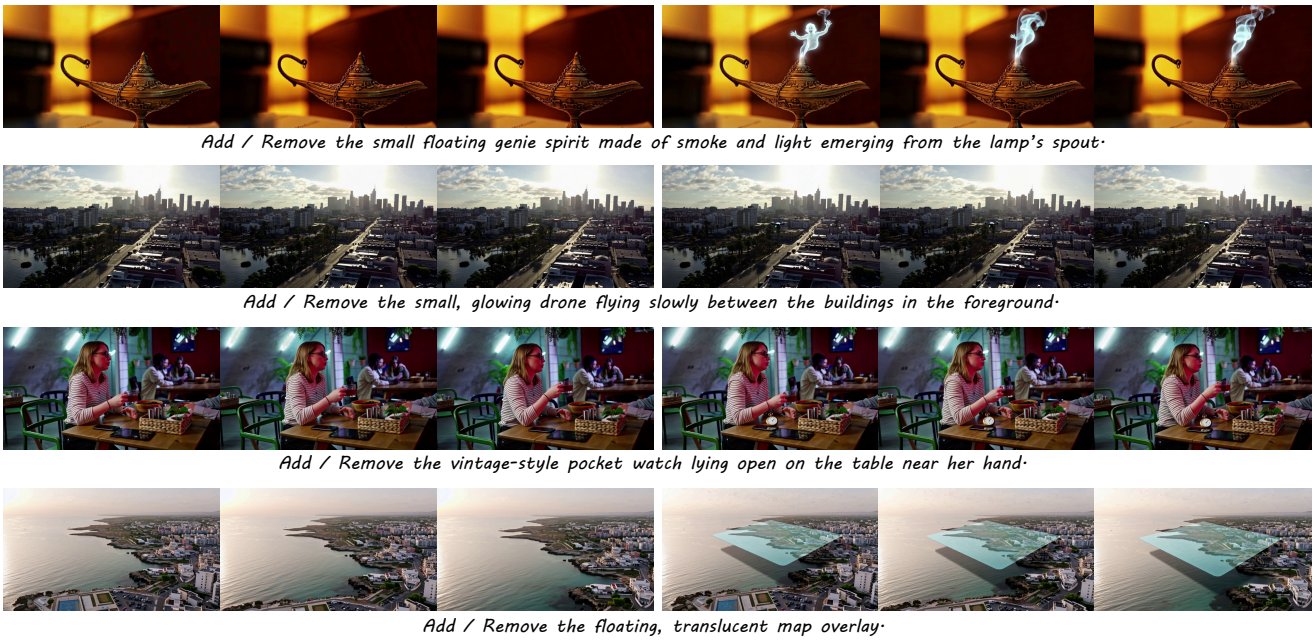

Figure S3. Additional visualization of local editing data on adding and removing. Given the depth corresponding to the source video, our pipeline can stably synthesize data for addition-type edits, which are then *reversely* used as training data for removal-type editing.

