DCM: Dual-Expert Consistency Model for Efficient and High-Quality Video Generation

Zhengyao Lv[2,3*] Chenyang Si[1*] Tianlin Pan[1,4] Zhaoxi Chen[5] Kwan-Yee K. Wong[2] Yu Qiao[3] Ziwei Liu[5]
NJU[1] HKU[2] Shanghai Artificial Intelligence Laboratory[3] UCAS[4] NTU[5]

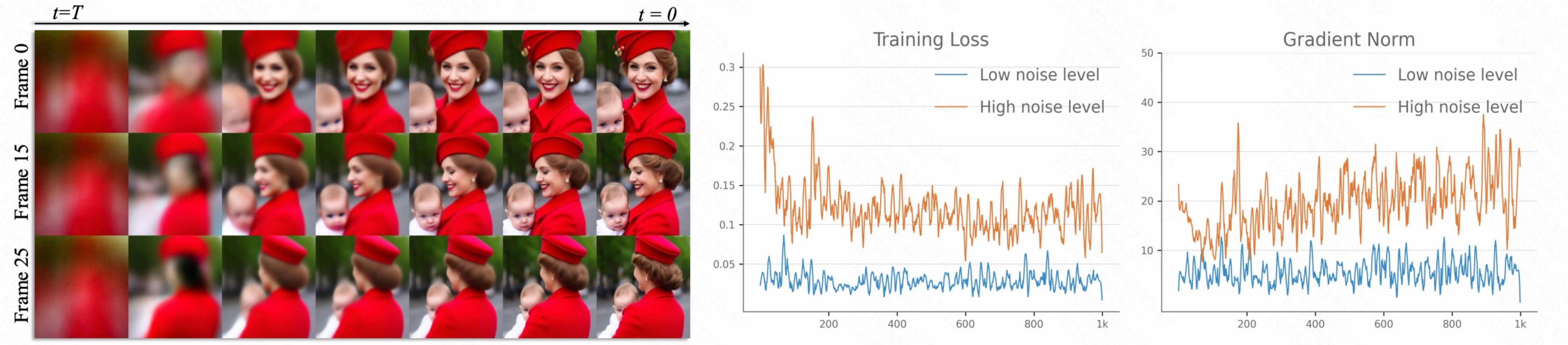
ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation

Task: In this paper, we propose **a parameter-efficient Dual-Expert Consistency Model (DCM)**, maintain visual quality with significantly reduced sampling steps, **down to as few as four steps**.

Motivation:

- Differences between adjacent steps are **substantial in early sampling stages**, while the changes become **more gradual in later stages.**
- The **magnitude of loss and gradient during distillation differ markedly between different noise levels**, suggesting that distilling a single student to capture both semantic layout and fine-detail synthesis may cause optimization interference and yield suboptimal results.

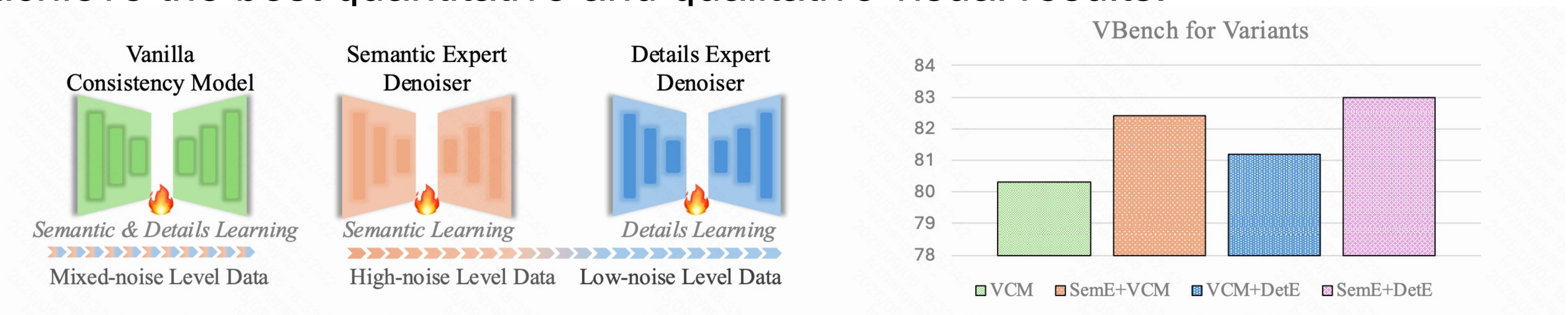


(a) Visualization of the video synthesis sampling process (b) The trend of training loss and gradient norm

