# Video Editing via Factorized Diffusion Distillation

Uriel Singer* · Amit Zohar* · Yuval Kirstain · Shelly Sheynin · Adam Polyak · Devi Parikh · Yaniv Taigman

* Equal Contribution.

## Highlights

We present Emu Video Edit (EVE), a video editing model setting a new state-of-the-art without relying on any supervised video editing data.

‘add two birds around it’ ‘remove the guitar’

Our approach:

- Train two adapters on top of the same text-to-image model: an **image editing** adapter and a **video generation** adapter.
- Attach the adapters to the T2I and align them using **Factorized Diffusion Distillation**.
- The resulting model sets a new SOTA and supports numerous video editing operations like local, global, style and background changes.

## Method

Video editing requires two main capabilities:

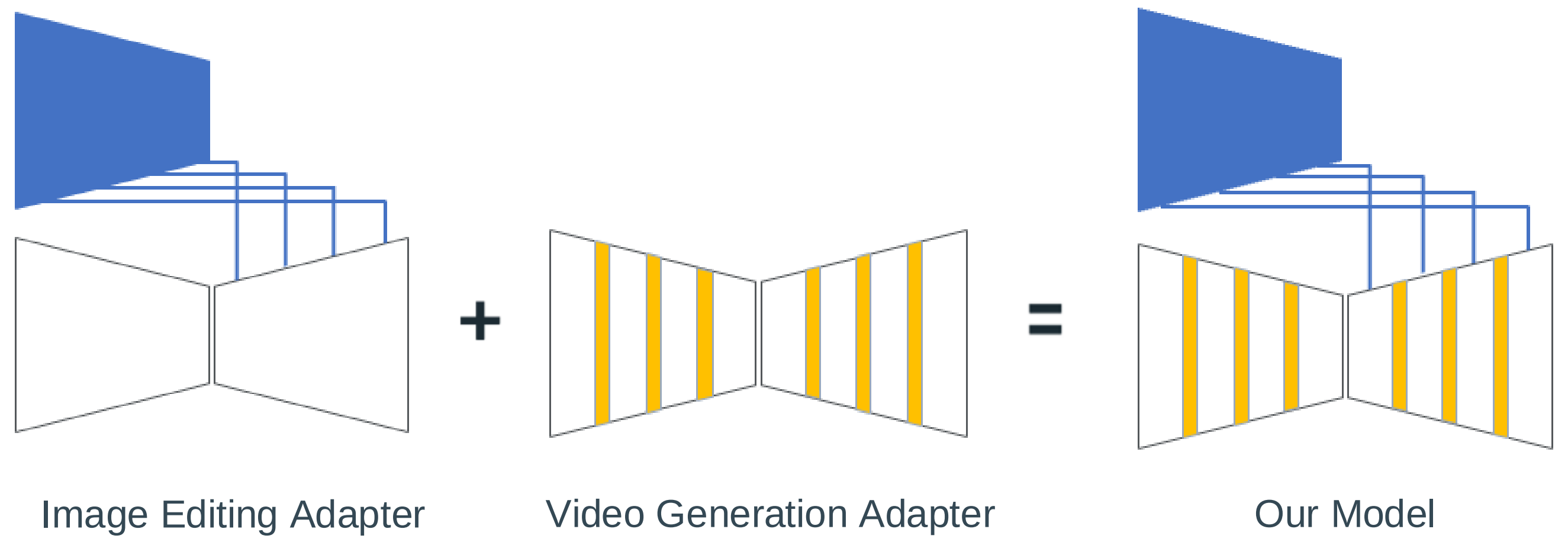
1. Precisely editing images.
2. Ensuring temporal consistency among frames.

We train a dedicated adapter on top of the **same** T2I for each capability:

