Scaling Instruction-Based Video Editing

with a High-Quality Synthetic Dataset

CVPR 2026

Qingyan Bai[1,2], Qiuyu Wang[2], Hao Ouyang[2], Yue Yu[1,2], Hanlin Wang[1,2],
Wen Wang[2,3], Ka Leong Cheng[2], Shuailei Ma[2,4], Yanhong Zeng[2],
Zichen Liu[1,2], Yinghao Xu[2], Yujun Shen[2], Qifeng Chen[1]
[1]HKUST [2]Ant Group [3]Zhejiang University [4]Northeastern University

A Growing Divide: Image vs. Video Editing

Instruction-based Image Editing

✅ **High Precision & User-Friendly:** Intuitive tools that respond accurately to simple text prompts.
✅ **Proven Success:** Models like **FLUX Kontext, GPT-Image, and Nano-Banana** have set new standards for high-quality results.
✅ **Nuanced Control:** Enables fine-grained, user-guided modifications to specific parts of an image.

Instruction-based Video Editing

⚠️ **Significant Lag:** Still far behind its image-editing counterparts in terms of usability and output quality.
⚠️ **Temporal Inconsistency:** Major challenge in maintaining visual coherence across frames, leading to flickering or distorted results.
⚠️ **Inherent Complexity:** Video editing involves multiple frames, motion dynamics, and long-term context, making it much harder to master.

The Fundamental Bottleneck: Data Scarcity

THE CORE CHALLENGE

The primary obstacle to advancing AI video editing is the profound scarcity of **large-scale, high-quality, and diverse paired data**. This lack of training material directly limits progress.

- **What is "Paired Data"?** Critical triplets: (Source Video, Editing Instruction, Edited Video).
- **Prohibitive Cost:** Manually creating such datasets is extremely expensive and time-consuming.
- **Performance Roadblock:** Insufficient data makes training high-performance models extremely difficult.

Current Approaches Fall Short

Inversion-based

(e.g., TokenFlow)

Pros: No need for paired data, offering flexibility in training.

Cons: Computationally intensive and struggles significantly with complex, non-rigid motion.

Feed-forward

(e.g., VEGGIE)

Pros: Operates end-to-end for faster inference speeds, suitable for real-time applications.

Cons: Faces critical face data scarcity issues and often results in jarring temporal inconsistencies.

Specialized "Expert System"

(e.g., Señorita)

Pros: Delivers high-quality results specifically for predefined tasks or domains.

Cons: Not easily scalable to new domains or use cases and requires significant manual development effort.

Introducing Ditto: A Scalable Solution

🎯 GOAL

To systematically address the critical data scarcity challenge that plagues the development of instruction-based video editing models.

🧩 WHAT IS DITTO?

A holistic framework designed to automate the generation of a large-scale, high-quality synthetic dataset for training advanced video editing models.

💡 CORE IDEA

Combine state-of-the-art image editors and in-context video generators, orchestrated by an intelligent agent, to create a fully automated data creation pipeline.

Data Generation

Efficiently produces diverse, high-fidelity synthetic video pairs to fuel model training.

VLM Agent

Acts as the central brain to translate user instructions into executable editing actions.

Editto Model

The resulting state-of-the-art video editing model, fine-tuned on the generated dataset.

OUTCOME: A Million-Scale Synthetic Dataset & a State-of-the-Art Video Editing Model

