

NEURIPS 2026 · VIDEO TEXT EDITING

# SteerVTE: Seamless Video Text Editing with Style and Glyph Control

Precise text replacement with source style and motion preserved

Kai Zeng · Moran Li · Zhengwei Wang · Yingchen Yu · Yiheng Lin · Ruichuan An · Ming Lu · Qi She · Wentao Zhang

Peking University · ByteDance China · ByteDance Singapore



# Video text editing and its failure modes



## Per-frame editing

Accurate text can still flicker.

## Image-to-video

Motion can change the source scene.

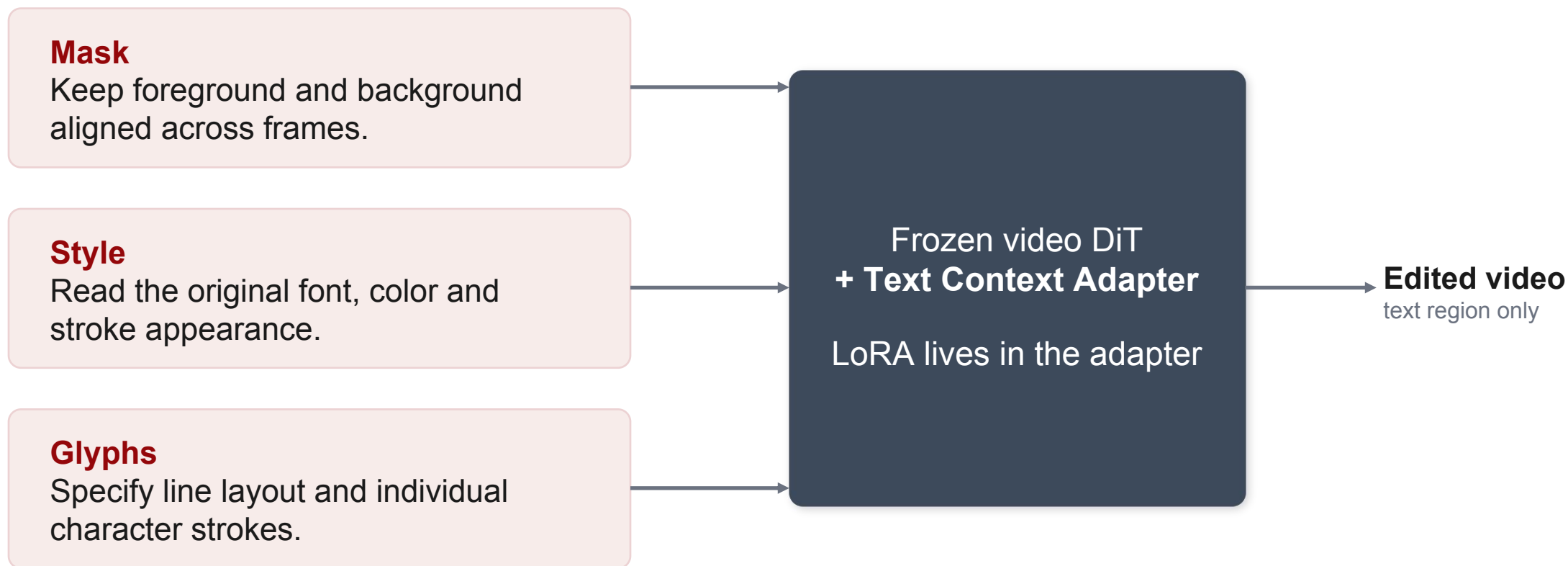
## Video-to-video

Weak glyph priors produce errors.