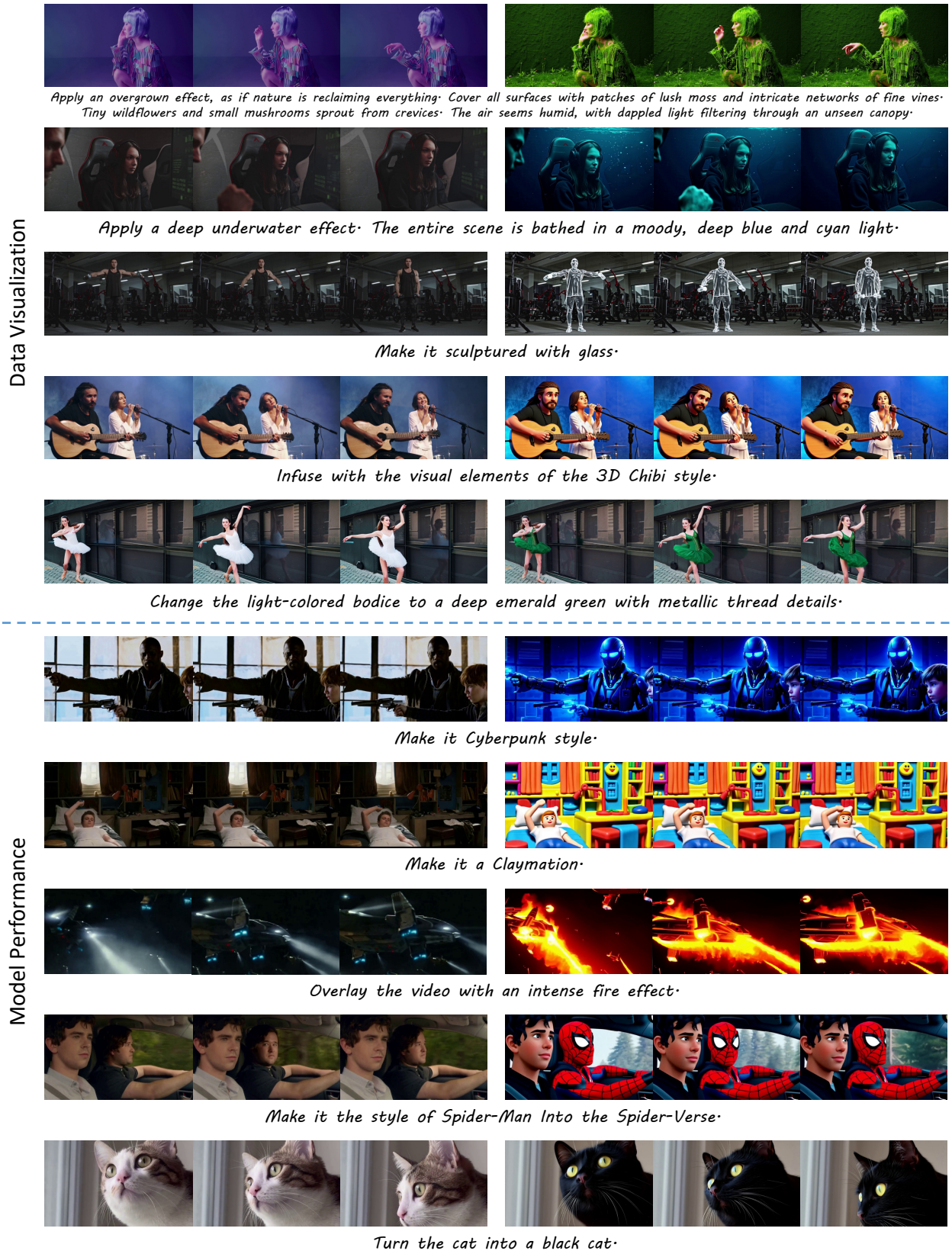

Figure S4. Additional visualization of data from the proposed dataset and model outputs. Please view the site in the supplementary materials for additional video samples.