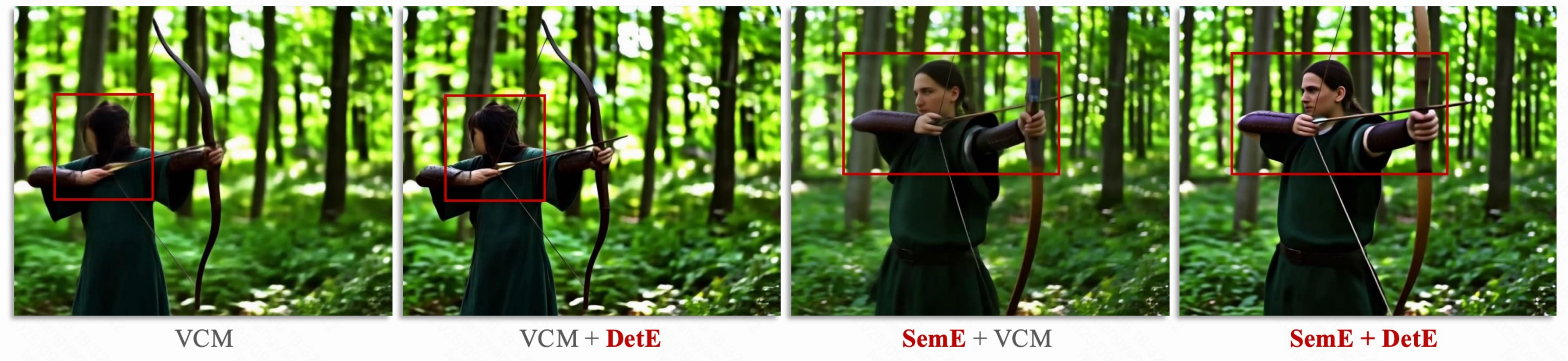
Investigation

Suboptimal Solution in Consistency Distillation

- Comparison of results of denoiser variants trained at different noise level samples.
- By **optimizing two expert denoisers to decouple the distillation process into semantic learning and detail learning**, and combining them during inference, we achieve the best quantitative and qualitative visual results.



(a) Variants trained on data with different noise levels (b) VBench metric for different variants

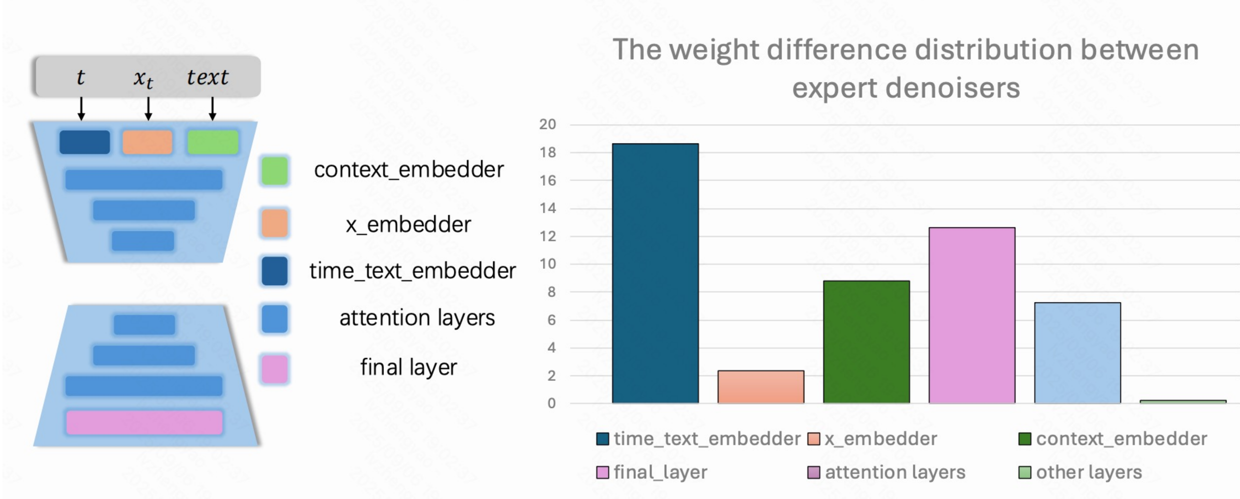


(c) Example visual results of different variants

Method

Parameter-efficient Dual-Expert Distillation

- While training two expert denoisers improves video quality, it significantly increases model parameters and GPU memory consumption during inference.
- We found that the primary differences in model parameters lie in the embedding layers and the linear layers within the attention layers.



