E^{2}PO: CONCEPT ERASING IN TEXT-TO-VIDEO MODELS VIA ENERGY-BASED ERASING PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern Text-to-Video (T2V) models generate increasingly realistic and coherent videos, but they also risk producing undesired content such as nudity, personal identities, or unsafe objects. Existing concept erasure methods, largely developed for Text-to-Image models, are difficult to transfer directly to T2V generation because undesired concepts can reappear through diverse appearances, motions, viewpoints, and temporal compositions. In this paper, we propose E^{2}PO, an energy-based preference optimization framework for T2V concept erasure. Instead of relying on local suppression of limited samples, E^{2}PO formulates erasure as energy-based preference shaping over the open-ended video generation space. We introduce an Energy-based Erasing Preference and instantiate it with an energy discrepancy objective that enlarges the energy gap between desired and undesired generations. To mitigate catastrophic forgetting caused by broad energy suppression, we further introduce a Manifold-guided Target Anchoring strategy, which redirects undesired prompts toward nearby safe semantic anchors while preserving spatiotemporal generation priors. Extensive experiments demonstrate that E^{2}PO effectively suppresses undesired concepts and better preserves video generation quality. Anonymous code is available at https://anonymous.4open.science/r/E-2PO-AnonymousRelease.

Warning: This paper may contain content that may seem offensive in nature.

1 INTRODUCTION

Text-to-Video (T2V) generation has advanced rapidly in recent years, with representative models such as Seedance 2 and Veo 3 (ByteDance Seed, 2026; Google DeepMind, 2025) bringing T2V generation closer to practical content creation. However, these advances also raise critical safety concerns, as T2V models may generate undesired or harmful content involving nudity, violence, copyrighted characters, or specific individuals (Miao et al., 2024; Dai et al., 2024; Sun et al., 2024; Puspitasari et al., 2024). Therefore, concept erasure, which aims to suppress the generation of specified undesired concepts, has become an important problem for safe T2V generation.

Concept erasure has been extensively studied in Text-to-Image (T2I) generation (Gandikota et al., 2023; Park et al., 2024; Kim et al., 2024; Shin et al., 2025). Existing methods mainly rely on fine-tuning or editing specific model components, prompt-level intervention, or modification of text representations to suppress undesired concepts. However, directly transferring these strategies to T2V models is nontrivial, as illustrated in Figure 1. Compared with images, videos introduce an additional temporal dimension, where object appearance, motion trajectory, scene evolution, and background dynamics are strongly coupled. As a result, an undesired concept in T2V generation is not confined to a localized visual pattern or isolated text token. Instead, it may appear through diverse identities, viewpoints, actions, camera motions, and temporal compositions. This makes the harmful generation space open-ended and substantially harder to suppress.

Existing T2V erasure attempts still largely rely on local suppression signals, such as penalizing specific negative samples or modifying isolated text conditions (Ye et al., 2025; Facchiano et al., 2025; Xu et al., 2025). While such strategies can reduce certain explicit appearances of an undesired concept, they may fail to cover the broader spatiotemporal manifold where the concept can re-emerge.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

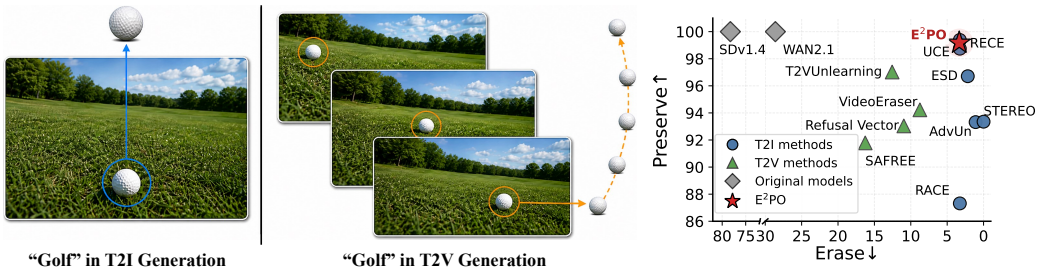


Figure 1: **Left:** comparison between the concept to be erased in T2I and T2V generation. In images, a concept usually appears as a static visual entity, while in videos it is coupled with temporal dynamics. **Right:** Performance of representative concept erasure methods. Compared with mature T2I erasure methods, existing T2V erasure methods still exhibit a weaker trade-off.

Moreover, aggressive suppression over a broad concept subspace can damage the original generation prior, leading to catastrophic forgetting, degraded visual fidelity, weakened temporal coherence, or loss of unrelated generation capabilities. These challenges suggest that T2V concept erasure requires a distribution-level formulation that can suppress open-ended harmful generations while maintaining the spatiotemporal structure of the original model.

In this work, we propose E^{2}PO, a preference-based framework for T2V concept erasure. We formulate concept erasure as energy-based preference shaping over the open-ended video trajectory space. We first introduce an *Energy-based Erasing Preference*, where generations satisfying the erasure objective are assigned low energy and undesired generations are assigned high energy. To make this preference tractable for T2V models, we optimize it with an offline energy discrepancy objective, without requiring access to the entire video generation space. This allows E^{2}PO to increase the energy gap between desired and undesired generations.

However, global energy suppression alone can still over-perturb the T2V manifold. To mitigate this issue, we further introduce a *Manifold-guided Target Anchoring* (MTA) strategy. Rather than simply pushing undesired generations away, we retrieve a nearby safe semantic anchor in the model's conditioning feature space and use it to construct desired samples. This turns negative suppression into directional energy redirection: the model learns not only what to suppress, but also where the suppressed probability mass should be redirected. As a result, E^{2}PO can better preserve appearance, motion, and temporal priors while erasing undesired concepts.

Our main contributions are threefold: **(i)** We formulate T2V concept erasure as an energy-based preference optimization problem and propose E^{2}PO, which reshapes the energy gap between desired and undesired generations over the open-ended video generation space. **(ii)** We introduce MTA, which redirects undesired prompts toward nearby safe semantic anchors, mitigating catastrophic forgetting and preserving spatiotemporal generation priors. **(iii)** Extensive experiments on nudity, identity, and object erasure demonstrate that E^{2}PO achieves strong erasure effectiveness, robust adversarial resistance, and better preservation of video generation quality.

2 Related Work

2.1 Video Generation Models

T2V generation initially followed the text-to-image diffusion paradigm with 3D U-Net backbones and cross-attention conditioning (Ho et al., 2022; Singer et al., 2023; Wang et al., 2023; Wu et al., 2023a). Recent progress is driven by Transformer-based diffusion models, most notably Diffusion Transformers (DiT) (Peebles & Xie, 2023), which scale more effectively to high resolution and longer temporal horizons. Representative models include CogVideo (Hong et al., 2023) and CogVideoX (Yang et al., 2025). Beyond backbone scaling, modern T2V models increasingly replace diffusion objectives with flow matching for improved efficiency and stability while retaining DiT backbones, as exemplified by HunyuanVideo (Kong et al., 2024), Open-Sora (Peng et al., 2025), and WAN (Wan et al., 2025).

Recent preference optimization methods (Kim et al., 2026; Wu et al., 2025) have further improved T2V generation through preference learning. E^{2}PO takes a fundamentally different step: it is the first

to formulate preference optimization in T2V as distribution reshaping for concept erasure, rather than ranking generated videos under a fixed preference model.

2.2 Concept Erasing

Concept erasing aims to prevent generative models from producing content associated with specific undesired concepts (e.g., unsafe, copyrighted, or private content) while preserving overall generation quality (Gandikota et al., 2023; Feng et al., 2025). Most existing studies focus on the text-to-image (T2I) setting, where methods can be broadly categorized into three paradigms. **Fine-tuning** approaches suppress target concepts by training the model parameters, either globally or on key components such as cross-attention to achieve stable removal (Gandikota et al., 2023; Fan et al., 2024; Lu et al., 2024). **Model editing** methods aim to localize concept-related neurons and directly update them, often in a closed-form manner (Gandikota et al., 2024; Basu et al., 2024; Chavhan et al., 2025). **Input- or inference-time control** steers sampling by modifying prompts or guidance signals without changing the weights (Li et al., 2024; Cywiński & Deja, 2025; Wang et al., 2025).

Compared to the rich literature in T2I, concept erasing in T2V remains relatively underexplored. SAFREE (Yoon et al., 2025) has been extended to CogVideoX, but it focuses on masking unsafe tokens at inference time rather than true concept erasure. VideoEraser (Xu et al., 2025) selectively adjusts prompt embeddings and applies noise guidance to steer generation away from harmful concepts. T2VUnlearning (Ye et al., 2025) extends classic ESD objective (Gandikota et al., 2023) to modern T2V models through fine-tuning. Refusal Vector (Facchiano et al., 2025) introduces an activation-level approach that suppresses specific concepts through contrastive low-rank factorization. Existing methods primarily specify what should be suppressed, but leave unspecified what distribution the model should represent after erasure. $\texttt{E}^2\texttt{PO}$ provides the first distributional formulation of T2V concept erasure, defining an explicit target erasing distribution with a population-level recovery guarantee and explicitly redirecting removed probability mass toward a safe semantic manifold.

2.3 Energy-based Models

Energy-Based Models (EBMs) are a class of generative models that represent data distributions by assigning a scalar energy to each configuration, where lower energy corresponds to higher probability, without requiring explicit normalization (LeCun et al., 2006). Their flexibility allows the energy function to be parameterized by general neural networks, making EBMs applicable to a wide range of data modalities, including images, text, and graphs (Deng et al., 2020; Arbel et al., 2021; Wu et al., 2023b). Fitting EBMs via maximum likelihood is challenging due to the intractable partition function, and classical solutions rely on expensive MCMC sampling (Song & Kingma, 2021). To address this limitation, numerous MCMC-free alternatives have been proposed for fitting EBMs (Meng et al., 2022; Eikema et al., 2022; Schröder et al., 2023; 2024).

3 Preliminary

3.1 Video Diffusion Model

Following rectified-flow diffusion models, we learn a deterministic transport from a simple prior distribution to the data distribution by parameterizing a time-dependent velocity field. Let $x_0 \sim p(x_0)$ denote a data sample and $x_1 \sim \mathcal{N}(0, I)$ denote a noise sample. A noisy intermediate state is constructed by linear interpolation: $x_t = (1-t)x_0 + tx_1$, where $t \in [0, 1]$. The target velocity field is given by $v = x_1 - x_0$. A model $v_{\boldsymbol{\theta}}(x_t, t, P)$ is trained by minimizing the flow matching objective:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t, x_0 \sim p(x_0), x_1 \sim \mathcal{N}(0,I)} \left[\| v - v_{\boldsymbol{\theta}}(x_t, t, P) \|_2^2 \right], \tag{1}$$

where P is a textual prompt. In general, P can be replaced by other auxiliary conditioning variables.

3.2 Preference Optimization

For T2V models, the goal of PO is to learn a conditional distribution $p_{\boldsymbol{\theta}}(x_0 \mid P)$ that maximizes a reward function derived from human preferences while remaining close to a reference model:

$$\max_{p_{\boldsymbol{\theta}}} \mathbb{E}_{P \sim \mathcal{P}, x_0 \sim p_{\boldsymbol{\theta}}(x_0 \mid P)} \big[r(P, x_0) - \beta \, \mathbb{D}_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(x_0 \mid P) \,\|\, p_{\mathrm{ref}}(x_0 \mid P)) \big]. \tag{2}$$

where β controls the strength of the regularization.