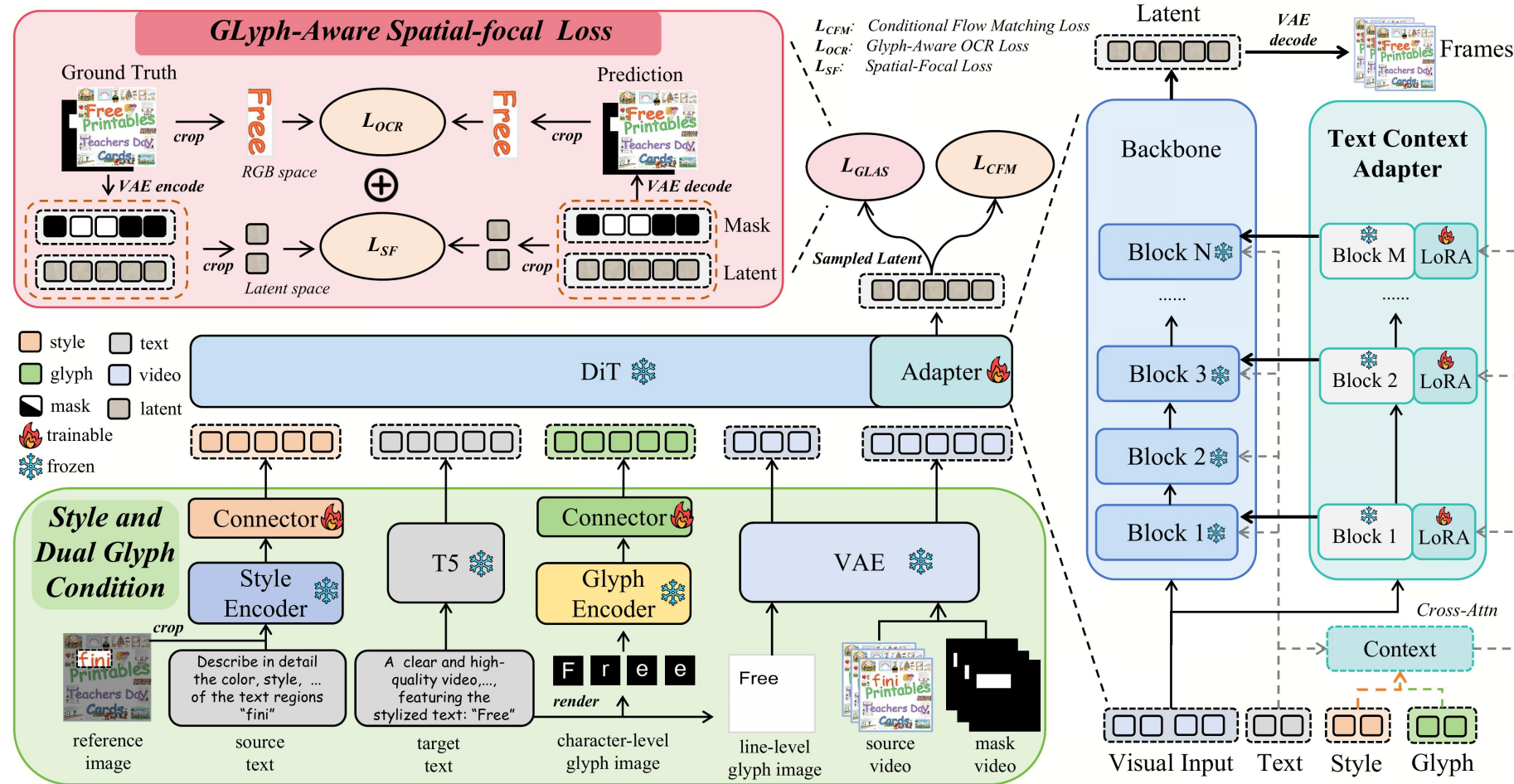
## KEY IDEA

# A frozen video prior with explicit text control

The adapter separates edit location, source appearance and target content. Connectors align the features; LoRA adapts the text context adapter in later stages.



# Style and glyph conditions complement each other



## VLM style encoding

Preserve text aspect ratio, spacing and stroke cues that square resizing can distort.