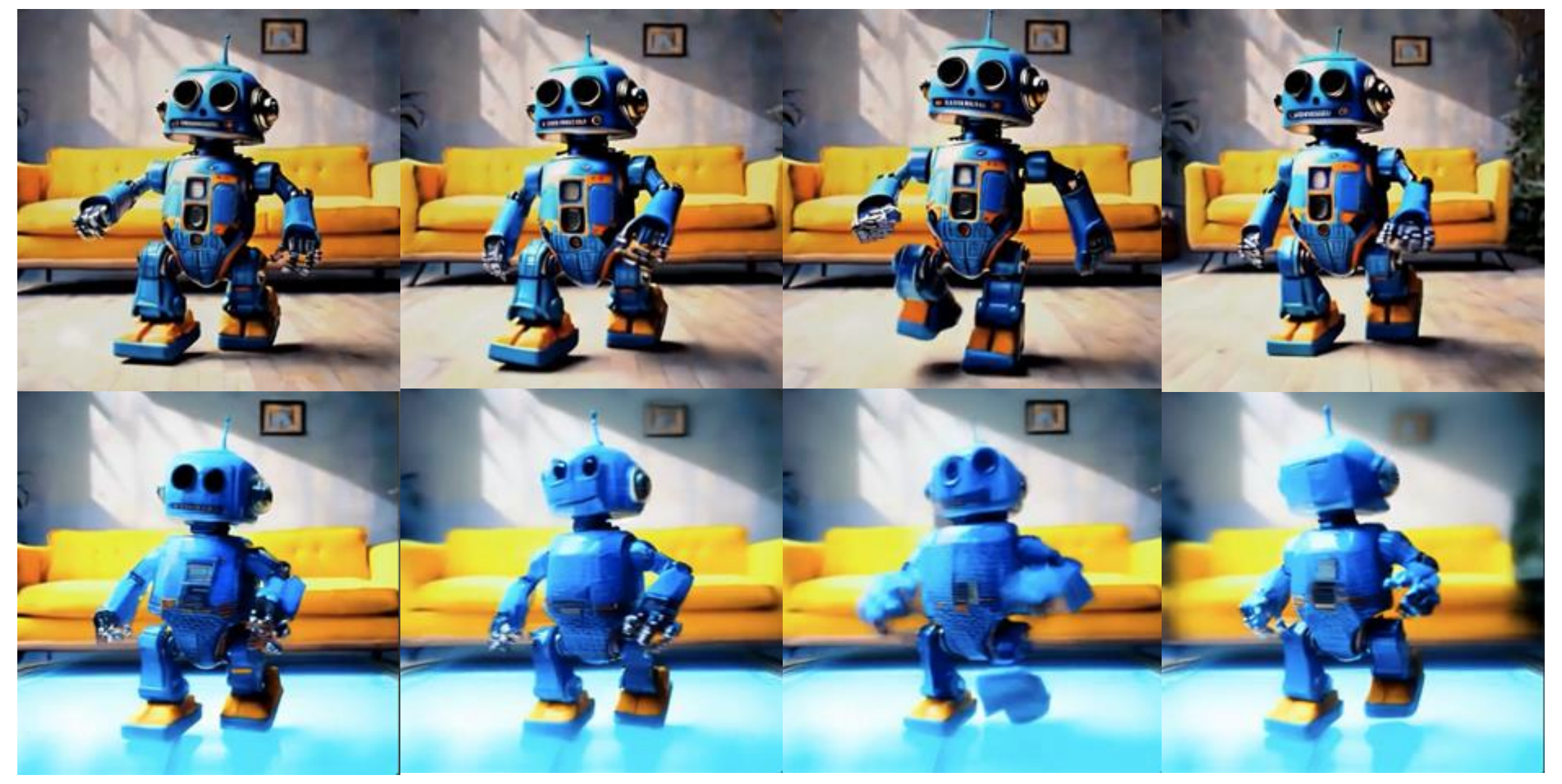
1. Image editing adapter
   - ControlNet on **Emu Edit**’s dataset.
2. Video generation adapter
   - Temporal layers on top of a frozen T2I model (like **Emu Video**)

If we attach both adapters simultaneously, we can perform video-editing:

- The image editing adapter edits each frame individually.
- The video generation adapter maintains temporal consistency.