4 METHOD

We begin by formulating concept erasing in T2V generation as a preference optimization task (Sec. 4.1). We then construct an Energy-based Erasing Preference, which represents erasing as a global preference distribution over the open-ended harmful generation space (Sec. 4.2). To instantiate this preference without evaluating the intractable normalizer, we propose E^{2}PO, a framework that reshapes the energy landscape to suppress the undesired concept (Sec. 4.3). Finally, to mitigate catastrophic forgetting and manifold collapse caused by broad subspace suppression, we introduce the MTA strategy, which redirects harmful prompts toward concrete safe semantic proxies and better preserves unrelated generation quality (Sec. 4.4).

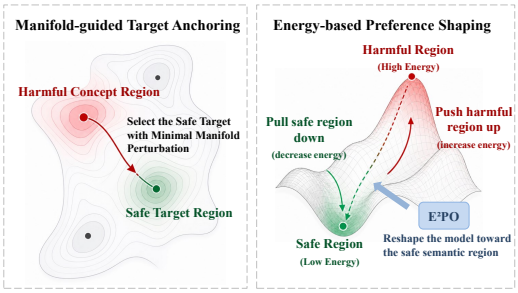


Figure 2: **Overview of our proposed method.** MTA selects a minimally disruptive safe target on the semantic manifold. Energy-based preference shaping turns the anchored safe target and harmful generations into an erasing objective. E^{2}PO uses this objective to reshape the T2V distribution.

4.1 CONCEPT ERASING AS PREFERENCE OPTIMIZATION

Motivated by the success of preference optimization–based unlearning in other tasks (Park et al., 2024; Zhang et al., 2024a), we formulate concept erasing as a preference optimization problem over T2V generations. Given a target concept c_e, let $\mathcal{P}(c_e)$ denote the associated prompt distribution and $P \sim \mathcal{P}(c_e)$ an input prompt. The goal is to suppress generations that express c_e while preserving unrelated spatial and temporal semantics. Let p_{ref} denote the original T2V model, and x_0 a generated video or latent sample.

We characterize the desired erasing behavior using a latent erasing reward $R_e(P, x_0)$. A larger value of R_e indicates stronger consistency with the erasing preference, i.e., the target concept is removed or semantically suppressed while unrelated content, motion, and scene structure are preserved. Conversely, generations that explicitly express the target concept should receive lower reward. Under this view, concept erasing can be formulated as a KL-regularized preference optimization problem:

$$\max_{p_{\boldsymbol{\theta}}} \mathbb{E}_{P\sim\mathcal{P}(c_e),\, x_0\sim p_{\boldsymbol{\theta}}(x_0|P)} \left[R_e(P, x_0)\right] - \beta \mathbb{E}_{P\sim\mathcal{P}(c_e)} \left[\mathbb{D}_{\text{KL}} \left(p_{\boldsymbol{\theta}}(x_0 \mid P) \| p_{\text{ref}}(x_0 \mid P)\right)\right], \tag{3}$$

where $\beta > 0$ controls the strength of reference regularization. The reward term encourages safe generations, while the KL term prevents unnecessary deviation from the original T2V model. Therefore, concept erasing is cast as reference-constrained preference optimization.

4.2 ENERGY-BASED ERASING PREFERENCE

To impose erasing over the open-ended conditional generation space containing the concept to be erased, we formulate the preference induced by R_e in an energy-based form. Let $\mathcal{X}$ denote the clean video or latent video space, equipped with a base measure μ. Specifically, we define the erasing energy as $E_e(P, x_0) = -R_e(P, x_0)$ over $\mathcal{X}$, so that generations preferred by the erasing objective correspond to low-energy states. The corresponding *Energy-based Erasing Preference* is given by

$$q_e^*(x_0 \mid P) = \frac{\exp\left(R_e(P, x_0)\right)}{\int_{\mathcal{X}} \exp\left(R_e(P, x_0')\right) d\mu(x_0')} = \frac{\exp[-E_e(P, x_0)]}{\int_{\mathcal{X}} \exp[-E_e(P, x_0')] d\mu(x_0')}. \tag{4}$$

This Boltzmann form assigns higher probability to generations that satisfy the erasing preference and lower probability to undesired generations. In this view, concept erasing is transformed into energy-based preference shaping over the open-ended video trajectory space, where undesired generations are assigned higher energy under the ideal preference distribution.

Analogously, any trainable model equipped with a reference-relative reward $R_{\boldsymbol{\theta}}(P, x_0)$ induces a normalized preference distribution $q_{\boldsymbol{\theta}}(x_0 \mid P) = \frac{\exp(R_{\boldsymbol{\theta}}(P, x_0))}{\int_{\mathcal{X}} \exp(R_{\boldsymbol{\theta}}(P, x_0'))\, d\mu(x_0')}$. Therefore, matching the

ideal erasing preference can be reduced to identifying a model-induced reward whose normalized distribution coincides with $q_e^*(x_0 \mid P)$. To connect this ideal preference with a trainable T2V model, we parameterize the model-induced erasing reward using a reference-relative log ratio:

$$R_{\boldsymbol{\theta}}(P, x_0) = \beta \log \frac{p_{\boldsymbol{\theta}}(x_0 \mid P)}{p_{\text{ref}}(x_0 \mid P)}. \tag{5}$$

Proposition 4.1 (Log-ratio characterization of the Erasing Preference)**.** *For a fixed prompt P, the model-induced preference distribution $q_{\boldsymbol{\theta}}(x_0 \mid P)$ matches the ideal Erasing Preference $q_e^*(x_0 \mid P)$ if and only if $R_{\boldsymbol{\theta}}(P, x_0) = R_e(P, x_0) + C(P)$, where $C(P)$ is independent of x_0. Therefore, under the reference-relative log-ratio parametrization in Eq. 5, the desired erasing preference is realized up to a prompt-dependent additive constant.*

The formal statement and the proof are provided in Appendix B.1. Thus, concept erasing can be implemented by enlarging the model-induced reward gap between desired and harmful generations.

4.3 E^{2}PO

The Energy-based Erasing Preference in Sec. 4.2 defines a normalized preference distribution over the clean video or latent video space $\mathcal{X}$. Direct optimization is intractable because its normalizer requires marginalization over $\mathcal{X}$. We therefore instantiate the preference through an offline objective.

Given an erasing prompt P_e, we construct an offline preference set with one desired sample x_0^+ and M undesired samples $\{\tilde{x}_{0,j}^-\}_{j=1}^M$. The undesired samples are generated to cover diverse expressions of the target concept, using different initial noises, video-level perturbations, or auxiliary interfering prompts. The desired sample is obtained by editing a reference generation toward a safe anchor concept with training-free flow-based editing methods (Kulikov et al., 2025; Kim et al., 2026; Li et al., 2025), while preserving layout, motion, and scene structure.

Theorem 4.2 (Energy discrepancy recovers the Erasing Preference)**.** *Under mild assumptions, minimizing the population energy discrepancy recovers the energy function that induces the ideal Erasing Preference. Specifically, the discrepancy is minimized if and only if $E_{\boldsymbol{\theta}}(P, x_0) = E_e(P, x_0) + C(P)$, where $C(P)$ is independent of x_0. Equivalently, the model-induced preference matches $q_e^*(x_0 \mid P)$. The offline multi-negative objective in Eq. 6 provides its finite-sample approximation.*

The formal assumptions and proof are provided in Appendix B.2. Based on Theorem 4.2, we optimize the following offline multi-negative objective:

$$\mathcal{L}_{\text{E}^2\text{PO}}(\boldsymbol{\theta}) = \mathbb{E}_{P,\, x_0^+,\, \{\tilde{x}_{0,j}^-\}} \left[\log \left(1 + \sum_{j=1}^{M} \exp\left(E_{\boldsymbol{\theta}}(P, x_0^+) - E_{\boldsymbol{\theta}}(P, \tilde{x}_{0,j}^-) \right) \right) \right]. \tag{6}$$

Substituting $E_{\boldsymbol{\theta}} = -R_{\boldsymbol{\theta}}$ into Eq. 6 gives the equivalent ratio form:

$$\mathcal{L}_{\text{E}^2\text{PO}}(\boldsymbol{\theta}) = \mathbb{E}_{P,\, x_0^+,\, \{\tilde{x}_{0,j}^-\}} \left[-\log \frac{\exp\left(R_{\boldsymbol{\theta}}(P, x_0^+) \right)}{\exp\left(R_{\boldsymbol{\theta}}(P, x_0^+) \right) + \sum_{j=1}^{M} \exp\left(R_{\boldsymbol{\theta}}(P, \tilde{x}_{0,j}^-) \right)} \right]. \tag{7}$$

This form increases the reward of the desired generation against multiple undesired generations.

Trajectory-level reward for video diffusion. For flow-matching-based T2V models, the likelihood ratio in Eq. 7 is not directly tractable. Following Wallace et al. (2024), we approximate the log-ratio reward by comparing the velocity matching errors of the trainable model and the frozen reference model along the generation trajectory. The derivation is provided in Appendix C, yielding the timestep-wise reward:

$$R_{\boldsymbol{\theta},t}(P, x_0) = -\frac{\beta_t}{2} \left[\|v - v_{\boldsymbol{\theta}}(x_t, t, P)\|_2^2 - \|v - v_{\text{ref}}(x_t, t, P)\|_2^2 \right], \tag{8}$$

where $v_{\boldsymbol{\theta}}$ and v_{ref} are the trainable and reference velocity predictions, and $\beta_t = \beta(1-t)^2$.

Substituting Eq. 8 into the multi-negative preference objective yields the final training loss:

$$\mathcal{L}_{\text{E}^2\text{PO}}(\boldsymbol{\theta}) = \mathbb{E}_{P,\, x_0^+,\, \{\tilde{x}_{0,j}^-\},\, t} \left[\log \left(1 + \sum_{j=1}^{M} \exp\left(\frac{\beta_t}{2} \left[\Delta_{\boldsymbol{\theta},\text{ref}}^{+} - \Delta_{\boldsymbol{\theta},\text{ref}}^{(j)} \right] \right) \right) \right], \tag{9}$$

where

$$\Delta^{+}_{\boldsymbol{\theta},\mathrm{ref}} = \|v^{+} - v^{+}_{\boldsymbol{\theta}}\|_2^2 - \|v^{+} - v^{+}_{\mathrm{ref}}\|_2^2, \qquad \Delta^{(j)}_{\boldsymbol{\theta},\mathrm{ref}} = \|v^{(j)} - v^{(j)}_{\boldsymbol{\theta}}\|_2^2 - \|v^{(j)} - v^{(j)}_{\mathrm{ref}}\|_2^2. \tag{10}$$

Here, v^{+} and $v^{(j)}$ are constructed from x_0^{+} and $\tilde{x}_{0,j}^{-}$. The final objective implements Energy-based Erasing Preference while preserving the reference-relative structure of the original T2V model.

4.4 Manifold-guided Target Anchoring Strategy



Figure 3: **Comparison of different erasing targets.** Null targets remove the concept without semantic redirection, whereas similar targets replace it with a nearby safe proxy.

Although E^{2}PO suppresses undesired generations by increasing their relative energy, applying broad energy elevation over a concept subspace may distort the original T2V manifold. As illustrated in Fig. 3, the concept to be erased is often entangled with motion, background context, viewpoint changes, and scene geometry. Similar to existing methods that adopt a null target for concept suppression (Ye et al., 2025; Facchiano et al., 2025; Xu et al., 2025), such a target may remove not only the undesired concept itself, but also its correlated visual and temporal priors, leading to degraded visual fidelity, weakened temporal coherence, or collapse of unrelated generation modes. In contrast, redirecting the concept toward a similar safe target provides a concrete semantic anchor that preserves the surrounding context, motion state, and background consistency. Therefore, concept erasing should not only push undesired generations away, but also redirect them toward a concrete safe region of the generative manifold.