The Ditto Pipeline: From Source to High-Quality Data

01. PRE-PROCESS

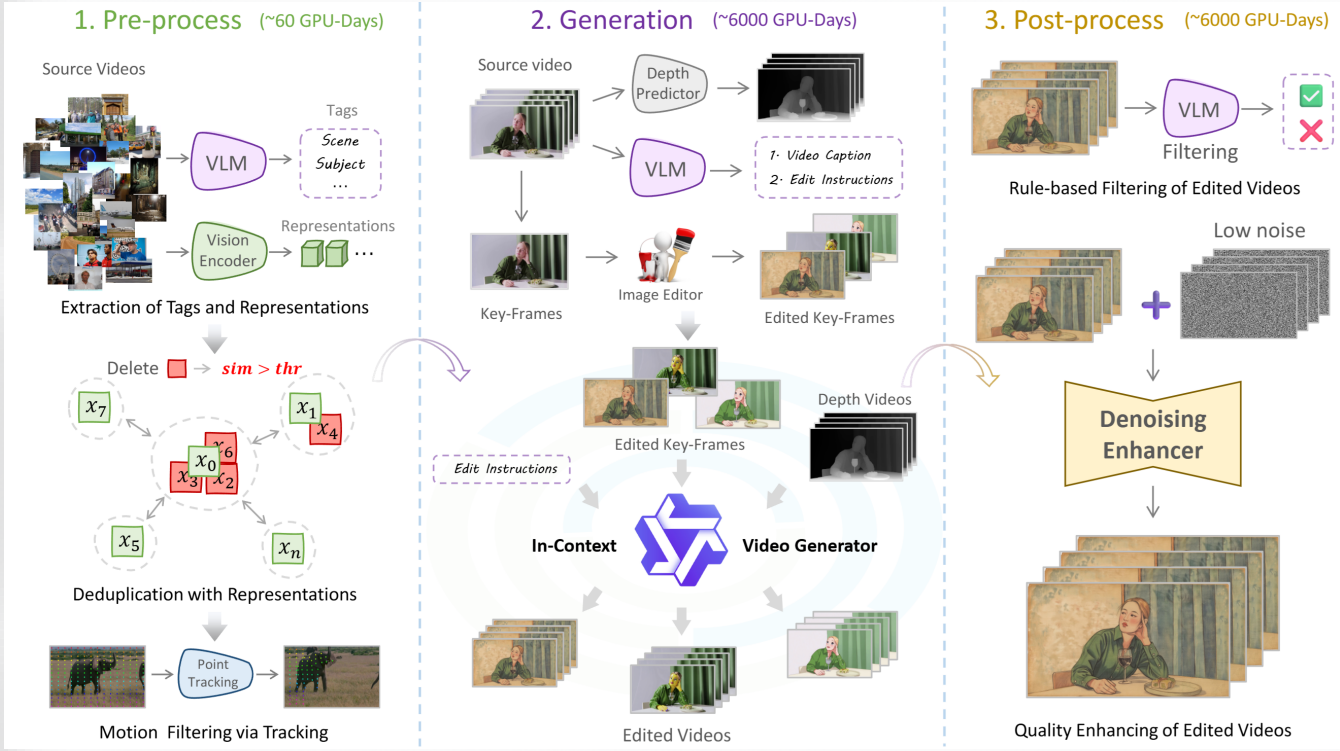
Curate a diverse video pool through automated deduplication and motion filtering to ensure source data quality.

02. GENERATION (Core Engine)

Synthesize video triplets by conditioning an in-context generator on automated instructions, edited key-frames, and depth maps.

03. POST-PROCESS

Guarantee final visual fidelity with a VLM-based filter and a dedicated denoising enhancer.