- Based on the above observations, we propose the parameter-efficient Dual-Expert distillation strategy.

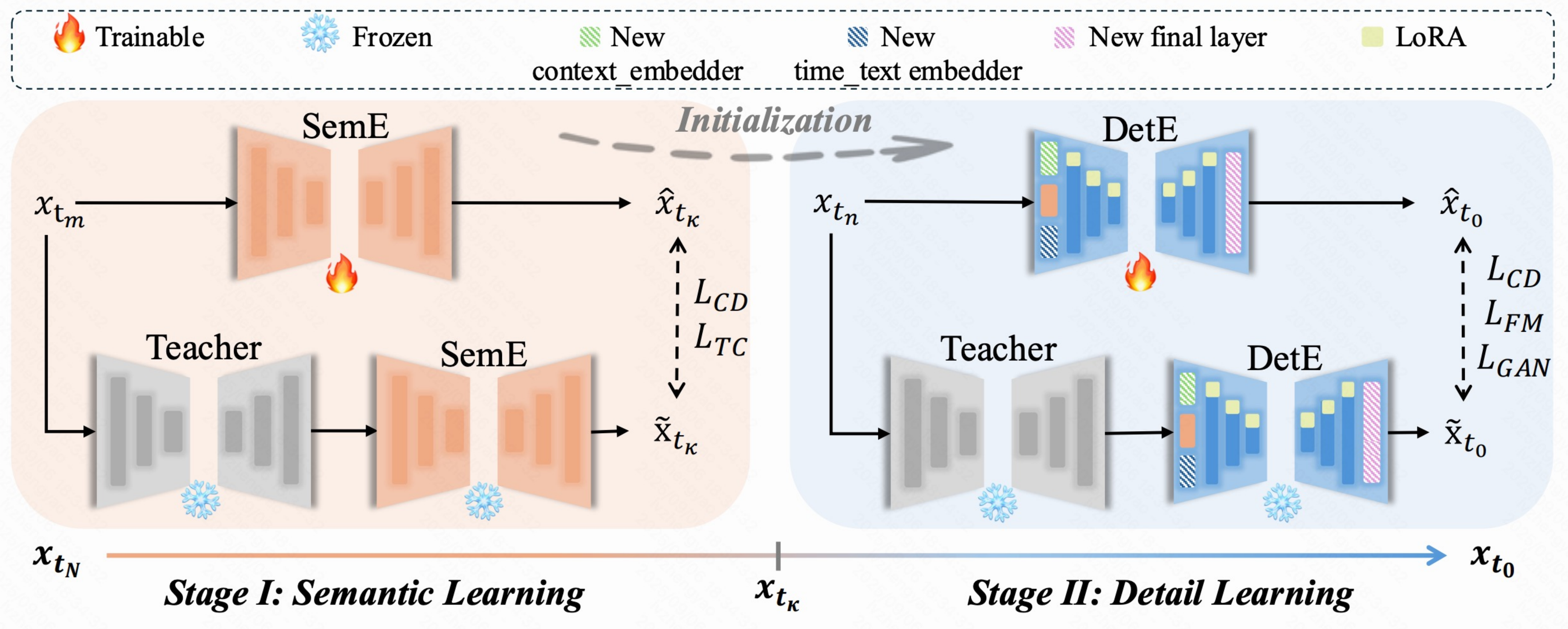


Figure 4. The training process of DCM consists of two stages. In the semantic learning stage, we train SemE on high-noise samples with consistency loss and temporal coherence loss as the learning objectives. In the detail learning stage, we initialize DetE with the weights of SemE and introduce a set of time-dependent layers and LoRA. DetE is then trained on low-noise samples, where **only the newly added**