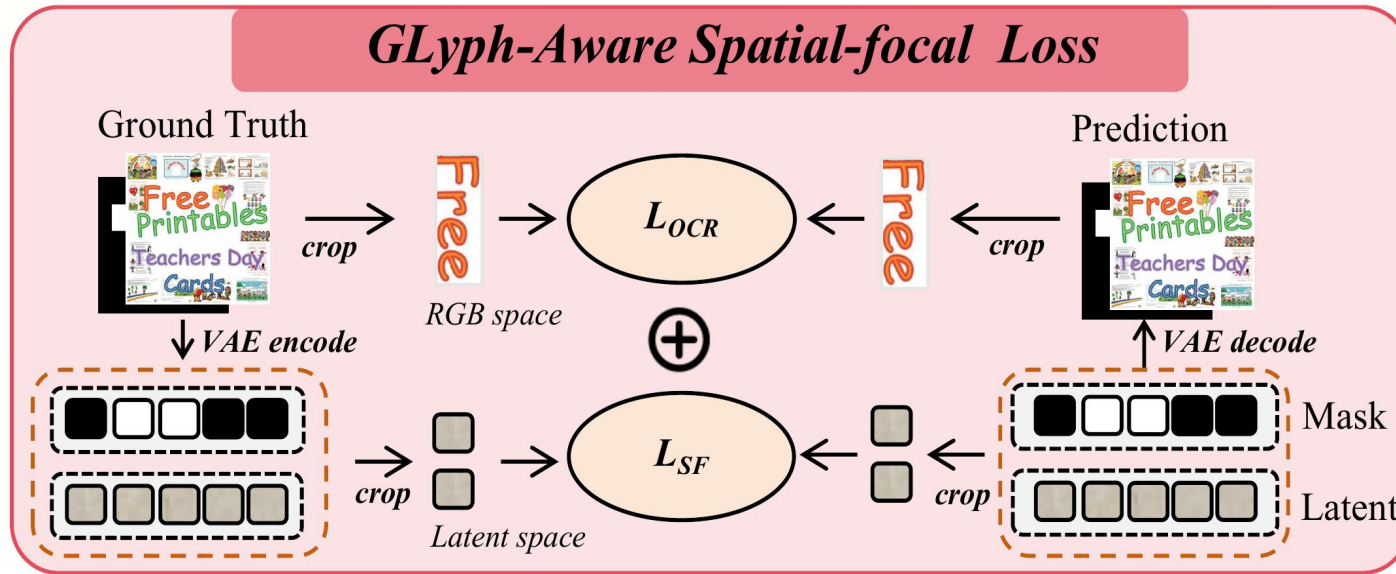
## Two glyph scales

Line-level features guide layout. OCR features preserve character strokes.

## Style stays separate

Neutral-font glyphs encode content. The VLM supplies source style.

# GLAS supervises text regions and character structure



## Spatial focus in latent space

Mask-derived weights concentrate learning on the sparse text region instead of the dominant background.

## Glyph guidance in pixel space

Decode sampled frames. Match OCR features and apply CTC with the target transcription.

## Total loss = flow matching + GLAS

$$\text{GLAS} = 3.0 \times \text{spatial-focal} + 0.1 \times \text{OCR}$$

Small text regions need stronger supervision than global latent matching alone provides.

## Why combine them?

OCR favors recognizability. Spatial focus helps retain the source style. The joint loss balances both.

# SteerVTE-1M pairs synthetic precision with real styles

(a) SynthTE Dataset

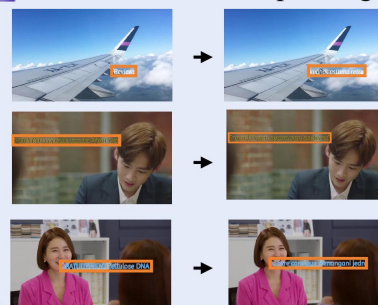
01 Text Configuration Sampling

|        | SynthTE-Easy | SynthTE-Hard        |
|--------|--------------|---------------------|
| Font   | Simple       | Complex, Artistic   |
| Color  | Black, White | Colorful            |
| Length | short, 1~20  | long, 1~30          |
| Size   | large, 30~50 | small, 25~40        |
| Effect | None         | stroke, bg, opacity |

02 Subtitle Texture Rendering

|     |                                 |
|-----|---------------------------------|
| Src | Reviews                         |
| Tgt | kulihat asthma retra            |
| Src | BURNING capsule tunda Wear Gai  |
| Tgt | Ingat! Umen seumururan acelerad |
| Src | GRATUITEMENT cellulose DNA      |
| Tgt | dintre consigue Menangani jedn  |

03 Random Video Compositing



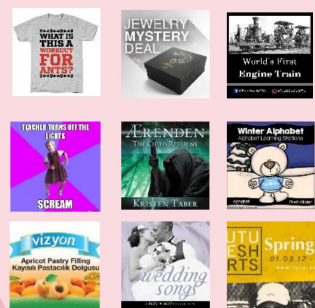
Hard: 200K videos

04 Ocr Filtering



(b) RealTE Dataset

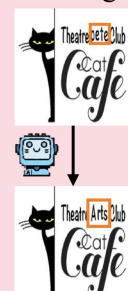
01 Text Image Dataset



02 Pre-OCR Filtering



03 Text Editing



300K images

04 Post-OCR Filtering



(c) Training Triplet

Source Video



Mask Video



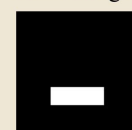
Target Video



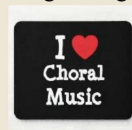
Source Image



Mask Image



Target Image



## 700K synthetic pairs

500K easy images + 200K hard videos. 1,933 fonts cover different layouts and effects.