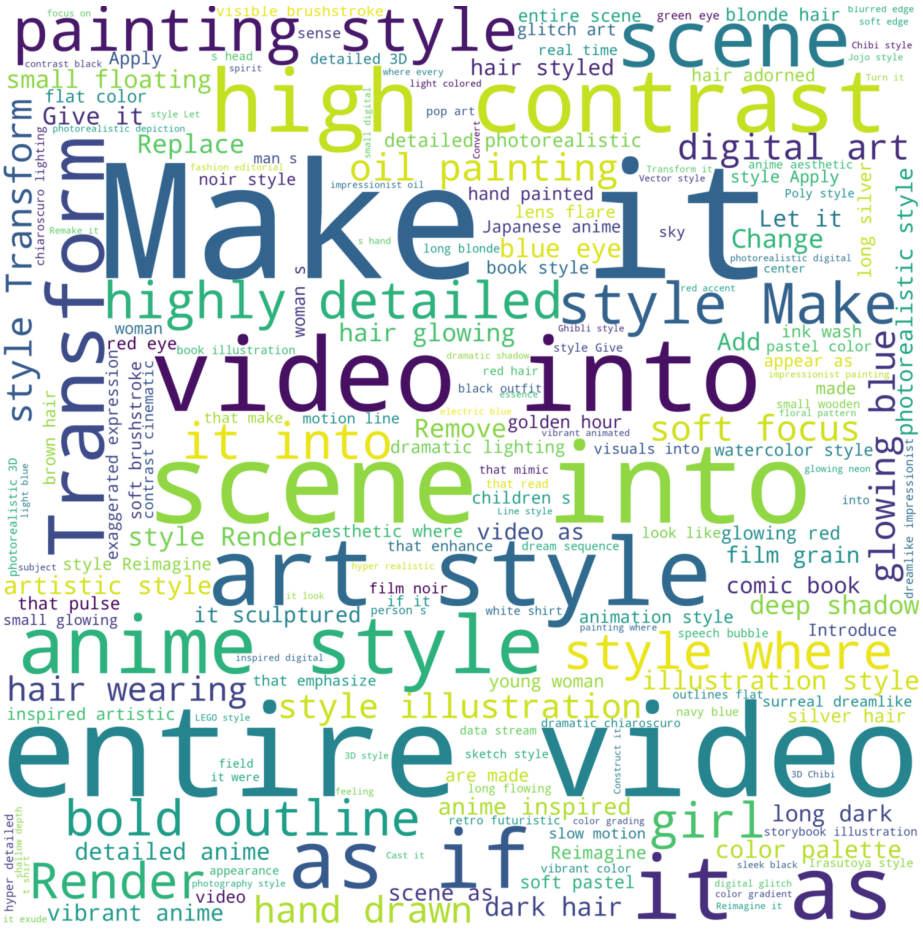

Figure S5. The word cloud of editing instructions.

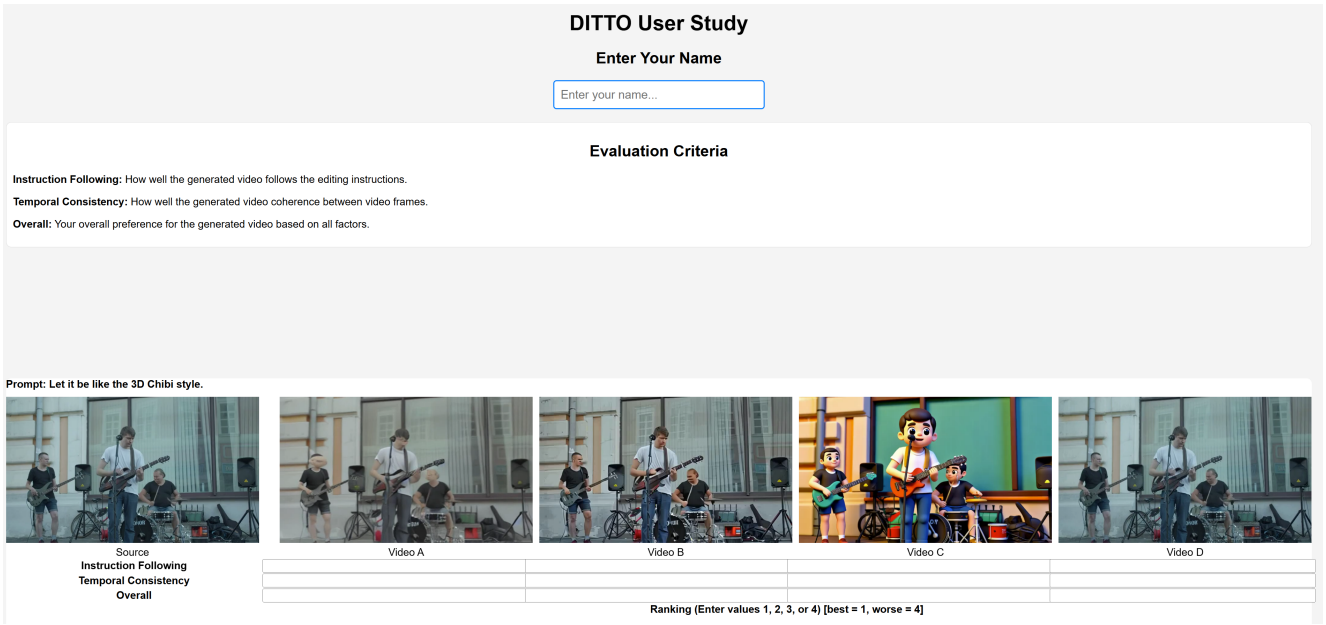

Figure S6. The interface of the user study.