Expert-specific Optimization Objective

- Temporal Coherence Loss for SemE

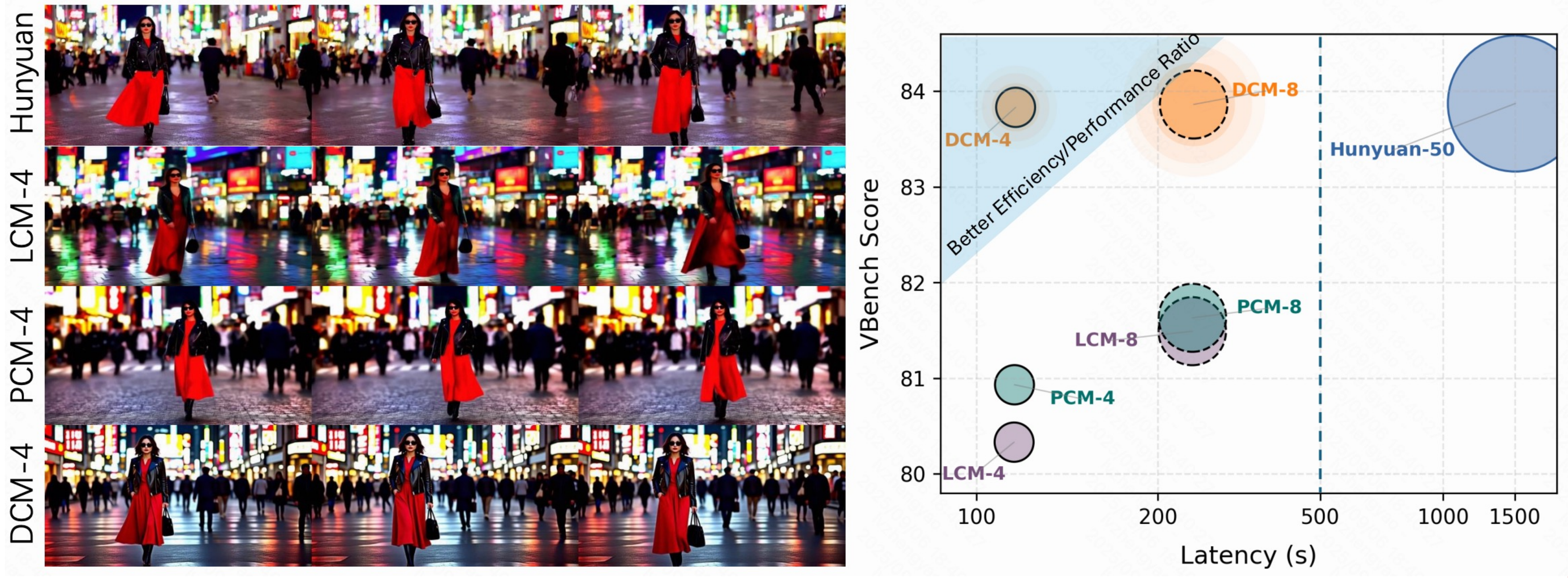
$$\boldsymbol{x}_{t_\kappa} = \Phi(\boldsymbol{x}_{t_m}, F_{\mathrm{SemE}}(\boldsymbol{x}_{t_m}, t_m, c), t_\kappa),$$
$$\hat{\boldsymbol{x}}_{t_\kappa} = \Phi(\hat{\boldsymbol{x}}_{t_{m-1}}, F^{-}_{\mathrm{SemE}}(\hat{\boldsymbol{x}}_{t_{m-1}}, t_{m-1}, c), t_\kappa),$$
$$\mathcal{L}_{TC} = ||(\boldsymbol{x}^{t_\kappa}_{l:L} - \boldsymbol{x}^{t_\kappa}_{0:L-l}) - (\hat{\boldsymbol{x}}^{t_\kappa}_{l:L} - \hat{\boldsymbol{x}}^{t_\kappa}_{0:L-l})||^2_2.$$

- Generative Adversarial and Feature Matching Loss for DetE

$$\mathcal{L}_{FM} = \mathbb{E}_{\boldsymbol{x},t_n} \|\Omega(\boldsymbol{x}_{fake}) - \Omega(\boldsymbol{x}_{real})\|^2_2,$$
$$\mathcal{L}_G = \mathbb{E}_{\boldsymbol{x},t_n}[1 - f_D(\Omega(\boldsymbol{x}_{fake}))] + \mathcal{L}_{FM},$$
$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{x},t_n}[f_D(\Omega(\boldsymbol{x}_{fake}))] + \mathbb{E}_{\boldsymbol{x},t_n}[1 - f_D(\Omega(\boldsymbol{x}_{real}))].$$

Results

Quantitative Comparison & Qualitative Comparison



Ablation Study

Figure 8. Impact of temporal coherence loss.