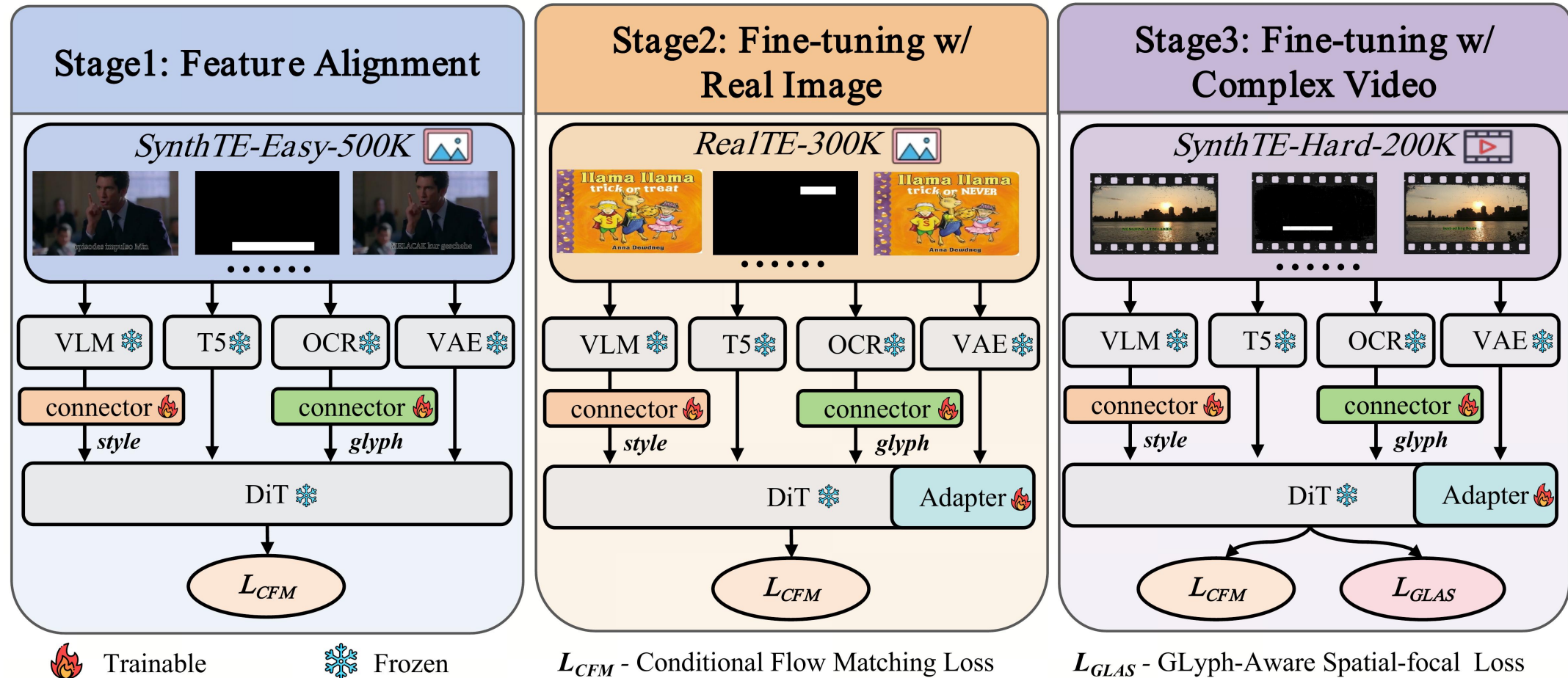
## Synthetic precision

Paired rendering fixes the background. Keep only samples with 100% OCR recognition.

## 300K real image pairs

OCR and style filtering retain diverse real typography. These images bridge the domain gap.

# Image pretraining prepares the model for video editing



## 1 Feature alignment

Connectors learn glyph and style features.