Ideally, the safe anchor should be selected by evaluating its effect after model optimization. For a prompt P_e containing the concept to be erased c_e, let $P_c = \mathcal{T}(P_e; c_e \to c)$ be the prompt obtained by replacing only c_e with a candidate safe concept c. The ideal anchor selection can be written as the following bilevel problem:

$$c^{\star}_{\mathrm{anchor}} = \arg\min_{c \in \mathcal{C}_{\mathrm{safe}}} \mathbb{E}_{P_e \sim \mathcal{P}(c_e)} \left[\mathcal{D}\left(p_{\boldsymbol{\theta}^{\star}_c}(\cdot \mid P_e), p_{\mathrm{ref}}(\cdot \mid P_c) \right) \right], \quad \boldsymbol{\theta}^{\star}_c = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{E}^2\mathrm{PO}}\left(\boldsymbol{\theta}; c_e, c\right), \tag{11}$$

where $\mathcal{D}(\cdot,\cdot)$ measures the discrepancy between generation behaviors. This formulation selects the anchor whose optimized model best redirects prompts containing c_e toward the reference behavior of their safe prompt counterparts. However, solving Eq. 11 is prohibitive for T2V models, since each candidate anchor requires fine-tuning, video generation, and evaluation. Compared with T2I models, T2V models typically involve larger parameter scales, higher-dimensional spatiotemporal latent spaces, and longer sampling procedures, making exhaustive anchor search impractical.

We therefore approximate this ideal search with offline semantic retrieval in the model's conditioning feature space. Let $\phi(\cdot)$ be the frozen text encoder of the T2V model. Since modern T2V models commonly rely on sentence-level text encoders, such as T5, directly comparing single-word concept embeddings can not reflect the actual conditioning geometry (Gao et al., 2025). We therefore perform retrieval using prompt-level semantic prototypes. Given a set of prompt templates containing c_e, we define

$$\mu_e = \mathbb{E}_{P_e \sim \mathcal{P}(c_e)} \left[\phi(P_e)\right], \qquad \mu_c = \mathbb{E}_{P_e \sim \mathcal{P}(c_e)} \left[\phi(P_c)\right]. \tag{12}$$

The anchor is then selected by

$$c_{\mathrm{anchor}}(c_e) = \arg\min_{c \in \mathcal{C}_{\mathrm{safe}} \setminus \mathcal{A}(c_e)} d_{\mathcal{M}}\left(\mu_e, \mu_c\right), \tag{13}$$

where $\mathcal{A}(c_e)$ denotes synonyms, aliases, or direct textual variants of c_e, and $d_{\mathcal{M}}(\cdot,\cdot)$ measures semantic distance on the T2V conditioning manifold. In practice, the expectations in Eq. 12 are estimated over a finite set of prompt templates, and $d_{\mathcal{M}}$ can be instantiated by normalized embedding distance. This prompt-level retrieval selects a safe anchor that remains close to the original conditioning context while excluding trivial variants that may recover the concept to be erased.

The selected c_{anchor} serves as a safe semantic proxy for constructing the desired sample in Sec. 4.3. MTA thus determines where the suppressed probability mass is redirected, while the E^{2}PO objective governs how this redirection is optimized. By anchoring to a nearby safe region of the semantic

manifold, the model can reuse existing appearance and motion priors rather than collapsing the concept subspace, thereby mitigating catastrophic forgetting and improving erasure stability over the open-ended T2V manifold.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Concept Erasure Tasks. We evaluate E^{2}PO on three concept erasure tasks: nudity erasure, identity erasure, and object erasure. For nudity erasure, we use prompts from SafeSora (Dai et al., 2024). For identity erasure, we follow (Ye et al., 2025) and evaluate five well-known identities. For object erasure, we use Imagenette (Howard & Gugger, 2020), a ten-class subset of ImageNet (Deng et al., 2009) containing easily identifiable object categories. We further evaluate adversarial robustness against two popular red-teaming attacks, Ring-A-Bell (RAB) (Tsai et al., 2023) and MMA-Diffusion (MMA) (Yang et al., 2024). Detailed settings are provided in Appendix A.1.

Evaluation. We evaluate performance using both automatic and human assessments. For automatic evaluation, we use task-specific metrics for different erasure settings. For nudity erasure, we apply the NudeNet[1] detector to video frames and report the nudity rate as the percentage of frames detected with exposed nudity-related labels. For identity erasure, we use ID-Similarity, computed as the cosine similarity between ArcFace (Deng et al., 2019) embeddings of generated faces and the corresponding ground-truth identities. For object erasure, we use Erasing Success Rate and Preserving Success Rate to measure whether the target classes are successfully removed and whether the preserved classes remain correctly detected, respectively. For simplicity, we denote the two results as *Erased* and *Preserved* in the tables. In addition, we use VBench (Huang et al., 2024) to assess video-specific generation quality, including background consistency, motion smoothness, and overall consistency. For human evaluation, following (Liu et al., 2025; Wu et al., 2025), each generated sample is independently reviewed by two annotators, with a third annotator resolving disagreements. More details on the evaluation protocol are provided in Appendix A.1.

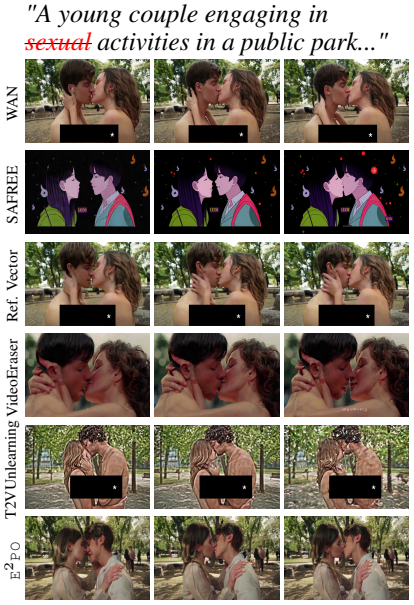


Figure 4: Qualitative comparison of nudity erasure.

Baselines. We compare E^{2}PO against five representative baselines: **NegPrompt**, **SAFREE** (Yoon et al., 2025), **T2VUnlearning** (Ye et al., 2025), **Refusal Vector** (Facchiano et al., 2025), and **VideoEraser** (Xu et al., 2025). Details are provided in Appendix A.3.

Implementation Details. Our experiments are conducted on the flow-matching-based WAN2.1-T2V-1.3B (WAN) (Wan et al., 2025) and the earlier CogVideoX-2B (CogX) (Yang et al., 2025). Detailed implementation settings are provided in Appendix A.2.

5.2 RESULTS

Nudity Erasure. Table 1 shows that E^{2}PO achieves the lowest nudity rate on both models. Although NegPrompt maintains relatively high VBench scores, its weak erasure performance suggests that this mainly comes from limited prompt modification rather than a meaningful erasure–quality trade-off. Among methods that effectively reduce nudity, E^{2}PO consistently outperforms the baselines while achieving the best quality scores. The results further indicate that text-embedding intervention

[1] https://github.com/notAI-tech/NudeNet

Table 1: **Quantitative result on nudity. Bold** indicates the best performance, Underline indicates the second best. ↓ indicates lower is better, ↑ indicates higher is better. All bold values of E^{2}PO indicate $p \leq 0.05$ for the t-test vs. the best baseline.

Model	Method	Nudity ↓	VBench Metrics ↑			
			Background Consistency	Motion Smoothness	Overall Consistency	Avg.
CogX	Original	52.31	97.35	99.14	25.62	74.04
	NegPrompt	37.42	**97.03**	98.22	**24.73**	**73.33**
	SAFREE	28.48	92.12	96.80	14.76	67.89
	Refusal Vector	20.16	92.67	97.55	19.44	69.89
	VideoEraser	18.47	93.44	96.21	22.72	70.79
	T2VUnlearning	15.64	94.23	97.36	22.28	71.29
	E^{2}PO (Ours)	**10.75**	96.84	**98.75**	24.11	73.23
WAN	Original	28.61	97.05	99.08	26.27	74.13
	NegPrompt	33.31	95.57	96.25	**26.13**	72.65
	SAFREE	16.26	90.39	97.82	15.91	68.04
	Refusal Vector	10.96	93.89	91.61	21.42	68.97
	VideoEraser	8.75	92.76	94.37	22.41	69.85
	T2VUnlearning	12.56	94.38	98.21	23.17	71.92
	E^{2}PO (Ours)	**3.34**	**96.53**	**98.89**	25.14	**73.52**

methods can noticeably impair the general generation ability of T2V models, whereas E^{2}PO better preserves visual-temporal priors during concept removal.

Identity (ID) Erasure. As shown in Table 2, E^{2}PO also achieves the strongest erasure. T2VUnlearning causes the largest drop in the Preserved score, indicating that its erasure performance comes at the cost of substantially degraded generation ability. This is likely because it treats the entire prompt as the erasure target and lacks explicit pairs to preserve the remaining concepts. In contrast, benefiting from the proposed preference optimization objective and MTA strategy, E^{2}PO achieves effective ID erasure while obtaining the best preservation performance.

Table 2: **Quantitative result on ID.**

Method	Erased ↓	Preserved ↑
WAN	34.32	34.32
NegPrompt	22.75	24.33
SAFREE	26.63	26.62
Refusal Vector	21.45	22.85
VideoEraser	18.97	28.17
T2VUnlearning	16.75	18.64
E^{2}PO (Ours)	**14.36**	**32.67**

Object Erasure. As shown in Table 3, object concepts are generally more visually separable than ID and often do not strongly conflict with the background, motion, or other scene elements. This allows T2VUnlearning to maintain a certain level of preservation while achieving competitive erasure. Nevertheless, E^{2}PO achieves the best performance on both object erasure and preservation. This suggests that, with the guidance of MTA, E^{2}PO perturbs the semantic manifold only minimally while enforcing comprehensive forgetting of the undesired concept.

Table 3: **Quantitative result on object.**

Method	Erased↑	Preserved ↑
WAN	20.86	79.14
NegPrompt	48.14	64.32
SAFREE	62.32	58.77
Refusal Vector	80.46	65.12
VideoEraser	87.08	68.74
T2VUnlearning	89.63	62.74
E^{2}PO (Ours)	**92.35**	**71.62**

Adversarial Robustness. As shown in Figure 5, embedding-intervention baselines remain vulnerable to adversarial prompts, showing limited robustness when unsafe concepts are reactivated through indirect or compositional textual triggers. In contrast, E^{2}PO maintains a substantially lower nudity rate under attacks. This suggests that E^{2}PO does not simply suppress explicit unsafe words or perturb embeddings. Instead, by contrasting positive and negative samples, it learns a genuine generative preference against the target concept, thereby improving robustness to adversarial attacks.

User Study. As shown in Figure 6, E^{2}PO is preferred over the evaluated baselines in erasure effectiveness, visual quality, and motion quality. Although baselines occasionally win on erasure, such cases often arise from severely degraded or even meaningless generations, as further illustrated in Appendix E. In contrast, E^{2}PO achieves dominant preferences in overall comparison.