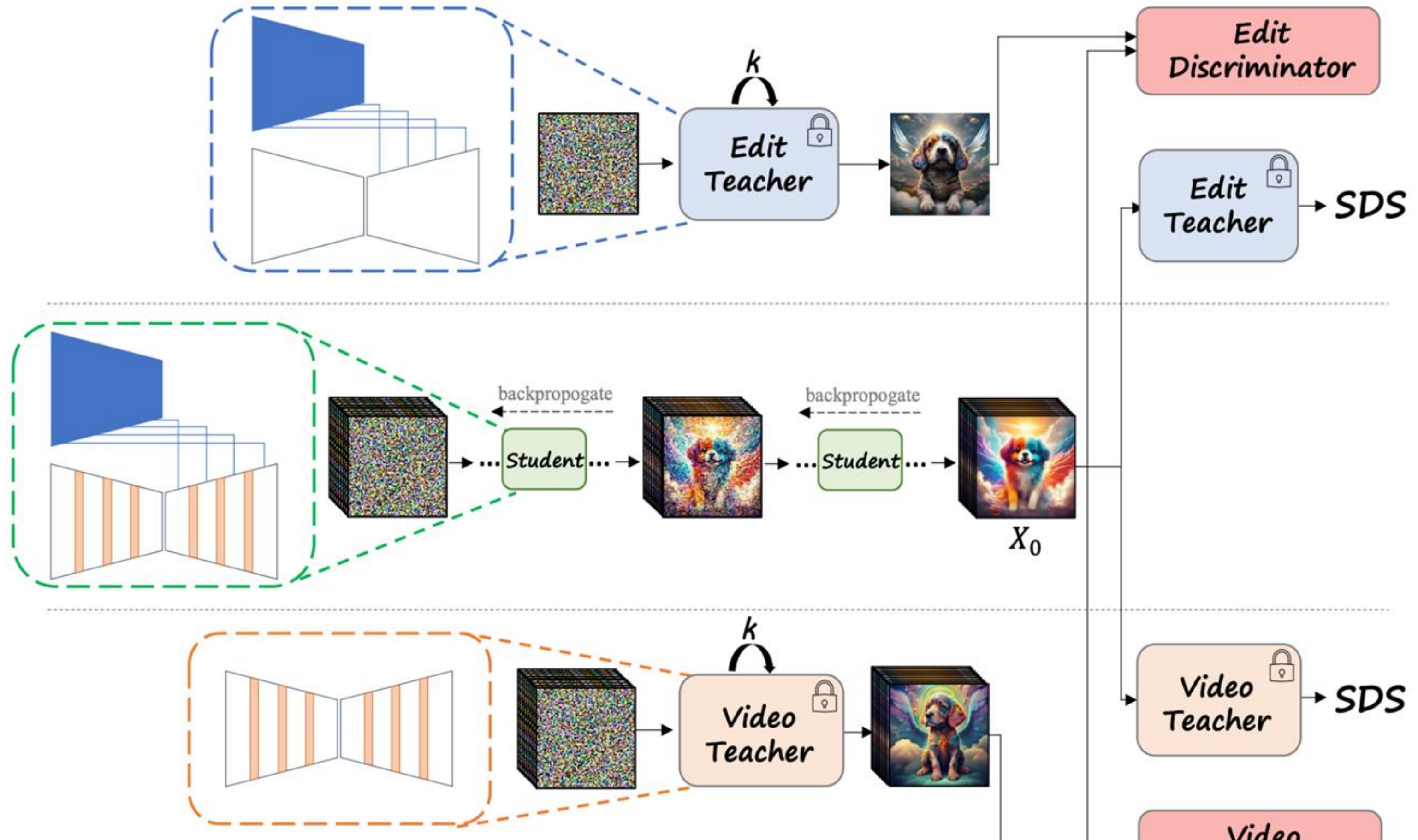
Even though the adapters use the same frozen T2I, combining them causes **severe artifacts**:


‘turn the floor into glass`

## Factorized Diffusion Distillation (FDD)

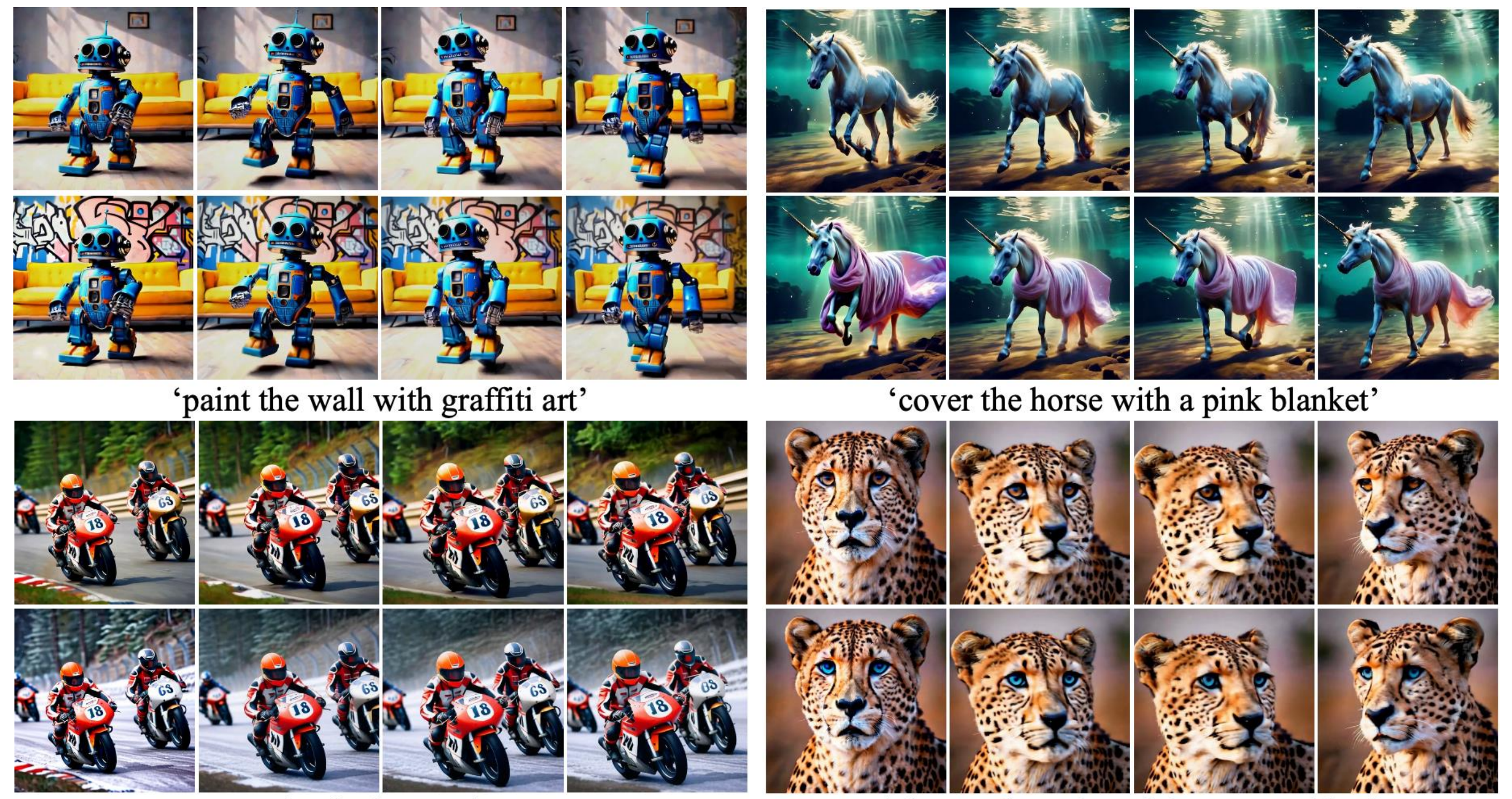
An unsupervised alignment procedure which solves the artifacts.

- We train LoRA weights over the T2I, **keeping the adapters frozen.**
- The model edits a video from **pure noise**.
- We perform Score Distillation Sampling, obtaining feedback from each adapter:
  - Image Editing adapter on **edit faithfulness** per frame.
  - Video Generation adapter on the video’s **temporal consistency**.
- To prevent blurriness, we add an adversarial loss for each teacher.



## Emu Video Edit (EVE)

- State-of-the-art in text-based video editing.
- Supports all 16 tasks that Emu Edit does for images:
  - Local & global changes.
  - Style & background operations.
  - Computer Vision Tasks.


‘paint the wall with graffiti art’ ‘cover the horse with a pink blanket’

‘make it snow’ ‘change the color of the eyes to blue’

EVE also supports Emu Edit tasks it wasn’t aligned on.

- Suggests the student aligns with the entire knowledge of the teacher

input
‘make a sketch of this video.’