## 2 Real-world appearance

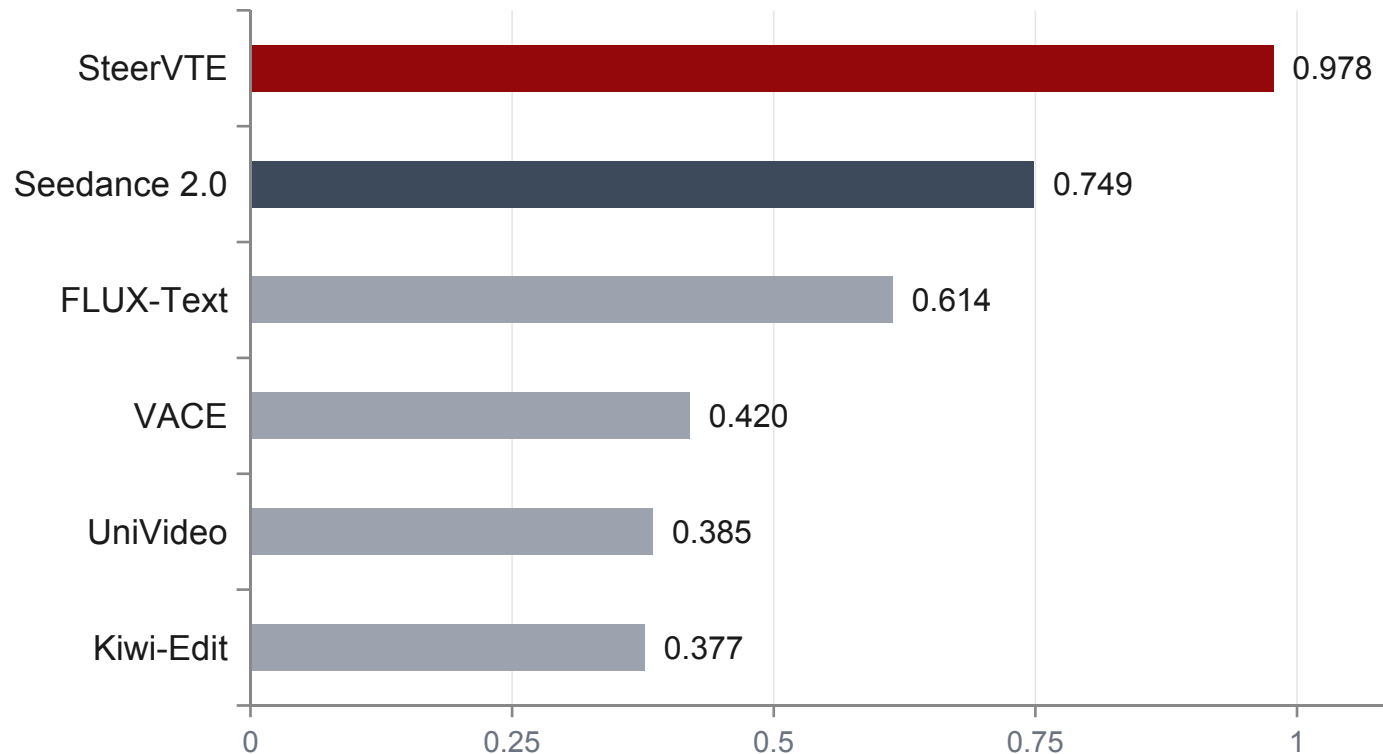
Adapter LoRA learns complex typography.

## 3 Temporal consistency

Video training adds GLAS supervision.

# 77% sentence accuracy versus 32% for Seedance 2.0

Real split: 100 videos. F-Score measures OCR text editing quality.



F-Score (higher is better). Values rounded to three decimals.

VTE-Bench: 100 Synth + 100 Real + 200 SceneText videos.

## 77% vs 32%

Sentence accuracy requires every character to match the target.

## 0.965 SSIM-bg

Similarity outside the edited region measures background preservation.

## 71% vs 23%

Quality rate: videos judged visually satisfactory by human raters.