Step 1: Source Video Filtering

Building a High-Quality Foundation

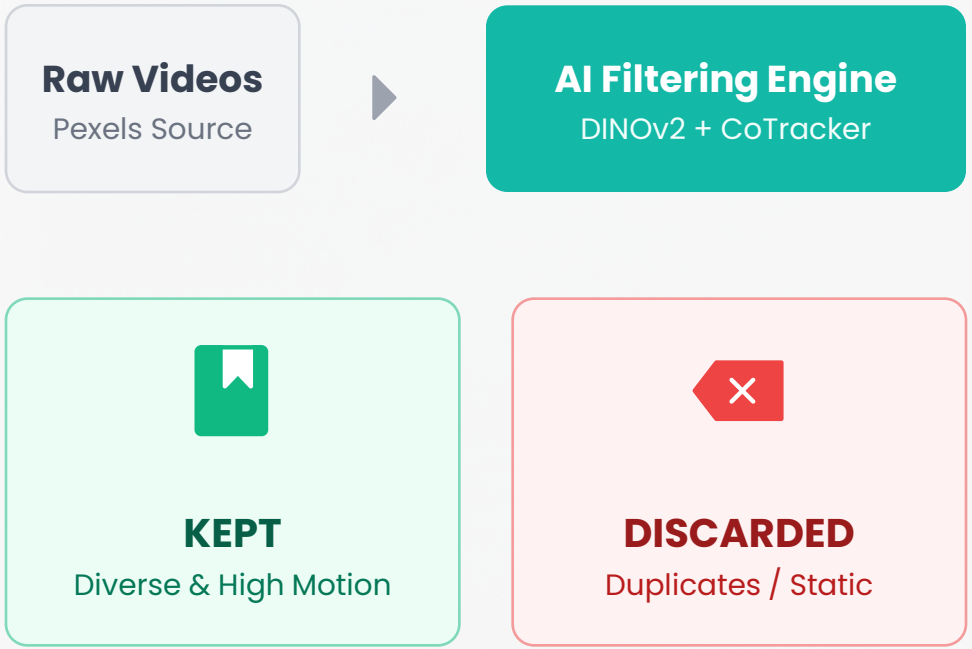
Data Sourcing

Curated professional-grade footage from Pexels to ensure high-resolution and copyright-safe content.

Dual-Stage Filtering Protocol

1. **Near-Duplicate Removal:** DINOv2 feature extraction eliminates redundant or overly similar videos.
2. **Motion Scale Analysis:** CoTracker tracks inter-frame motion vectors to filter out static/low-dynamics clips.

OUTCOME: Clean, diverse & dynamic source pool



Step 2: Instruction Generation

Automating the "What to Edit"

The Challenge

Creating a large, diverse, and meaningful set of editing instructions manually is resource-intensive and impossible to scale for training AI models.

The Solution

01. Dense Captioning: Use a powerful VLM (Qwen2.5-VL) to generate a detailed, fine-grained description of the video's visual content.

02. Instruction Generation: Leverage the generated caption and video to synthesize plausible, creative editing instructions.

Key Outcome

A scalable pipeline that produces a high-quality dataset of **contextually grounded instructions**, supporting both **global (e.g., style transfer)**and **local (e.g., object modification)** video editing tasks.

Step 3: Visual Context Preparation

Guiding the Generation with Rich Context

Edited Key-Frame (Appearance Guidance)

• Select a key-frame and edit it using a state-of-the-art image editor (e.g., Qwen-Image).
• This frame acts as a visual prototype, defining the target style, color palette, and key content elements.

Depth Video (Spatio-temporal Structure)

• Generate a depth video from the original source material.
• It functions as a dynamic structural scaffold, ensuring the preservation of the original scene's spatial geometry and temporal motion flow.

Input Context: Source Frame → Edited Key-Frame & Depth Map

Step 4: In-Context Video Generation

The Heart of Ditto: Synthesizing the Edited Video

Core Component

The central engine is an **in-context video generator (VACE)**. It serves as the core logic responsible for translating user edits into a coherent, full-length video.

Key Inputs

The VACE model processes three distinct inputs to produce the final output:

- Text Prompt (p)
- Edited Key-Frame (f'_k)
- Depth Video (V_d)

Propagation Process

The generator intelligently propagates the single key-frame edit across the entire video sequence, while strictly adhering to the original motion and structural information derived from the depth video.

Optimized Efficiency

To enable practical use, we applied model quantization and knowledge distillation. These optimizations drastically reduce computational costs to **just 20%** of the original model.