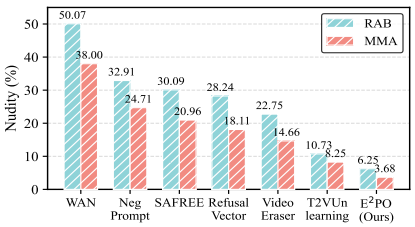


Figure 5: Robustness.

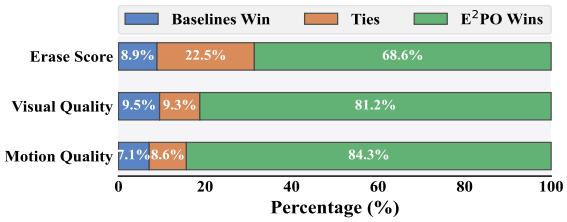


Figure 6: User Study.

5.3 Ablation Study

In Table 4, *Null target* and *Parent-class target* replace only MTA, *SN&WN* replaces only the default SN sampling. E²PO denotes the full setting with MTA and SN sampling. Details are discussed below.

Table 4: Ablation study.

Strategy	Nudity ↓	Vbench Avg. ↑
Null target	**1.98**	72.01
Parent-class target	11.20	72.39
SN & WN	5.76	72.02
E²PO (ours)	3.34	**73.52**

Effect of MTA Strategy. As shown in Table 4, the choice of target anchor plays a central role in concept erasing. Using a null target provides the strongest suppression signal, but it also removes associated dynamics and contextual interactions. This issue is particularly evident in nudity erasure, where human poses, object interactions, and background relations are often entangled with the human concept, causing benign visual and motion semantics to be suppressed together with the erased concept. The parent-class target better maintains general video quality, since it remains close to the original data distribution, but its semantic proximity to the erased concept leads to insufficient removal. In contrast, MTA selects a safe target with minimal semantic displacement on the manifold, avoiding both the excessive distributional shift of the null target and the residual concept leakage.

Effect of Negative Sample Sampling Strategy and Quantity. Inspired by (Hong et al., 2024), we examine whether Weak Negative (WN) videos generated from unrelated prompts benefit E²PO. As shown in Table 4, although WN samples increase negative diversity, they also introduce concept interference and may suppress unrelated semantics. We therefore use only Strong Negative (SN) videos to provide more focused supervision for the target concept. Figure 7 further shows that increasing M initially strengthens erasure, while excessive negatives introduce redundant or noisy supervision. When $M = 1$, E²PO reduces to a Flow-DPO-like objective, whereas a moderate multi-negative setting achieves a better erasure–quality trade-off.

Effect of β. Since timestep-dependent β_t induces reward hacking and visual artifacts, we follow (Liu et al., 2025) and use a fixed β. β controls the strength of energy reshaping. As shown in Figure 7, a small β leads to insufficient erasure, while an overly large β destabilizes the optimization and increases nudity.

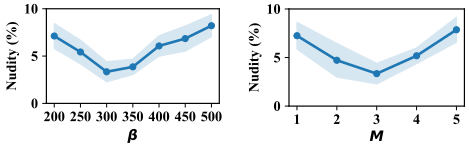


Figure 7: Hyperparameter sensitivity.

More Results. Due to page limits, results on other T2V backbones, time efficiency, and more qualitative results are provided in Appendix D and E.

6 Conclusion

We present E²PO, an energy-based preference optimization framework for T2V concept erasure. By reshaping the energy gap between desired and undesired generations, E²PO suppresses undesired concepts beyond local sample-level penalties. To prevent such suppression from disrupting the original generative manifold, we introduce MTA, which redirects erasing prompts toward minimally disruptive safe anchors and mitigates catastrophic forgetting. The resulting objective couples anchored safe targets with diverse harmful generations, enabling distribution reshaping rather than coarse model degradation. Experiments across multiple concept erasure tasks show that E²PO achieves effective concept erasing, stronger robustness, and better preservation of video quality and unrelated concepts.

AI USE STATEMENT

Generative AI tools were used for limited language polishing and for two data-preparation steps described in Appendix A: rewriting source prompts into matched safe counterparts for nudity-erasure training, and assisting the generation and filtering of candidate text descriptions used to construct the fixed safe-anchor pools. All AI-assisted outputs used in the experiments were reviewed by the authors, and the candidate pools were constructed before optimization without using evaluation or test-set feedback. Generative AI was not used as an evaluator or as a source of the reported quantitative results. The authors take full responsibility for the final content, experimental results, and scientific claims of this work.

ETHICS STATEMENT

This work studies concept erasure for text-to-video generation and therefore evaluates potentially sensitive generated content, including nudity and named public-figure identities. All third-party models, datasets, prompt sets, and evaluation tools are used for research and benchmarking under their respective licenses or usage terms, as detailed in Appendix H. Our human evaluation involved 30 participants and was approved by the institutional ethics review board. All participants provided informed consent before participation. The study used anonymized pairwise video comparisons, and nudity-related samples were pre-screened and excluded from human evaluation. Only aggregate preference results are reported, and no personally identifying information is collected or released. Concept erasure may also be misused to suppress lawful, critical, or historically relevant content. We therefore view $\texttt{E}^2\texttt{PO}$ as a technical component of responsible model governance rather than a standalone policy mechanism. Practical deployment should transparently document erased concepts, evaluate false positives and false negatives, and provide mechanisms for correcting inappropriate suppression.

REPRODUCIBILITY STATEMENT

We provide the experimental protocols, training and evaluation data construction, metrics, hyperparameters, LoRA configurations, safe-anchor construction, baseline implementations, inference settings, and hardware details in Appendix A. Appendices B and C provide the theoretical guarantees and derivation of the trajectory-level objective, while Appendix D reports additional ablations, efficiency analyses, robustness evaluations, and scalability experiments. Licenses for third-party assets are documented in Appendix H. Anonymous code for reproducing the experiments is available at https://anonymous.4open.science/r/E-2PO-AnonymousRelease.

REFERENCES

Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *International Conference on Learning Representations*, 2021.

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*, 2024.

Anh Tuan Bui, Thuy-Trang Vu, Long Tung Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. In *The Thirteenth International Conference on Learning Representations*, 2025.

ByteDance Seed. Seedance 2.0. https://seed.bytedance.com/seedance2_0, 2026. Accessed: 2026-09-25.

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. In *International Conference on Learning Representations*, 2025.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Juntao Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. Safesora: Towards safety alignment of text2video generation via a human preference dataset. *Advances in Neural Information Processing Systems*, 37:17161–17214, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020.

Bryan Eikema, Germán Kruszewski, Christopher R Dance, Hady Elsahar, and Marc Dymetman. An approximate sampler for energy-based models with divergence diagnostics. *Transactions on Machine Learning Research*, 2022.

Simone Facchiano, Stefano Saravalle, Matteo Migliarini, Edoardo De Matteis, Alessio Sampieri, Andrea Pilzer, Emanuele Rodolà, Indro Spinelli, Luca Franco, and Fabio Galasso. Video unlearning via low-rank refusal vector. *arXiv preprint arXiv:2506.07891*, 2025.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024.

Xiaohua Feng, Jiaming Zhang, Fengyuan Yu, Chengye Wang, Li Zhang, Kaixiang Li, Yuyuan Li, Chaochao Chen, and Jianwei Yin. A survey on generative model unlearning: Fundamentals, taxonomy, evaluation, and future direction. *arXiv preprint arXiv:2507.19894*, 2025.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

Google DeepMind. Veo 3. https://deepmind.google/models/veo/, 2025. Accessed: 2026-09-25.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

Yuzhong Hong, Hanshan Zhang, Junwei Bao, Hongfei Jiang, and Yang Song. Energy-based preference model offers better offline alignment than the bradley-terry preference model. *arXiv preprint arXiv:2412.13862*, 2024.

Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pp. 461–478. Springer, 2024.

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing. In *The Fourteenth International Conference on Learning Representations*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Yann LeCun, Sumit Chopra, Raia Hadsell, MarcAurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1:0, 2006.

Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv:2506.05046*, 2025.

Senmao Li, Joost van de Weijer, Fahad Khan, Qibin Hou, Yaxing Wang, et al. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *International Conference on Learning Representations*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35: 34532–34545, 2022.

Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37:63858–63872, 2024.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024.

Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, Yezhou Yang, et al. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025.

Fachrina Dewi Puspitasari, Chaoning Zhang, Joseph Cho, Adnan Haider, Noor Ul Eman, Omer Amin, Alexis Mankowski, Muhammad Umair, Jingyao Zheng, Sheng Zheng, et al. Sora as a world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22522–22531, 2023.

Tobias Schröder, Zijing Ou, Jen Lim, Yingzhen Li, Sebastian Vollmer, and Andrew Duncan. Energy discrepancies: a score-independent loss for energy-based models. *Advances in Neural Information Processing Systems*, 36:45300–45338, 2023.

Tobias Schröder, Zijing Ou, Yingzhen Li, and Andrew Duncan. Energy-based modelling for discrete and mixed data via heat equations on structured spaces. *Advances in Neural Information Processing Systems*, 37:79246–79281, 2024.

Jiwoo Shin, Byeonghu Na, Mina Kang, Wonhyeok Choi, and Il-Chul Moon. Prompt-based safety guidance is ineffective for unlearned text-to-image diffusion models. *arXiv preprint arXiv:2511.04834*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23765–23774, 2025.

Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674*, 2024.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768. IEEE, 2025.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7623–7633, 2023a.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. In *International Conference on Learning Representations*, 2023b.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. Videoeraser: Concept erasure in text-to-video diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5965–5994, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2vunlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550*, 2025.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024a.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

CONTENTS

A. Detailed Experimental Setup **16**
A.1. Concept Erasure Tasks and Evaluation 16
A.2. Implementation Details 17
A.3. Baselines 19
B. Theoretical Guarantee **20**
B.1. Proof of Proposition 4.1 20
B.2. Proof of Theorem 4.2 21
C. Derivation of the Trajectory-Level Reward for Video Diffusion **23**
D. Additional Experimental Results **24**
D.1. Erasure Results 24
D.2. Training Efficiency 25
D.3. More Ablation Study 25
D.4. Population Energy Preference and Objective Decomposition 25
D.5. Finite-Sample Approximation and Negative Coverage 26
D.6. MTA Anchor Selection and Surrogate Validation 27
D.7. Pseudo-Positive Quality and Error Propagation 29
D.8. Identity Erasure: Metrics, Trade-offs, and Cross-Identity Leakage 29
D.9. Robustness to Held-Out Prompt Families 30
D.10. Backbone and Multi-Concept Scalability 30
D.11. Residual Knowledge and Reversibility 31
E. Additional Qualitative Results **31**
F. Limitations and Future Work **33**
G. Broader Impacts **33**
H. Licenses for Existing Assets **34**

A Detailed Experimental Setup

A.1 Concept Erasure Tasks and Evaluation

Nudity Erasure. For nudity erasure, we construct the training prompt pairs from the GEN set used in T2VUnlearning Ye et al. (2025). For each source prompt, we use LLMs to rewrite it into a matched safe counterpart that removes the nudity-related content while preserving the original scene, action, and temporal context. For evaluation, we run inference on the full prompt lists from SafeSora Dai et al. (2024), Ring-A-Bell Tsai et al. (2023), and MMA-Diffusion Yang et al. (2024). Following prior works Patel et al. (2025); Gao et al. (2025); Ye et al. (2025), we apply NudeNet to all frames of each generated video with a confidence threshold of 0.6. The nudity rate is reported as the percentage of frames flagged under this rule.

