VADER: Video Diffusion Alignment via Reward Gradient

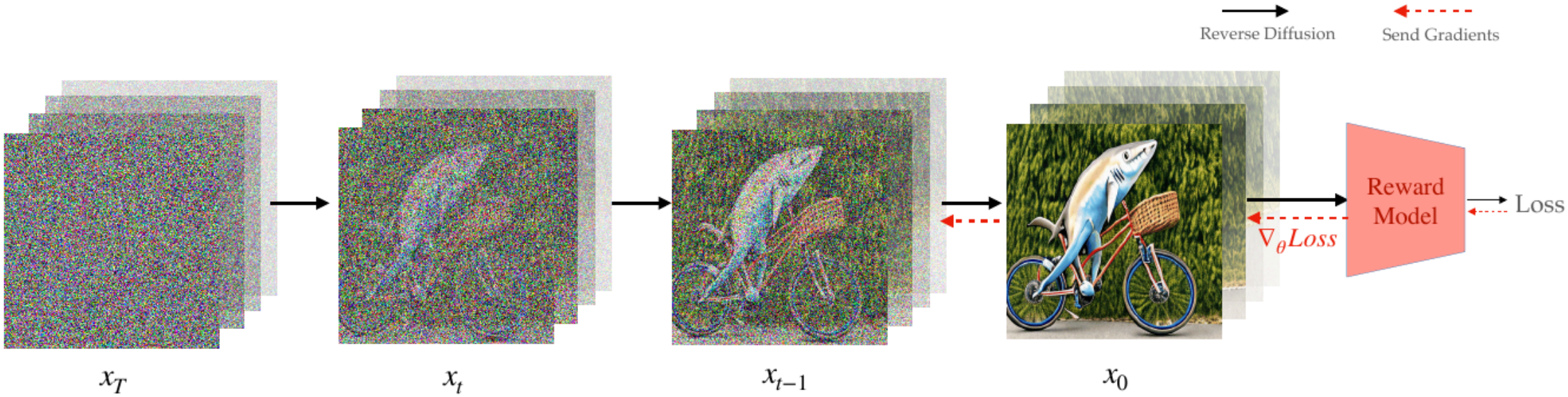


Figure 2: VADER aligns various pre-trained video diffusion models by backpropagating gradients from the reward model, to efficiently adapt to specific tasks.

Pseudo Code

Algorithm 1 VADER

Require: Diffusion Model weights θ
Require: Reward function $R(\cdot)$
Require: Denoising Scheduler f
(eg - DDIM, EDM)
Require: Gradient cutoff step K

while training **do**
  for t = T,..,1 **do**
    pred = $\epsilon_\theta(x_t, c, t)$
    if t $>$ K **then**
      pred = stop_grad(pred)
    end if
    x_{t-1} = f.step(pred, t, x_t)
  end for
  $g = \nabla_\theta R(x_0, c)$
  $\theta \leftarrow \theta - \eta * g$
end while

Pseudo Code

Algorithm 1 VADER

Require: Diffusion Model weights θ
Require: Reward function $R(\cdot)$
Require: Denoising Scheduler f
(eg - DDIM, EDM)
Require: Gradient cutoff step K

while training **do**
  for t = T,..,1 **do**
    pred = $\epsilon_\theta(x_t, c, t)$
    if t $>$ K **then**
      pred = stop_grad(pred)
    end if
    x_{t-1} = f.step(pred, t, x_t)
  end for
  $g = \nabla_\theta R(x_0, c)$
  $\theta \leftarrow \theta - \eta * g$
end while

truncated backpropagation

Image-Level Reward Model

1. Score Model (e.g., Human Preference Score, PickScore, LAION Aesthetics predictor)
2. Object-Detection Model (YOLO)

These reward models evaluate **each frame**, not a whole video.

Video-Level Reward Model

1. Action Classifier (VideoMAE)
2. Masked Frame Predictor (V-JEPA)

Masked frame predictor evaluates the consistency of the video generation.

Example 1: HPS + Aesthetic

VideoCrafter

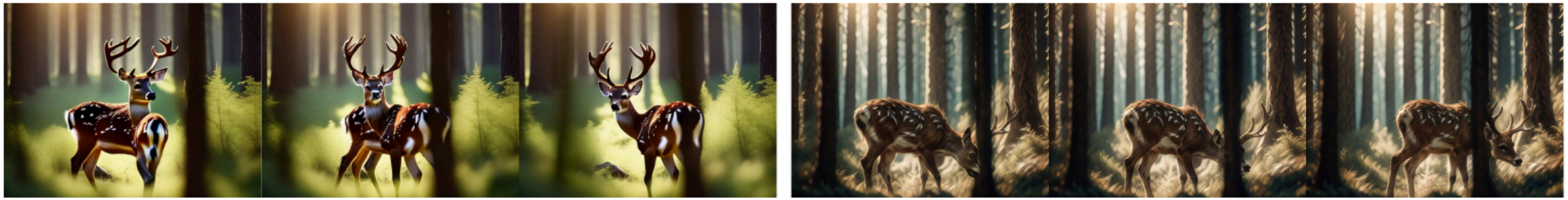
VADER (Ours)

"The raccoon is wearing a red coat and holding a snowball."