Step 5: Curation & Enhancement

Ensuring Quality at Scale

VLM-Based Curation

• An autonomous VLM agent acts as a judge to perform rejection sampling.
• It evaluates each generated triplet against criteria like instruction fidelity, visual quality, and safety.

Quality Enhancement via Denoising

• The curated videos are refined using a specialized fine denoiser.
• This removes subtle artifacts and enhances textures, ensuring high-fidelity visual output.

BEFORE: Raw Output
Potential artifacts, grainy textures

AFTER: Denoised & Polished
Clean, high-detail, and professional

The Result: Ditto-1M Dataset

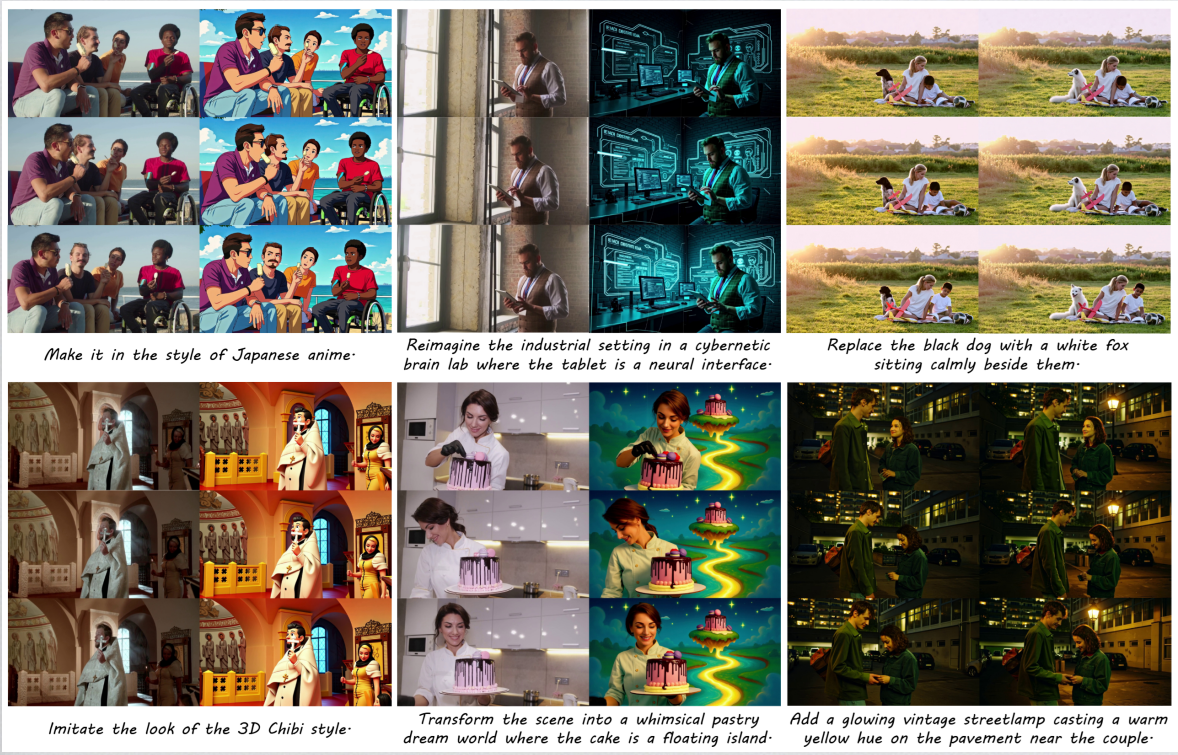
A Million-Scale Editing Dataset

Scale
Over 1 million high-quality source-instruction-edited video triplets.

Composition
~700k global edits and ~300k local edits, offering rich diversity.

Quality
Uniform 1280x720 resolution, with 101 frames at 20 FPS.

Accessibility
Fully open-source to facilitate and encourage community research.