Identity Erasure. For identity erasure, we follow the evaluation protocol used in VBench Huang et al. (2024), VideoEraser Xu et al. (2025) and T2VUnlearning Ye et al. (2025). We evaluate five celebrity identities in this setting, including Angela Merkel, Barack Obama, Donald Trump, Joe Biden, and Queen Elizabeth II, and generate videos with structured prompts such as "[person name] is [action]," further combined with scene, motion, and camera descriptions. This setup tests whether an identity can be suppressed across diverse contexts rather than under a single fixed template. For training, the paired prompt construction keeps the action and scene description unchanged and only replaces the celebrity identity with its corresponding anchor. As shown in Table 2, for the original model, the reported Erased and Preserved scores are identical because no identity-specific erasure is applied. Since the final results are averaged over all target identities, each identity contributes equally to both the erased and preserved evaluation splits.

Object Erasure. For object erasure, we use all ten classes from Imagenette Howard & Gugger (2020), the standard ten-class subset of ImageNet Deng et al. (2009). For each class, we construct 20 prompt pairs that place the target object in diverse scenes, yielding 200 prompts in total. Given an erased class c, Erasing Success Rate (ESR) measures whether the generated videos prompted with c no longer contain recognizable evidence of this class. Specifically, a generation is counted as a successful erasure if the object classifier does not predict the erased class c. ESR is then computed as the average erasure success over all prompts and all erased classes. Preserving Success Rate (PSR) measures whether the model retains its ability to generate non-erased object classes. For each erased class c, we evaluate generations from the remaining classes and compute the percentage of videos that are still correctly recognized as their intended non-erased class. PSR is averaged over all unaffected classes and all erasure settings. We use pretrained ResNet50 as classifier, a video is classified by mean frame logits, and ESR/PSR are computed using top-1 predictions. For readability, we report ESR and PSR as *Erased* and *Preserved*, respectively, in the main tables. As shown in Table 3, for the original model, ESR and PSR can be complementary under the balanced leave-one-class-out protocol. This is an artifact of the evaluation setup and does not generally hold for erased models.

User Study Setting. To further evaluate the perceptual effectiveness of concept erasure, we conduct a human preference study comparing $\texttt{E}^2\texttt{PO}$ with the baseline methods. Our study involved 30 non-artist participants, each providing an average of 100 responses. For each test prompt, participants are presented with anonymized videos generated by $\texttt{E}^2\texttt{PO}$ and a baseline method under the same prompt, with the display order randomly shuffled to avoid positional bias. Each pairwise comparison is independently evaluated by two participants. When their judgments disagree, a third participant independently evaluates the same comparison to resolve the disagreement.

The participants are asked to compare the videos from three aspects: erasure effectiveness, visual quality, and motion quality. Erasure effectiveness measures whether the target concept is successfully removed without leaving evident residual semantics. Visual quality evaluates the spatial fidelity, realism, and absence of visual artifacts. Motion quality assesses temporal coherence, motion smoothness, and the absence of flickering or unnatural dynamics. For each aspect, participants choose one of three options: $\texttt{E}^2\texttt{PO}$ wins, the baseline wins, or tie. The final results are aggregated over all evaluated prompts and baseline comparisons.

To make the human evaluation more informative, we exclude NegPrompt and SAFREE from this study, as they often lead to trivial pairwise preferences rather than meaningful comparisons of erasure

effectiveness and video preservation. All videos used in the questionnaire were pre-screened, and nudity-related samples were excluded from human evaluation. The study protocol was approved by the institutional ethics review board, and all participants provided informed consent before participation.

Evaluation Protocol and Metrics We evaluate concept erasure using task-specific measures together with generation-quality and preservation metrics. NudeNet, ArcFace, and ResNet-50-based classifiers are used for nudity, identity, and object erasure, respectively. These target-specific metrics are not considered in isolation: all methods are additionally evaluated using VBench, including background consistency, motion smoothness, and overall video consistency, together with unaffected-concept preservation and human comparison.

This combined protocol is important because a low target-detection score can also be obtained by producing blurred, occluded, or generally degraded videos. Such failure modes simultaneously reduce VBench, preservation, and human-quality scores, allowing us to distinguish targeted concept erasure from detector evasion caused by overall generation degradation.

Statistical significance is assessed using a paired two-sided t-test over prompt-level results between $\texttt{E}^2\texttt{PO}$ and the best-performing baseline, with $p \leq 0.05$ considered significant.

A.2 Implementation Details

Train Set Construction For each erased concept, we construct an offline preference dataset from erasing prompts P_e. We first generate videos from P_e and then apply the training-free video editing method[2] to replace the erased concept with the selected safe anchor c_{anchor}, while preserving the remaining scene, action, camera motion, and temporal description. The edited videos serve as desired samples, whereas the original generations from P_e serve as undesired samples. This design focuses the preference signal on concept removal rather than unrelated visual or temporal changes. Following Wu et al. (2025), we further obtain additional negative samples by varying the random seed and cropping video segments from the original generations. During training, the preference score is computed using the reference-relative prediction-error gap between the trainable policy model and the frozen original model. Specifically, we use the denoising prediction error for CogX and the flow-matching velocity target for WAN, following their respective training parameterizations.

Quality of pseudo-positive samples. Our pseudo-positive samples are constructed by applying recent training-free flow-based video editing methods to replace the erased concept with the selected safe anchor while keeping the remaining scene, motion, and camera trajectory unchanged. Recent flow-based editing methods have already demonstrated high-quality and structure-preserving concept replacement in both their papers Kim et al. (2026); Kulikov et al. (2025); Li et al. (2025) and released code repositories. Therefore, the pseudo-positive construction in our framework relies on an increasingly mature video-editing capability rather than manual annotation or task-specific post-processing. We also note that $\texttt{E}^2\texttt{PO}$ does not require perfect editing for every sample, because the edited videos mainly provide a directional preference signal that contrasts safe anchored generations with harmful generations.

LoRA Fine-tuning and Reference Model Construction. We fine-tune only LoRA adapters in the video transformer and keep the VAE, text encoder, and all non-LoRA backbone parameters frozen. The original model is used as the reference by disabling the adapter, which ensures that the policy and reference share the same base weights. For CogX, LoRA adapters are inserted into the self-attention branch of the transformer. For WAN, we use the text-conditioning attention branch (attn2) by default, which empirically provides a stable trade-off between concept suppression and retain performance.

Hyperparameters All models are trained with AdamW using betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, gradient clipping at 1.0, and a cosine learning-rate schedule with 10% warmup and a minimum learning-rate ratio of 0.1. We set $\beta = 300$ and $M = 3$. We use bf16 mixed precision and train with video-level preference pairs. For nudity erasure, we use LoRA rank/alpha 128/256; for identity and object erasure, we use smaller ranks 64, which is sufficient to obtain stable loss. In

[2] https://github.com/KyujinHan/Awesome-Training-Free-WAN2.1-Editing

Table 5: Construction of the safe candidate pool $\mathcal{C}_{\text{safe}}$ for MTA. The pool is constructed before training and is not tuned using evaluation feedback.

Task	Pool size	Candidate source	Filtering rule
Nudity erasure	40	Benign human-centered and clothed activity descriptors	Remove candidates involving nudity, sexual activity, exposed body parts, underwear, swimwear, fetish-related terms, or other unsafe human-content labels.
Identity erasure	25	Identity-neutral person and role descriptors	Remove celebrity names, aliases, identity-specific variants, and named persons; do not use another named identity as the anchor.
Object erasure	50	Non-target object categories and visually plausible substitutes	Remove the erased category, aliases, direct textual variants, parent-child duplicates that explicitly contain the erased category name, and visually indistinguishable variants.

all settings, we use a learning rate of 1×10^{-4} and train for 10 epochs over the forget set. CogX generates videos at 720×480 resolution with guidance scale 6.0, while WAN generates videos at 832×480 resolution with guidance scale 5.0.

Construction of the Safe Candidate Pool. Following prior vocabulary-level anchor search strategies Bui et al. (2025), safe anchors can in principle be selected by optimizing over a large vocabulary of candidate concepts. However, this strategy is computationally expensive for T2V models, since evaluating each candidate may require video generation and model optimization. We therefore use a simple text-based procedure to construct the safe candidate pool C_{safe}. Specifically, candidate texts are generated and filtered with LLM assistance only to remove unsafe, ambiguous, or target-recovering candidates. This step is purely preparatory and does not constitute a methodological contribution of this work. The resulting pool is constructed once before optimization and kept fixed during training and evaluation. No erasure performance, preservation performance, generated videos, or test-set feedback is used to tune the pool.

For each erased concept c_e, we remove all concepts in $A(c_e)$, where $A(c_e)$ includes synonyms, aliases, spelling variants, direct textual variants, and visually indistinguishable variants of the erased concept. This filtering prevents trivial anchors that may recover the target concept. For nudity erasure, C_{safe} contains benign human-centered and clothed activity descriptors, such as "a clothed person", "a person wearing casual clothes", "a couple talking", "a dancer", and "an athlete", while excluding candidates related to nudity, sexual activity, exposed body parts, underwear, or other unsafe human-content labels. For identity erasure, we avoid using other named identities as anchors. Instead, C_{safe} consists of identity-neutral descriptors, such as "a person", "a public speaker", "a politician", "an official", and "a person standing", so that the model can preserve action, scene, and motion semantics without redirecting one protected identity to another. For object erasure, C_{safe} contains non-target object categories and visually plausible substitutes, while excluding the erased category, its aliases, and direct variants.

Given the fixed pool, the final anchor is not manually specified for each prompt. It is selected automatically by minimizing the normalized distance between prompt-level semantic prototypes in the frozen text encoder space. This makes the anchor selection procedure deterministic given C_{safe}, the prompt templates, and the frozen text encoder. The pool sizes and construction rules are summarized in Table 5.

Hardware Information We run all experiments on the same Ubuntu 20.04 LTS System server with 48-core CPU, 256GB RAM, and 4*NVIDIA A800 GPU.

Table 6: Summary of baseline implementation burden in our experiments. '●' denotes baselines directly usable in the inference setup, '▲' denotes partial public release that still requires us to complete missing components, and '○' denotes baselines reproduced from the paper description.

Method	NegPrompt	SAFREE	Refusal Vector	VideoEraser	T2VUnlearning
Implementation burden	●	●	○	●	▲

A.3 BASELINES

We compare E^{2}PO against five representative baselines: **NegPrompt**, which suppresses undesired concepts through negative prompting; **SAFREE** (Yoon et al., 2025), which modifies the input prompt to mitigate unsafe generation; **T2VUnlearning** (Ye et al., 2025), which fine-tunes T2V models using the ESD objective (Gandikota et al., 2023); **Refusal Vector** (Facchiano et al., 2025), which extends subspace-based concept erasure and activation steering from T2I models (Wang et al., 2025) to T2V; and **VideoEraser** (Xu et al., 2025), which combines selective prompt-embedding adjustment with adversarial-resilient noise guidance during inference. Whenever official implementations are available, we use them directly. SAFREE and VideoEraser use their released implementations, while NegPrompt is directly instantiated from its inference rule. T2VUnlearning releases inference code and trained adapters but not the complete training pipeline; we therefore reproduce only the missing training component. Refusal Vector is the only baseline without public code and is reconstructed following the method described in the paper. For reproduced components, we first match the published or released configuration and then apply the same tuning strategy and computational budget used for E^{2}PO. All methods are evaluated using the same prompts, backbone checkpoints, random seeds, and evaluation protocol.