# Curriculum and GLAS address different failure modes

## Training schedule on VTE-Bench-Synth

| Schedule          | Accuracy   | FVD ↓        |
|-------------------|------------|--------------|
| Video only        | 37%        | 17.88        |
| Mixed data        | 53%        | 15.49        |
| <b>Curriculum</b> | <b>61%</b> | <b>14.55</b> |

## Loss ablation in the video training stage

| Loss terms           | Accuracy   | Style-Sim ↑   |
|----------------------|------------|---------------|
| Flow matching        | 51%        | 0.5268        |
| + OCR only           | 65%        | 0.4794        |
| + Spatial-focal only | 56%        | 0.5691        |
| <b>+ Both (GLAS)</b> | <b>61%</b> | <b>0.5936</b> |

### A staged schedule helps

Image pretraining builds text priors. Video fine-tuning then improves temporal consistency.

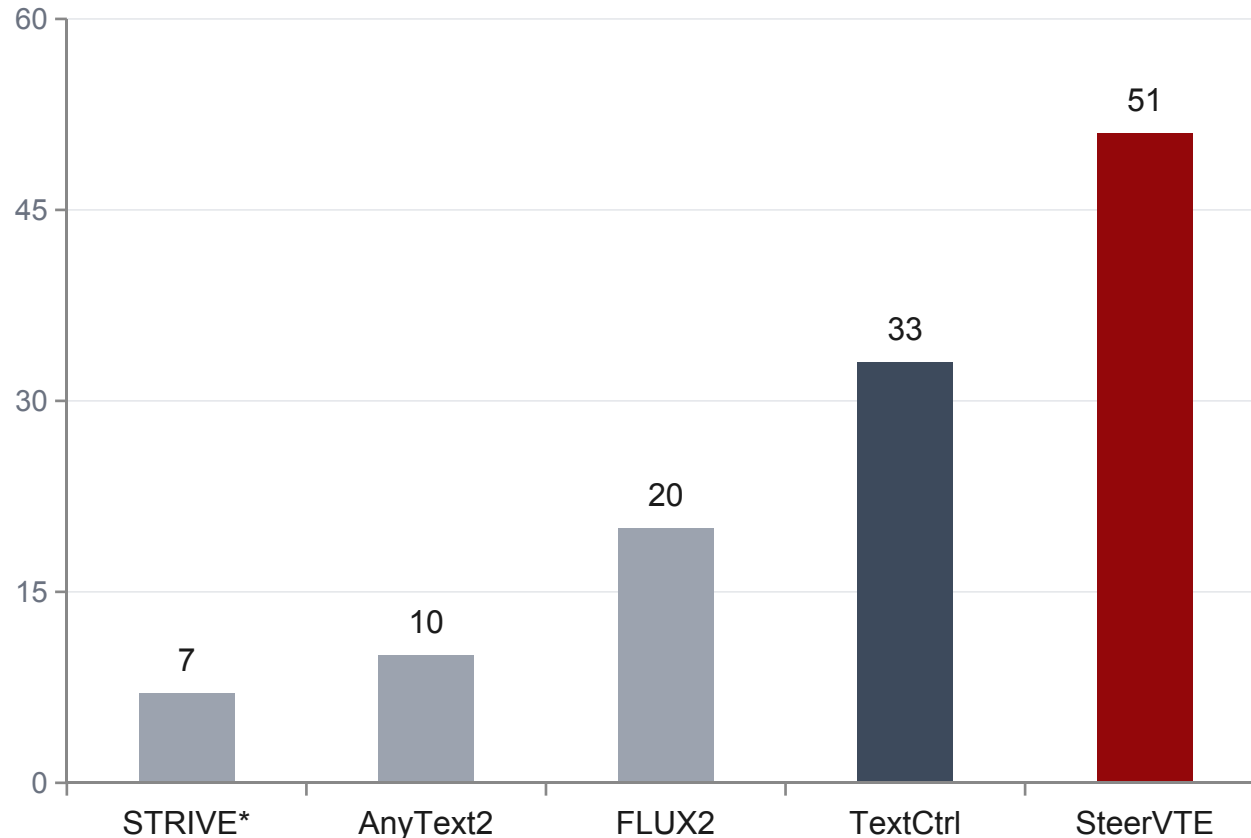
### OCR alone has a trade-off

Recognition rises to 65%, but style similarity drops to 0.4794.

### The joint objective balances both

GLAS reaches 61% accuracy with the strongest style similarity in this loss ablation.

# Scene-text accuracy reaches 51% on unseen videos



Sentence accuracy (%) on 200 SceneText videos.  
STRIVE\* denotes the reproduced implementation.

## No paired scene-text video training

Training uses synthetic subtitle videos and real text images. The test set adds text attached to scene surfaces.