"The fox is wearing a red hat and playing with leaves."

Example 2: PickScore

“A strong lion and a graceful lioness resting together in the shade of a big tree on a wide grassland.”



“A peaceful deer eating grass in a thick forest, with sunlight filtering through the trees.”

Example 3: YOLO

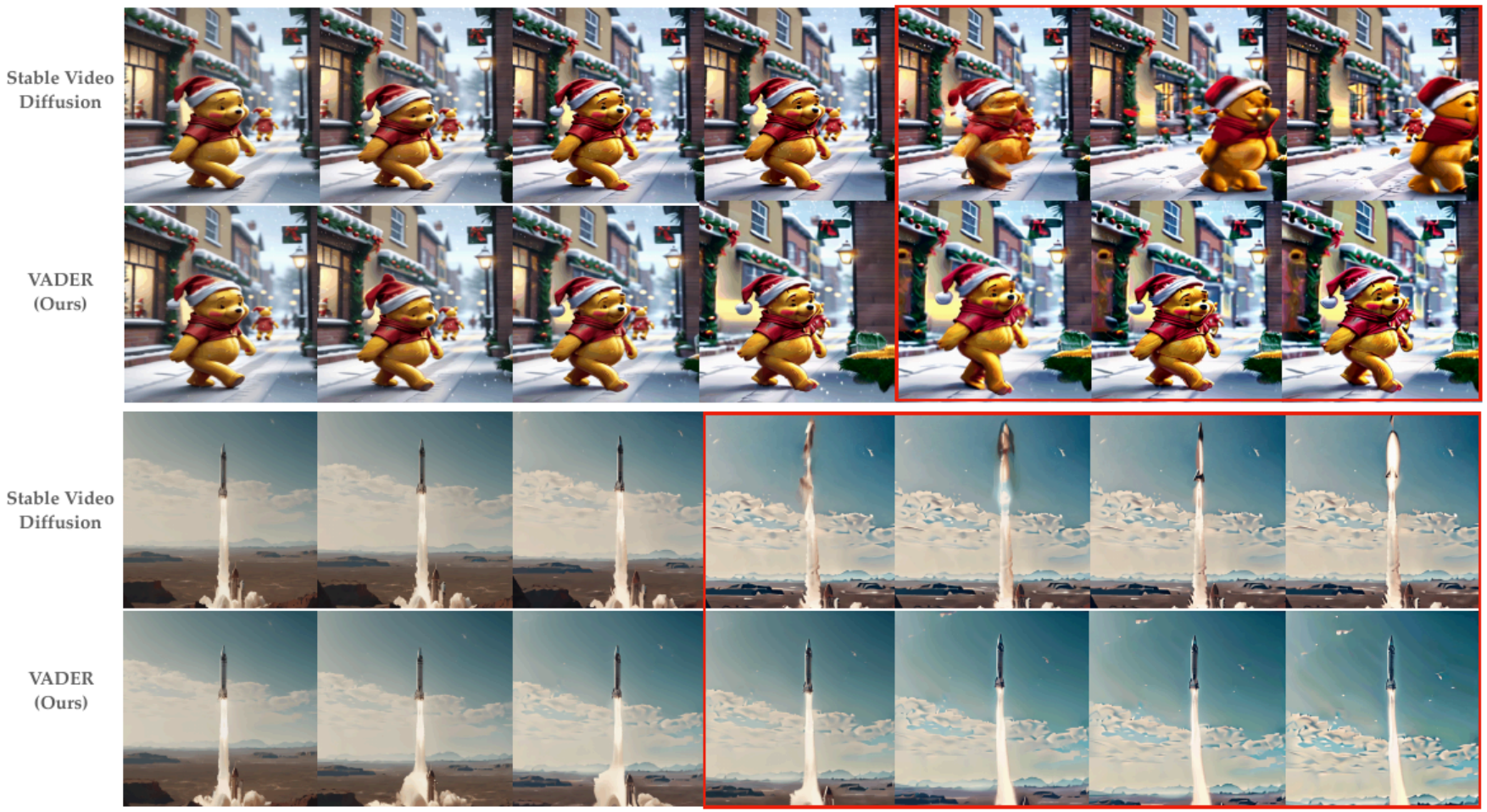
Before | VADER (Ours)

"A book and a cup of tea on a blanket in a sunflower field."

"A book and a cup of hot chocolate on a windowsill with a snowy view."

"A book and a cup of coffee on a rustic wooden table in a cabin."

Example 4: V-JEPA



Data Efficiency & Computational Efficiency