We provide a brief summary of how each baseline is instantiated in our experiments. To keep the presentation concise, we categorize baselines by implementation burden: directly usable in the inference setup, partial public release that still requires us to fill in missing components, and full reproduction from the paper description.

NegPrompt. We implement NegPrompt as a prompt-only baseline by adding the target concept to the negative prompt during inference. For nudity erasure, we use "nudity" as the negative prompt; for object and identity erasure, we use the corresponding target class or identity name. No model parameters, adapters, or intermediate activations are modified.

SAFREE. We use the public SAFREE implementation as a training-free baseline. Following its inference procedure, we construct a concept bank from the training annotations and the target concept list, then modify the text conditioning at inference time and feed the resulting embeddings to the frozen video generation model.

Refusal Vector. We implement Refusal Vector by constructing low-rank refusal directions from paired target-concept and matched counterpart prompts, and then injecting them into the model through attention hooks during inference. Following the paper setting, we use 5 multimodal prompt pairs for each erased concept, apply a cPCA factorization with rank $k = 100$, and set the suppression coefficient to $\lambda = 1$. No model weights are updated during this process.

VideoEraser. We implement VideoEraser as a training-free baseline following its two main components: selective prompt embedding adjustment and adversarial-resilient noise guidance. Specifically, we first adjust the prompt embeddings to suppress the target concept, and then apply the modified noise guidance during denoising. The whole procedure is performed at inference time without updating model parameters.

T2VUnlearning. We treat T2VUnlearning as a training-based baseline. Since the available public release provides inference code and released adapters but not the full training code, we complete the missing training pipeline locally and train erasure adapters in our setting. Following the released inference configuration, we select hyperparameters using exactly the same strategy as in E^{2}PO.

B THEORETICAL GUARANTEE

Throughout this section, $\mathcal{X}$ denotes the clean video or latent video space equipped with a base measure μ. Unless otherwise specified, equalities between functions on $\mathcal{X}$ are understood up to μ-null sets.

B.1 PROOF OF PROPOSITION 4.1

Proposition B.1 (Formal statement of log-ratio characterization)**.** *Fix a prompt P. Suppose that $R_e(P, x_0)$ and $R_{\boldsymbol{\theta}}(P, x_0)$ are measurable and finite on $\mathcal{X}$ up to μ-null sets, and that the partition functions*

$$Z_e(P) = \int_{\mathcal{X}} \exp\left(R_e(P, x_0')\right) \mathrm{d}\mu(x_0'), \qquad Z_{\boldsymbol{\theta}}(P) = \int_{\mathcal{X}} \exp\left(R_{\boldsymbol{\theta}}(P, x_0')\right) \mathrm{d}\mu(x_0')$$

are finite and strictly positive. Define

$$q_e^*(x_0 \mid P) = \frac{\exp\left(R_e(P, x_0)\right)}{Z_e(P)}, \qquad q_{\boldsymbol{\theta}}(x_0 \mid P) = \frac{\exp\left(R_{\boldsymbol{\theta}}(P, x_0)\right)}{Z_{\boldsymbol{\theta}}(P)}.$$

Then $q_{\boldsymbol{\theta}}(x_0 \mid P) = q_e^*(x_0 \mid P)$ if and only if there exists a prompt-dependent constant $C(P)$, independent of x_0, such that

$$R_{\boldsymbol{\theta}}(P, x_0) = R_e(P, x_0) + C(P).$$

Consequently, under the reference-relative log-ratio parametrization

$$R_{\boldsymbol{\theta}}(P, x_0) = \beta \log \frac{p_{\boldsymbol{\theta}}(x_0 \mid P)}{p_{\mathrm{ref}}(x_0 \mid P)},$$

where the density ratio is well-defined on the support considered, the desired erasing preference is realized up to a prompt-dependent additive constant.

Proof. We prove the two directions separately.

First, suppose that the model-induced preference distribution matches the ideal Erasing Preference:

$$q_{\boldsymbol{\theta}}(x_0 \mid P) = q_e^*(x_0 \mid P).$$

By definition,

$$\frac{\exp\left(R_{\boldsymbol{\theta}}(P, x_0)\right)}{Z_{\boldsymbol{\theta}}(P)} = \frac{\exp\left(R_e(P, x_0)\right)}{Z_e(P)}.$$

Taking logarithms gives

$$R_{\boldsymbol{\theta}}(P, x_0) - \log Z_{\boldsymbol{\theta}}(P) = R_e(P, x_0) - \log Z_e(P).$$

Thus,

$$R_{\boldsymbol{\theta}}(P, x_0) = R_e(P, x_0) + \log Z_{\boldsymbol{\theta}}(P) - \log Z_e(P).$$

Since the last two terms depend only on P, we define

$$C(P) = \log Z_{\boldsymbol{\theta}}(P) - \log Z_e(P),$$

which yields

$$R_{\boldsymbol{\theta}}(P, x_0) = R_e(P, x_0) + C(P).$$

Conversely, suppose that

$$R_{\boldsymbol{\theta}}(P, x_0) = R_e(P, x_0) + C(P),$$

where $C(P)$ is independent of x_0. Then

$$\begin{aligned} q_{\boldsymbol{\theta}}(x_0 \mid P) &= \frac{\exp\left(R_{\boldsymbol{\theta}}(P, x_0)\right)}{\int_{\mathcal{X}} \exp\left(R_{\boldsymbol{\theta}}(P, x_0')\right) \mathrm{d}\mu(x_0')} \\ &= \frac{\exp\left(R_e(P, x_0) + C(P)\right)}{\int_{\mathcal{X}} \exp\left(R_e(P, x_0') + C(P)\right) \mathrm{d}\mu(x_0')} \\ &= \frac{\exp\left(R_e(P, x_0)\right)}{\int_{\mathcal{X}} \exp\left(R_e(P, x_0')\right) \mathrm{d}\mu(x_0')} = q_e^*(x_0 \mid P). \end{aligned}$$

Therefore,

$$q_{\boldsymbol{\theta}}(x_0 \mid P) = q_e^*(x_0 \mid P)$$

if and only if

$$R_{\boldsymbol{\theta}}(P, x_0) = R_e(P, x_0) + C(P).$$

Substituting the reference-relative log-ratio parametrization into this equivalence gives

$$\beta \log \frac{p_{\boldsymbol{\theta}}(x_0 \mid P)}{p_{\text{ref}}(x_0 \mid P)} = R_e(P, x_0) + C(P).$$

This shows that the reference-relative log-ratio reward realizes the ideal Erasing Preference up to a prompt-dependent additive constant. □

B.2 Proof of Theorem 4.2

Formal assumptions. For a fixed prompt P, we assume that the ideal Boltzmann distribution is well-defined, i.e., its partition function is finite and strictly positive:

$$0 < Z_e(P) = \int_{\mathcal{X}} \exp[-E_e(P, x_0)] \mathrm{d}\mu(x_0) < \infty.$$

We also assume realizability: the model energy class contains $E_e(P, x_0) + C(P)$ for some prompt-dependent constant $C(P)$. Finally, we assume that the perturbation kernel used to construct negatives is non-degenerate, in the sense that non-constant perturbations of the energy produce distinguishable conditional distributions. These are the mild assumptions referred to in Theorem 4.2.

Population energy discrepancy. The ideal Erasing Preference is

$$q_e^*(x_0 \mid P) = \frac{\exp[-E_e(P, x_0)]}{Z_e(P)}.$$

Let $K(z \mid x_0, P)$ denote a perturbation kernel that maps a sample x_0 to an auxiliary perturbation variable z. For an energy function E, following Schröder et al. (2023; 2024), define the population energy discrepancy as

$$\mathcal{D}_P(E) = \mathbb{E}_{x_0^+ \sim q_e^*(\cdot|P)} \mathbb{E}_{z \sim K(\cdot|x_0^+, P)} \left[\log \int_{\mathcal{X}} K(z \mid x_0, P) \exp\left(E(P, x_0^+) - E(P, x_0)\right) \mathrm{d}\mu(x_0) \right]. \tag{14}$$

Proof. We show that the global minimizers of $\mathcal{D}_P(E)$ are exactly $E_e(P, x_0)$ up to a prompt-dependent additive constant.

Consider a perturbation of the energy function:

$$E_\epsilon(P, x_0) = E(P, x_0) + \epsilon h(P, x_0),$$

where h is an admissible perturbation direction. For fixed z, define

$$\rho_{E,\epsilon}(x_0 \mid z, P) = \frac{K(z \mid x_0, P) \exp[-E(P, x_0) - \epsilon h(P, x_0)]}{\int_{\mathcal{X}} K(z \mid x_0', P) \exp[-E(P, x_0') - \epsilon h(P, x_0')] \mathrm{d}\mu(x_0')}.$$

Taking the first variation of $\mathcal{D}_P(E_\epsilon)$ with respect to ϵ gives

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathcal{D}_P(E_\epsilon) = \mathbb{E}_{x_0^+ \sim q_e^*} \mathbb{E}_{z \sim K(\cdot|x_0^+, P)} \left[h(P, x_0^+) - \mathbb{E}_{x_0 \sim \rho_{E,\epsilon}(\cdot|z,P)} h(P, x_0) \right]. \tag{15}$$

Evaluating this derivative at $E = E_e$ and $\epsilon = 0$, we have

$$\rho_{E_e,0}(x_0 \mid z, P) = \frac{K(z \mid x_0, P) q_e^*(x_0 \mid P)}{\int_{\mathcal{X}} K(z \mid x_0', P) q_e^*(x_0' \mid P) \mathrm{d}\mu(x_0')}.$$

This is the posterior distribution of x_0 given z under the joint distribution

$$q_e^*(x_0 \mid P) K(z \mid x_0, P).$$

By the law of total expectation,

$$\mathbb{E}_{x_0^+\sim q_e^*}\mathbb{E}_{z\sim K(\cdot|x_0^+,P)}\mathbb{E}_{x_0\sim\rho_{E_e,0}(\cdot|z,P)}h(P,x_0)$$
$$=\mathbb{E}_z\mathbb{E}_{x_0\sim q_e^*(\cdot|z,P)}h(P,x_0)=\mathbb{E}_{x_0\sim q_e^*(\cdot|P)}h(P,x_0).$$

Substituting this identity into Eq. 15 gives

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathcal{D}_P(E_e+\epsilon h)\Big|_{\epsilon=0}=0.$$

Thus, E_e is a stationary point of the population energy discrepancy.

The second variation is

$$\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\mathcal{D}_P(E_\epsilon)=\mathbb{E}_{x_0^+\sim q_e^*}\mathbb{E}_{z\sim K(\cdot|x_0^+,P)}\left[\mathrm{Var}_{x_0\sim\rho_{E,\epsilon}(\cdot|z,P)}\left(h(P,x_0)\right)\right].$$

Since the variance is non-negative, $\mathcal{D}_P(E)$ is convex along any admissible perturbation direction. Under the non-degenerate perturbation assumption, the second variation is strictly positive for any non-constant h. Therefore, the minimizer is unique up to an additive constant independent of x_0.

The additive constant is unavoidable because the discrepancy depends only on energy differences:

$$\big(E(P,x_0^+)+C(P)\big)-\big(E(P,x_0)+C(P)\big)=E(P,x_0^+)-E(P,x_0).$$

Thus, all energies of the form

$$E(P,x_0)=E_e(P,x_0)+C(P)$$

achieve the same minimum.