## 18-point accuracy gain

Sentence accuracy: 51% vs 33% (TextCtrl).

## 41% quality rate

TextCtrl: 6%. Higher recognizability alone does not guarantee a satisfactory video.

# Sentence editing across three video frames

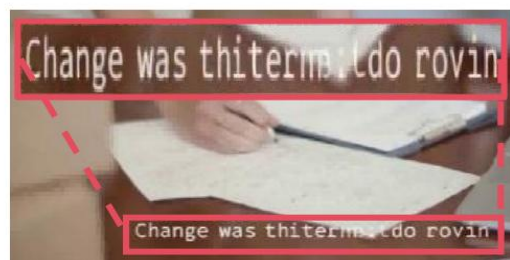
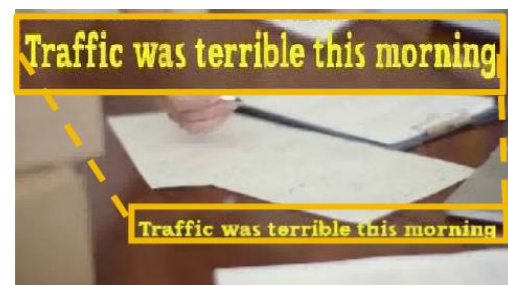
**Target text: “Traffic was terrible this morning”**

**Source**

**FLUX-Text**

**Seedance 2.0**

**SteerVTE**



Inspect character stability across frames and unintended changes outside the target region.

# Short-text editing with background preservation

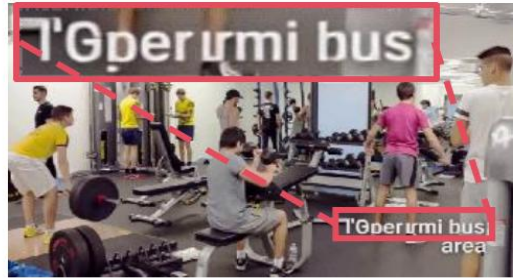
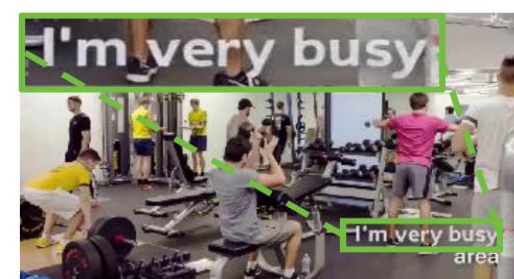
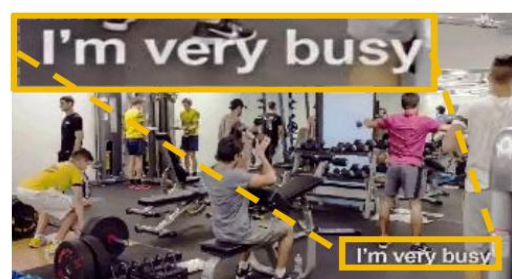
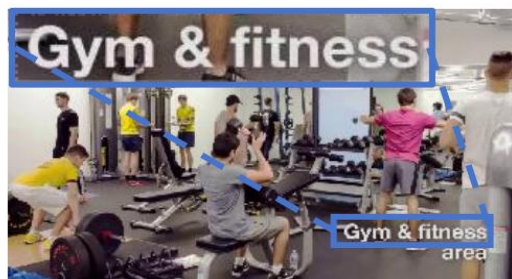
Target text: "I'm very busy"