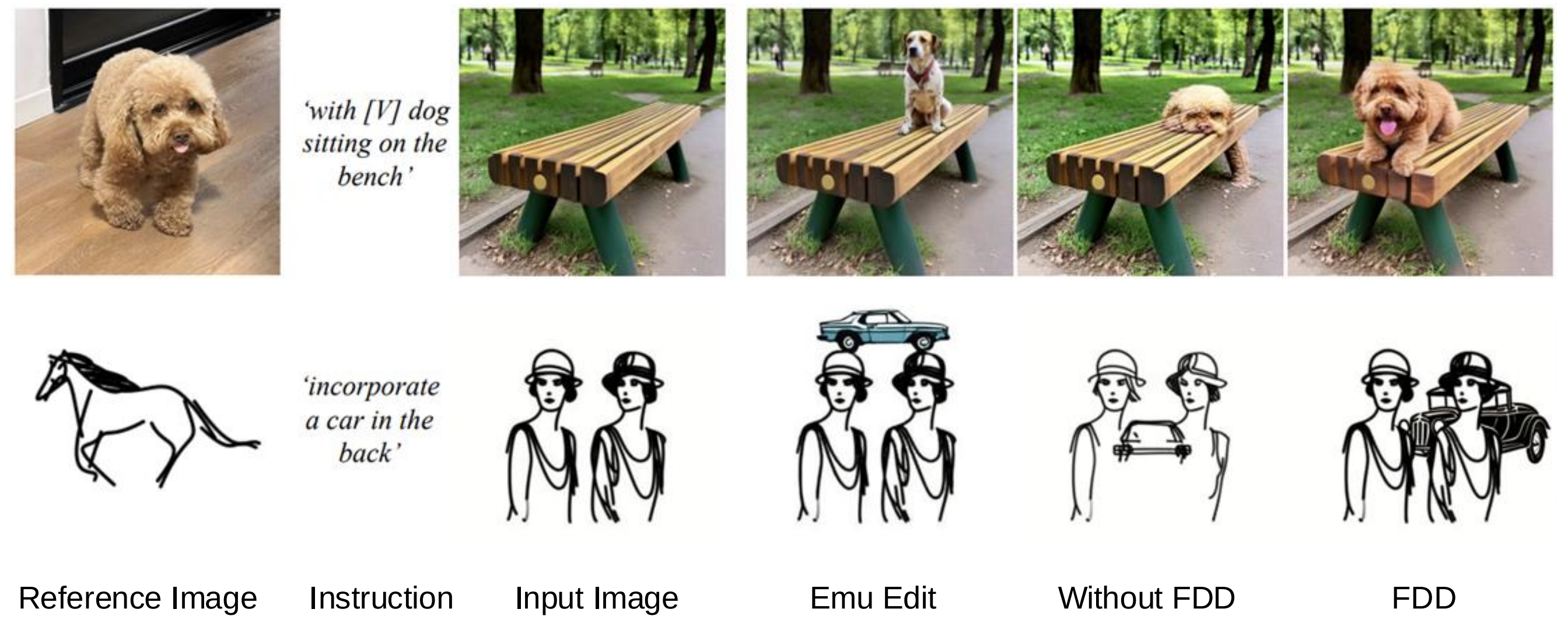
## Evaluation

Comparison with baselines on TGVE.

| Method | PickScore↑ | CLIPFrame↑ | ViCLIP$_{dir}$ ↑ | ViCLIP$_{out}$ ↑ | Text | Struct. | Quality | Avg. |
|---|---|---|---|---|---|---|---|---|
| TAV [37] | 20.36 | 0.924 | 0.162 | 0.243 | 72.4 | 74.0 | 85.2 | 77.2 |
| SDEdit [24] | 20.18 | 0.896 | 0.172 | 0.253 | 75.7 | 67.4 | 79.0 | 74.0 |
| STDF [42] | 20.40 | **0.933** | 0.110 | 0.226 | 81.3 | 65.8 | 70.1 | 72.4 |
| Fairy [36] | 19.80 | **0.933** | 0.164 | 0.208 | 77.3 | 62.8 | 75.0 | 71.7 |
| InsV2V [6] | **20.76** | 0.911 | 0.208 | **0.262** | 57.9 | 55.9 | 65.1 | 59.6 |
| EVE (Ours) | **20.76** | 0.922 | **0.221** | **0.262** | – | – | – | – |

## Additional Adapter Combinations

- FDD can be applied to **any group** of adapters
- We apply FDD on image editing + personalization/style LoRA
  - Enables personalized image editing & stylized image editing


‘with [V] dog sitting on the bench’
‘incorporate a car in the back’
Reference Image
Instruction
Input Image
Emu Edit
Without FDD
FDD

## TGVE+ Benchmark

- An extension of the TGVE benchmark.
- We add three new tasks:
  - Adding objects
  - Removing objects
  - Changing the color/texture of objects
- **TGVE+ is publicly available on HF datasets.**

‘add a large fireworks display.’ ‘change lotus to poppy flower.’

Ablation study on our different contributions.

| Method | Text | Struct. | Quality | Avg. |
|---|---|---|---|---|
| Random Init | 96.7 | 70.1 | 94.7 | 87.2 |
| w/o alignment | 77.6 | 91.4 | 89.8 | 86.3 |
| w/o SDS | 77.6 | 87.5 | 92.1 | 85.7 |
| w/o discriminators | 74.3 | 84.2 | 83.9 | 80.8 |
| w/o K-Bin Sampling | 57.6 | 49.7 | 51.6 | 53.0 |

Check out the Project Page!