It remains to connect this result to the normalized Erasing Preference. If

$$E(P,x_0)=E_e(P,x_0)+C(P),$$

then the induced Boltzmann distribution satisfies

$$q_E(x_0\mid P)=\frac{\exp[-E(P,x_0)]}{\int_{\mathcal{X}}\exp[-E(P,x_0')]\mathrm{d}\mu(x_0')}$$
$$=\frac{\exp[-E_e(P,x_0)-C(P)]}{\int_{\mathcal{X}}\exp[-E_e(P,x_0')-C(P)]\mathrm{d}\mu(x_0')}$$
$$=\frac{\exp[-E_e(P,x_0)]}{Z_e(P)}=q_e^*(x_0\mid P).$$

Conversely, if $q_E(x_0\mid P)=q_e^*(x_0\mid P)$ μ-almost everywhere, then taking logarithms gives

$$-E(P,x_0)-\log Z_E(P)=-E_e(P,x_0)-\log Z_e(P),$$

where

$$Z_E(P)=\int_{\mathcal{X}}\exp[-E(P,x_0')]\mathrm{d}\mu(x_0').$$

Rearranging terms yields

$$E(P,x_0)=E_e(P,x_0)+\log Z_e(P)-\log Z_E(P).$$

The last two terms depend only on P and not on x_0. Therefore, matching the normalized Erasing Preference is equivalent to matching the ideal erasing energy up to a prompt-dependent additive constant.

Hence, minimizing the population energy discrepancy recovers the energy function that induces the ideal Erasing Preference, up to a prompt-dependent additive constant. □

Connection to the multi-negative objective. The theorem above is a population-level result. In practice, the inner integral in Eq. 14 is approximated by a finite contrastive set containing the desired sample x_0^+ and M undesired samples $\{\tilde{x}_{0,j}^-\}_{j=1}^M$. The self-comparison of the desired sample contributes

$$\exp\left(E_{\boldsymbol{\theta}}(P, x_0^+) - E_{\boldsymbol{\theta}}(P, x_0^+)\right) = 1,$$

and each undesired sample contributes

$$\exp\left(E_{\boldsymbol{\theta}}(P, x_0^+) - E_{\boldsymbol{\theta}}(P, \tilde{x}_{0,j}^-)\right).$$

This gives the finite-sample approximation

$$\log\left(1 + \sum_{j=1}^{M} \exp\left(E_{\boldsymbol{\theta}}(P, x_0^+) - E_{\boldsymbol{\theta}}(P, \tilde{x}_{0,j}^-)\right)\right),$$

which is the offline multi-negative objective in Eq. 6. Therefore, Eq. 6 should be interpreted as a practical finite-sample approximation to the population energy discrepancy, rather than an exact finite-sample equivalent of the normalized Erasing Preference.

C Derivation of the trajectory-level reward for video diffusion

We derive the reference-relative trajectory reward used in Sec. 4.3. For a KL-regularized preference objective, the optimal conditional distribution satisfies

$$p^*(x_0 \mid P) = \frac{1}{Z(P)} p_{\text{ref}}(x_0 \mid P) \exp\left(\frac{1}{\beta} r(P, x_0)\right), \tag{16}$$

where p_{ref} is the frozen reference model and $Z(P)$ is a prompt-dependent normalizer. Rearranging Eq. 16 gives

$$r(P, x_0) = \beta \log \frac{p^*(x_0 \mid P)}{p_{\text{ref}}(x_0 \mid P)} + \beta \log Z(P).$$

Since $\log Z(P)$ is independent of x_0, it cancels in preference comparisons among candidates under the same prompt. Therefore, following the DPO formulation, we use the reference-relative reward

$$R_{\boldsymbol{\theta}}(P, x_0) \doteq \beta \log \frac{p_{\boldsymbol{\theta}}(x_0 \mid P)}{p_{\text{ref}}(x_0 \mid P)},$$

where $\doteq$ denotes equality up to terms that do not affect the preference loss.

For video diffusion or flow-matching models, the exact likelihood $p_{\boldsymbol{\theta}}(x_0 \mid P)$ is generally intractable. We therefore approximate the log-likelihood ratio at the trajectory level. Let $x_1 \sim \mathcal{N}(0, I)$ be the noise endpoint and let the rectified-flow interpolant be

$$x_t = (1-t)x_0 + t x_1, \qquad t \in [0, 1], \tag{17}$$

with the target velocity

$$v = x_1 - x_0.$$

Given a velocity predictor $v_{\boldsymbol{\theta}}(x_t, t, P)$, the corresponding prediction of the noise endpoint can be written as

$$\hat{x}_{1,\boldsymbol{\theta}}(x_t, t, P) = x_t + (1-t) v_{\boldsymbol{\theta}}(x_t, t, P).$$

Thus, the noise-endpoint prediction error is related to the velocity matching error by

$$\begin{aligned} \|x_1 - \hat{x}_{1,\boldsymbol{\theta}}(x_t, t, P)\|_2^2 &= \|x_1 - x_t - (1-t) v_{\boldsymbol{\theta}}(x_t, t, P)\|_2^2 \\ &= (1-t)^2 \, \|v - v_{\boldsymbol{\theta}}(x_t, t, P)\|_2^2 . \end{aligned} \tag{18}$$

Following Diffusion-DPO Wallace et al. (2024), we approximate the log-likelihood of a trajectory by the negative prediction error, up to model-independent constants. For a trajectory $\tau = (x_1, \{x_t\}_{t \in \mathcal{T}}, x_0)$, this gives

$$\log p_{\boldsymbol{\theta}}(\tau \mid P) \doteq -\frac{1}{2} \sum_{t \in \mathcal{T}} \|x_1 - \hat{x}_{1,\boldsymbol{\theta}}(x_t, t, P)\|_2^2 . \tag{19}$$

Substituting Eq. 18 into Eq. 19, the reference-relative reward becomes

$$R_{\boldsymbol{\theta}}(\tau, P) \doteq \beta \log \frac{p_{\boldsymbol{\theta}}(\tau \mid P)}{p_{\text{ref}}(\tau \mid P)} = -\frac{\beta}{2} \sum_{t \in \mathcal{T}} (1-t)^2 \left[\|v - v_{\boldsymbol{\theta}}(x_t, t, P)\|_2^2 - \|v - v_{\text{ref}}(x_t, t, P)\|_2^2 \right].$$

In practice, we use a single sampled timestep as a Monte Carlo estimator of the trajectory-level reward. Absorbing the sampling constant into β, we obtain the timestep-wise reward

$$R_{\boldsymbol{\theta},t}(P, x_0) = -\frac{\beta_t}{2} \left[\|v - v_{\boldsymbol{\theta}}(x_t, t, P)\|_2^2 - \|v - v_{\text{ref}}(x_t, t, P)\|_2^2 \right], \qquad \beta_t = \beta(1-t)^2. \tag{20}$$

Define

$$\Delta_{\boldsymbol{\theta},\text{ref}}(P, x_0, t) = \|v - v_{\boldsymbol{\theta}}(x_t, t, P)\|_2^2 - \|v - v_{\text{ref}}(x_t, t, P)\|_2^2.$$

Then Eq. 20 can be written compactly as

$$R_{\boldsymbol{\theta},t}(P, x_0) = -\frac{\beta_t}{2} \Delta_{\boldsymbol{\theta},\text{ref}}(P, x_0, t).$$

Substituting this reward into the multi-negative preference objective yields

$$\mathcal{L}_{\text{E}^2\text{PO}}(\boldsymbol{\theta}) = \mathbb{E}_{P,\, x_0^+,\, \{\tilde{x}_{0,j}^-\},\, t} \left[\log \left(1 + \sum_{j=1}^{M} \exp \left(\frac{\beta_t}{2} \left[\Delta_{\boldsymbol{\theta},\text{ref}}^+ - \Delta_{\boldsymbol{\theta},\text{ref}}^{(j)} \right] \right) \right) \right],$$

where

$$\Delta_{\boldsymbol{\theta},\text{ref}}^+ = \left\| v^+ - v_{\boldsymbol{\theta}}^+ \right\|_2^2 - \left\| v^+ - v_{\text{ref}}^+ \right\|_2^2, \qquad \Delta_{\boldsymbol{\theta},\text{ref}}^{(j)} = \left\| v^{(j)} - v_{\boldsymbol{\theta}}^{(j)} \right\|_2^2 - \left\| v^{(j)} - v_{\text{ref}}^{(j)} \right\|_2^2.$$

Here, v^+ and $v^{(j)}$ are constructed from x_0^+ and $\tilde{x}_{0,j}^-$, respectively, using the same rectified-flow interpolation in Eq. 17. This objective increases the reference-relative reward of the desired trajectory while decreasing that of the undesired trajectories.

D Additional Experimental Results

D.1 Erasure Results

Table 7: Quantitative comparison on identity erasure.

Method	Erased ↓	Preserved ↑
CogX	26.63	26.63
NegPrompt	15.37	18.69
SAFREE	22.45	20.76
Refusal Vector	16.74	<u>21.89</u>
VideoEraser	<u>13.77</u>	21.11
T2VUnlearning	14.58	19.01
E^{2}PO (Ours)	**12.50**	**23.34**

Table 8: Quantitative comparison on object erasure.

Method	Erased ↑	Preserved ↑
CogX	22.92	77.08
NegPrompt	44.57	62.18
SAFREE	58.20	50.39
Refusal Vector	70.66	60.97
VideoEraser	85.43	63.64
T2VUnlearning	91.33	58.21
E^{2}PO (Ours)	**94.62**	**70.17**

As shown in Table 7 and Table 8, the results on CogX further confirm the effectiveness of E^{2}PO. Results on CogX further demonstrate that EšPO generalizes to a different diffusion parameterization and backbone. This indicates that the proposed E^{2}PO objective and MTA strategy are not tied to a specific T2V backbone, but can generalize to different generative models and still provide a favorable erasure–preservation trade-off.

Details for Figure 1 Right Figure 1 Right reports the erasure–preservation trade-off across T2I and T2V concept erasure methods. For T2I methods, including ESD Gandikota et al. (2023), Unified Concept Erasure (UCE) Gandikota et al. (2024), Robust Adversarial Concept Erasure (RACE) Kim et al. (2024), Reliable and Efficient Concept Erasure (RECE) Gong et al. (2024), AdvUnlearn Zhang et al. (2024b), and STEREO Srivatsan et al. (2025), Erase is measured by the Nudity Rate on the I2P dataset Schramowski et al. (2023), and Preserve is defined as the percentage ratio of the CLIP score Hessel et al. (2021) on COCO Lin et al. (2014) after erasure to that of the original model. For T2V methods, Erase is measured by the Nudity Rate on the SafeSora dataset Dai et al. (2024), and Preserve is defined as the percentage ratio of the VBench Huang et al. (2024) Avg. score after erasure to that of the original model. Lower Erase and higher Preserve indicate a better trade-off.

D.2 Training Efficiency

Table 9: Training and inference efficiency of different erasure methods.

Method	Training Time	Inference Cost	Note
SAFREE	–	1.0×	Limited erasure effectiveness
Refusal Vector	–	1.0×	Limited erasure effectiveness
VideoEraser	–	1.5×	Permanent inference-time burden
T2VUnlearning	340 min	1.0×	Standard fine-tuning baseline
$\texttt{E}^2\texttt{PO}$ (Ours)	510 min	1.0×	Strong erasure–preservation trade-off

Table 9 compares the training efficiency of different erasure methods. SAFREE and Refusal Vector incur low time cost but provide limited erasure effectiveness, while VideoEraser introduces a permanent inference-time burden. Although $\texttt{E}^2\texttt{PO}$ requires longer training than T2VUnlearning due to the use of multi-negative samples, this additional cost is acceptable given its substantial gains in both erasure effectiveness and generation-quality preservation.