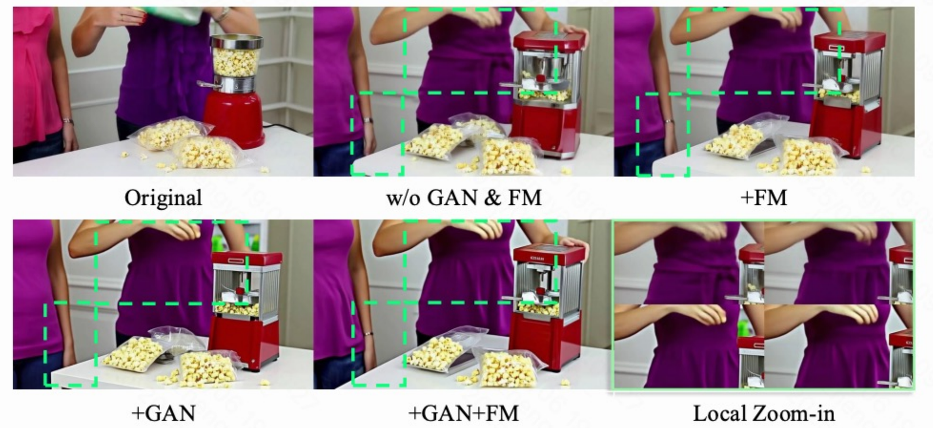


Figure 9. Impact of the GAN loss and Feature Matching term.

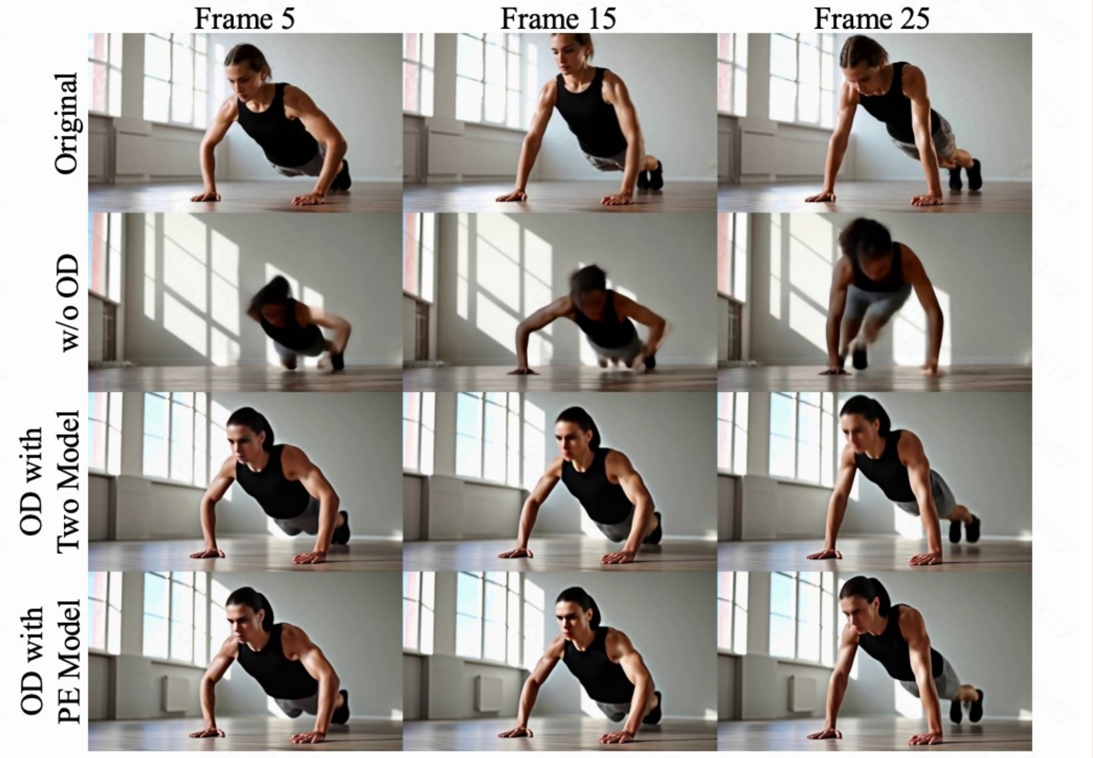


Figure 7. Impact of optimization decoupling and parameter-efficient distillation.

For More Detail

Project Page GitHub Repo Paper