Examples of video edits from the Ditto-1M dataset

From Data to Model: Editto

Training a State-of-the-Art Video Editor

Backbone Architecture

We use the in-context video generator **VACE** as our starting point to leverage its strong visual foundation.

Project Goal

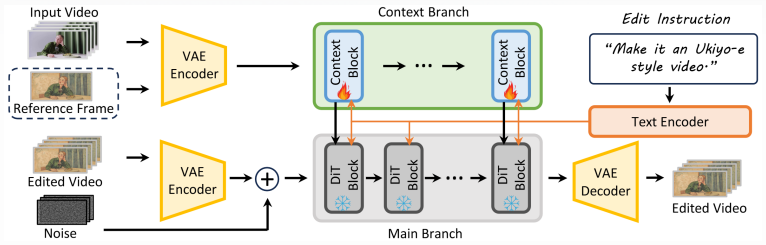
Transform the visual-conditioned generator into a proficient editor that operates purely on **textual instructions**.

Core Challenge

Directly bridging the semantic gap from visual to textual conditioning often leads to unstable and inconsistent results.

Our Solution: MCL

A novel training strategy called **Modality Curriculum Learning (MCL)**that smoothly transitions the model from visual to text control.



Editto Model Training Pipeline

The MCL strategy guides the model to gradually abandon visual dependencies and learn to follow precise textual editing commands.