Source

FLUX-Text

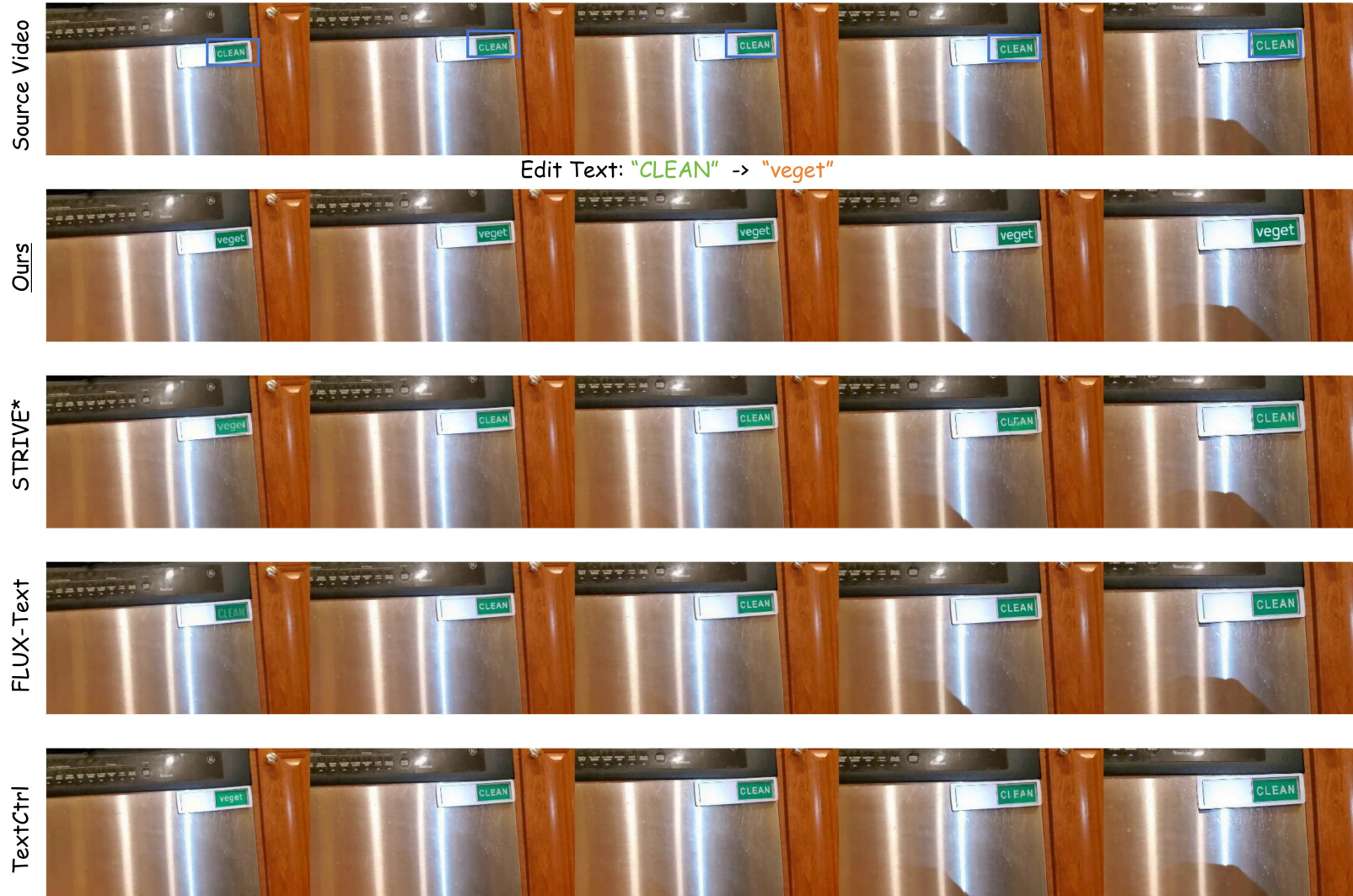
Seedance 2.0

SteerVTE



Compare the target caption and the surrounding people, equipment and background text.

# Text attached to a moving scene surface



## Editing target

"CLEAN" becomes "veget".

## What to inspect

Follow the sign across five frames. Check its perspective, placement and surrounding surface.

## Comparison rows

Source, SteerVTE, STRIVE\*, FLUX-Text and TextCtrl. All original rows are retained.

# Scene text under camera and subject motion



## Editing target

"London2012" becomes  
"CAPARISON".

## What to inspect

Follow the banner across five frames. Check text stability while the player and camera move.

## Comparison rows

Source, SteerVTE, STRIVE\*, FLUX-Text and TextCtrl. All original rows are retained.

## TAKEAWAYS

# Precise text control within a video prior

## Explicit text control

Mask, source style and dual-granularity glyphs steer a frozen video DiT.

## Training for accuracy and consistency

A 1M-triplet curriculum and GLAS loss support glyph fidelity, source style and temporal consistency.

**77% Real accuracy. 51% SceneText accuracy.**

Current scope: English and single-line text. Long-video editing remains limited.



[zengkaiya.github.io/SteerVTE](https://zengkaiya.github.io/SteerVTE)