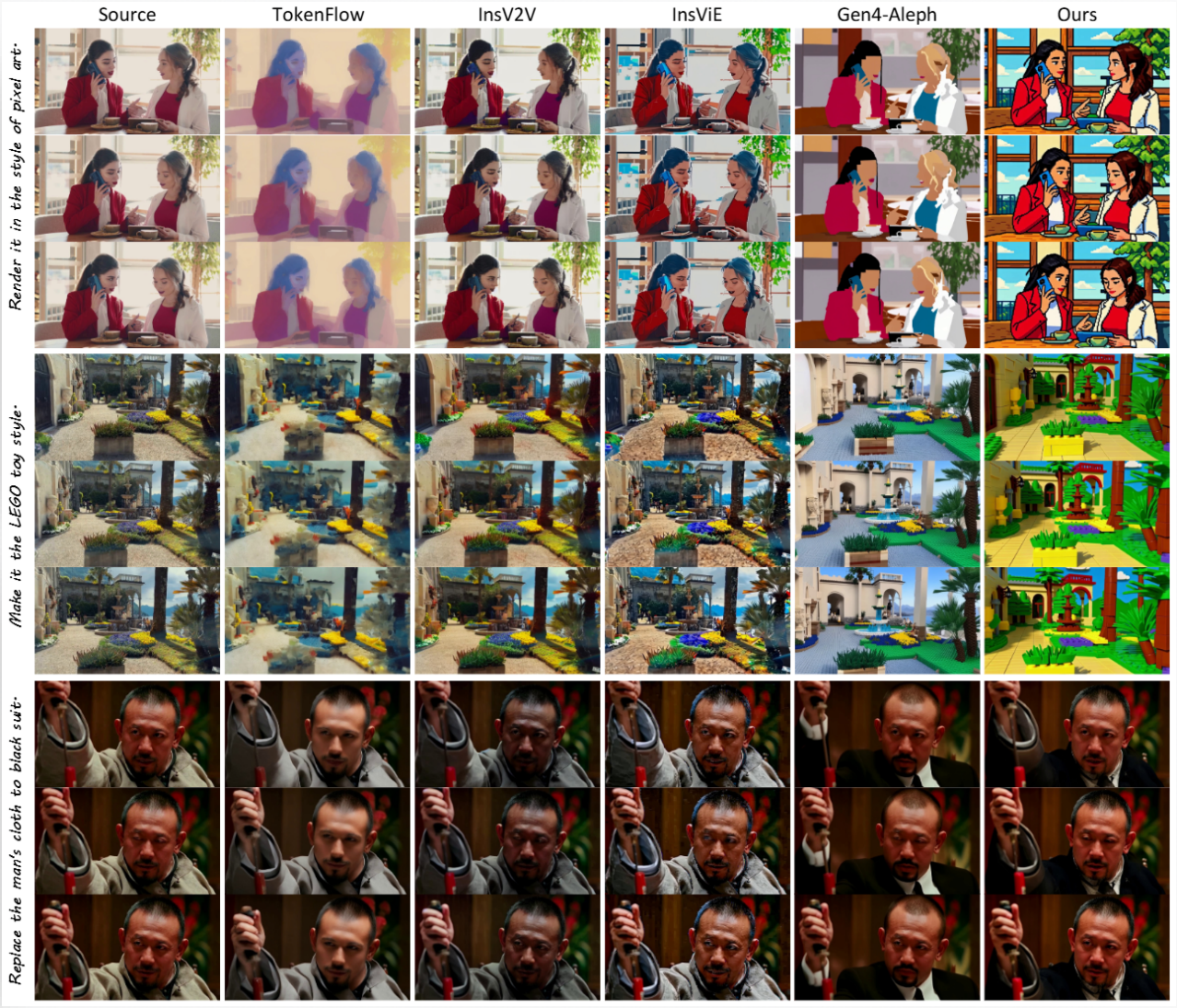
Visually Superior Edits

Complex Stylization

Our model generates temporally coherent videos that accurately match the target style (e.g., pixel art), while competitors produce blurry or inconsistent results.

Local Attribute Changes

Precisely edits the target object (e.g., changing a shirt to a black suit) while preserving identity and background details.



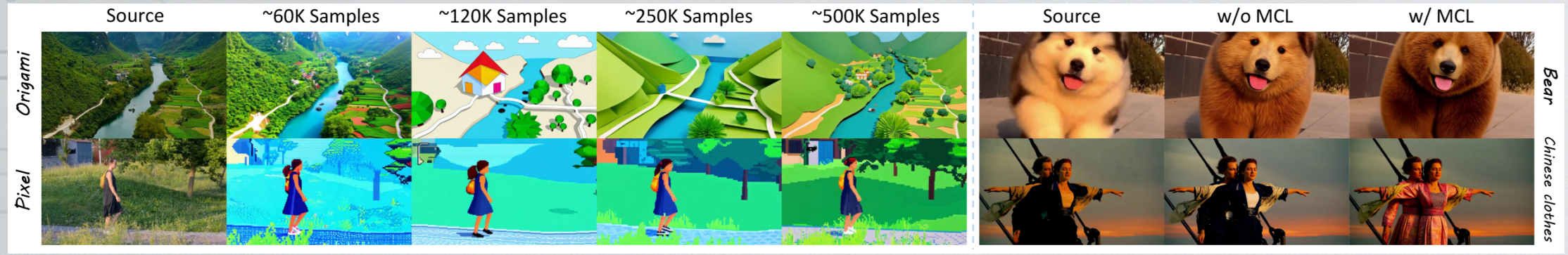
Ablation Studies: Validating Key Components

Data Scale

Performance scales effectively with the size of the training data. Experimental results consistently show that incorporating more high-quality data leads to significantly better generalization and instruction-following accuracy.

Modality Curriculum Learning (MCL)

Ablation analysis demonstrates that without MCL, the model struggles to interpret the full semantic intent of multi-modal instructions. This component is crucial for bridging the gap between different input modalities.



Conclusion & Contributions

01

Ditto Pipeline

A novel, scalable synthesis pipeline for high-fidelity video editing data.

02

Ditto-1M Dataset

A million-scale, open-source dataset to facilitate future research in video editing.

03

Editto Model

A state-of-the-art editing model demonstrating superior performance against existing alternatives.

04

Modality Curriculum Learning

An effective training strategy to build a purely instruction-driven video editor.

Thank You!

Project Page: [https://editto.net/] | Contact: [qingyanbai@hotmail.com]