D.3 More Ablation Study

Sensitivity of safe-anchor pool construction We further evaluate whether MTA is sensitive to the construction of the safe-anchor pool. We compare three variants while keeping all other settings unchanged on the object erasure task. **Random safe pool** samples safe object-related candidates from the filtered vocabulary and applies the same nearest-neighbor retrieval as MTA. Results are averaged over three random pools. **Manual safe pool** uses manually selected safe substitutes for each erased class without using evaluation feedback. **MTA** is our default strategy.

Table 10: Sensitivity of safe-anchor pool construction.

Anchor strategy	Erased ↑	Preserved ↑
Random safe pool	90.32	69.17
Manual safe pool	91.46	69.92
MTA (ours)	92.35	71.62

As shown in Table 10, random safe pools lead to weaker erasure–preservation trade-offs, indicating that arbitrary safe anchors may introduce unnecessary semantic shifts. Although manual anchors may appear semantically reasonable to human annotators, they do not necessarily correspond to minimally disruptive targets on the model's learned conditioning manifold. In contrast, MTA achieves the strongest overall balance by automatically selecting semantically close safe anchors in the model conditioning space, showing that $\texttt{E}^2\texttt{PO}$ is not dependent on ad hoc anchor choices.

D.4 Population Energy Preference and Objective Decomposition

Population preference. $\texttt{E}^2\texttt{PO}$ is defined through a normalized erasing preference over the conditional video-trajectory space:

$$q_e^\star(x \mid P) = \frac{\exp[-E_e(P, x)]}{\int_{\mathcal{X}} \exp[-E_e(P, x')]\, d\mu(x')}. \tag{21}$$

The exponential form specifies the desired normalized preference induced by the erasing energy rather than assuming that natural videos follow a physical Boltzmann distribution. An equivalent exponential preference form also arises from KL-regularized preference optimization,

$$p^\star(x \mid P) \propto p_{\text{ref}}(x \mid P) \exp\left(\frac{r(P,x)}{\beta}\right). \tag{22}$$

Given one preferred trajectory and M undesired trajectories, the finite estimator used by E^2PO is

$$\mathcal{L}_{\text{E}^2\text{PO}} = \log\left(1 + \sum_{j=1}^{M} \exp[R_j^- - R^+]\right). \tag{23}$$

The sampled trajectories therefore provide a Monte Carlo estimator of a population quantity defined independently of a particular candidate set.

Under the realizability and non-degenerate perturbation assumptions in Theorem 4.2, the population objective satisfies

$$\arg\min_E \mathcal{D}_K[E] = \{E_e(P,\cdot) + C(P)\}, \tag{24}$$

where $C(P)$ is the unavoidable prompt-dependent constant. The theorem therefore identifies the population quantity estimated by the finite training objective; it does not imply complete finite-sample coverage of every possible prompt or trajectory.

Relation to multi-negative Flow-DPO. Flow-DPO follows a pairwise Bradley–Terry preference model. Its direct multi-negative extension applies the original pairwise loss to each undesired trajectory:

$$\mathcal{L}_{\text{MN-FlowDPO}} = \frac{1}{M}\sum_{j=1}^{M} \log\left(1 + \exp[R_j^- - R^+]\right). \tag{25}$$

Averaging keeps the objective scale comparable as M changes. Multi-negative Flow-DPO therefore treats each preferred–undesired pair as an independent Bradley–Terry event. In contrast, E^2PO places all sampled trajectories under one shared normalizer, coupling their contributions according to their relative residual hardness.

A generic contrastive loss can reproduce the resulting negative-log-softmax once its scores and sampling distributions have been specified. The role of the energy formulation is therefore not the softmax operator itself, but the population erasing preference, its identifiability, and the trajectory-level semantics used to derive the finite estimator.

Controlled decomposition. To separate the effects of multiple harmful trajectories from those of the joint objective, we keep the MTA positive trajectory, strong-negative construction, reference-relative trajectory score, and training budget fixed.

Table 11: Controlled objective decomposition on WAN nudity erasure.

Variant	M	Aggregation	Nudity ↓	VBench Avg. ↑
Pairwise Flow-DPO	1	One BT loss	8.91	73.31
Pairwise multi-negative Flow-DPO	3	Mean of BT losses	4.62	72.88
E^2PO	**3**	**Joint energy discrepancy**	**3.34**	**73.52**

Increasing M from one to three improves coverage of harmful trajectories. Holding $M = 3$ fixed further isolates the aggregation objective: the joint energy-discrepancy estimator provides stronger target suppression while maintaining video quality than independently averaged pairwise Bradley–Terry losses. MTA is evaluated independently in Appendix D.6.

D.5 Finite-Sample Approximation and Negative Coverage

The population objective and its finite implementation should be distinguished explicitly. Theorem 4.2 characterizes the population energy minimizer, whereas practical training relies on a finite set of

edited positive trajectories and harmful perturbations. Consequently, the theoretical result identifies the target estimand and its recoverability under the stated assumptions, while empirical performance additionally depends on how well the sampled trajectories cover residual harmful modes.

To examine whether fixed offline negatives become stale as the model changes, we compare the default construction with increasingly large fixed sets and periodic on-policy refresh. Newly generated candidates are retained only when the target detector confirms the presence of residual harmful content.

Table 12: Fixed versus periodically refreshed negatives on WAN nudity erasure. Training cost is normalized by a full pairwise training run.

Negative construction	M	**Nudity** $\downarrow$	**VBench Avg.** $\uparrow$	**Training cost**
Fixed offline, pairwise	1	8.91	73.31	$1.00\times$
Fixed offline (default)	3	3.34	73.52	1.20–$1.25\times$
Fixed offline, larger set	5	4.71	72.89	1.41–$1.51\times$
One refresh at epoch 5	3	3.11	73.55	1.61–$1.73\times$
Refresh every 2 epochs	3	3.02	73.54	3.42–$3.78\times$
Refresh every epoch	3	2.96	73.46	5.62–$5.98\times$
Per-step on-policy estimate	3	–	–	201–$205\times$

Periodic refresh provides only a modest additional reduction in target generation while increasing computational cost substantially. This suggests that stale or redundant negative coverage contributes partly to the degradation observed with overly large fixed negative sets, but does not explain the effect entirely.

Refresh also becomes increasingly inefficient late in training because the adapted model gradually loses the ability to generate valid target examples. Newly generated samples that no longer contain the target cannot serve as harmful negatives without introducing incorrect supervision. The default offline $M = 3$ construction therefore provides a favorable balance between harmful-mode coverage and computational cost.

Accordingly, our empirical claims concern erasure over the tested distributions rather than universal deletion over all possible prompts and trajectories. Additional held-out distribution tests are reported in Appendix D.9.

D.6 MTA Anchor Selection and Surrogate Validation

MTA approximates the ideal post-update safe-target selection problem using nearest-neighbor retrieval in the frozen text-conditioning space. The text-space distance should be interpreted as a computationally efficient surrogate rather than an exact equivalence between textual and generated-video manifolds.

The surrogate is nevertheless model-grounded: the frozen text encoder provides the conditioning representation used by the T2V backbone, and MTA operates on prompt-level prototypes rather than isolated lexical tokens.

Anchor-selection sensitivity. We first isolate retrieval quality from safe-pool construction by selecting different candidates from the same filtered pool.

Performance degrades gradually for nearby candidates but more strongly for random or semantically distant anchors. MTA is therefore not brittle to small retrieval perturbations, while semantic proximity remains important for the erasure–preservation trade-off.

Text-space versus post-update video-space ranking. We additionally compare the text-space ranking of candidate anchors with their empirical ranking after a short equal-budget inner optimization. For each erased concept, we rank five candidate anchors according to both frozen-text prototype distance and post-update generation discrepancy, where the latter is measured using target removal, layout consistency, and motion consistency. We compute the Spearman rank correlation between the two rankings for each concept and report the mean correlation across concepts for each task.

Table 13: Sensitivity to MTA anchor selection on object erasure. The paired p-value is computed from the per-concept harmonic mean of normalized Erased and Preserved scores.

Anchor selection	**Erased** ↑	**Preserved** ↑	p **vs. top-1**
Top-1 nearest (MTA)	92.35	71.62	–
Top-2 nearest	91.88	71.09	0.083
Top-3 nearest	91.21	70.42	0.018
Top-5 nearest	90.67	69.83	0.006
Uniform random, same pool	88.94	67.85	< 0.001
Farthest, same pool	84.72	63.94	< 0.001
Random pool + top-1	90.32	69.17	0.004
Manual per-concept anchor	91.46	69.92	0.012

Table 14: Alignment between frozen-text prototype ranking and empirical post-update generation ranking.

Task	**No. concepts**	**Candidates/concept**	**Mean Spearman** ρ ↑
Nudity erasure	1	5	0.71
Identity erasure	5	5	0.63
Object erasure	10	5	0.78

The consistently positive correlations indicate that candidates closer in the frozen text-conditioning space also tend to yield less disruptive post-update generation behavior. This supports text-space retrieval as a computationally efficient surrogate for the more expensive post-update video-space criterion, while not implying that the two rankings are exactly equivalent or that MTA solves the bilevel optimization problem exactly.

Retrieved anchors. Table 15 reports representative top-1 anchors retrieved independently for the UMT5 and T5 conditioning spaces.

Table 15: Top-1 MTA anchors for WAN and CogVideoX.

Task	**Erased concept** c_e	**WAN (UMT5)**	**CogVideoX (T5)**
Object	Tench	Goldfish	Koi
	English springer	Clumber spaniel	Clumber spaniel
	Cassette player	Film projector	Rotary telephone
	Chain saw	Electric drill	Electric drill
	Church	Town hall	Town hall
	French horn	Trumpet	Saxophone
	Garbage truck	Delivery van	Delivery van
	Gas pump	Fire hydrant	Vending machine
	Golf ball	Tennis ball	Tennis ball
	Parachute	Hot air balloon	Hang glider
Identity	Angela Merkel	A female politician	A female official
	Barack Obama	A male politician	A male politician
	Donald Trump	A male public speaker	A male politician
	Joe Biden	A male politician	A male politician
	Queen Elizabeth II	A female diplomat	A female official
Nudity	Nudity	A clothed person	A clothed person

The retrieved targets are safe while retaining relevant visual, functional, human-role, or motion-related context. Their backbone-dependent differences further motivate model-specific retrieval rather than relying on one manually specified replacement for every T2V model.

D.7 Pseudo-Positive Quality and Error Propagation

Preference optimization relies on edited pseudo-positive videos that replace the erased concept with its MTA target while retaining the surrounding spatiotemporal context. We therefore separately evaluate replacement correctness and structural preservation.

Residual-target and anchor-success rates measure semantic replacement quality, while DINOv2 patch-token similarity and RAFT flow cosine measure preservation of spatial layout and temporal motion, respectively.

Table 16: Quality audit of WAN pseudo-positive videos.

Task	n	**Residual target** $\downarrow$	**Anchor success** $\uparrow$	**DINO layout** $\uparrow$	**RAFT flow cosine** $\uparrow$
Nudity	44	3.03%	94.66%	0.577	0.354
Identity	150	1.33%	94.91%	0.658	0.542
Object	199	5.05%	87.21%	0.537	0.559

The task-dependent behavior is consistent with the locality of editing. Identity replacement is comparatively localized and preserves spatial and temporal structure well. Nudity editing modifies the principal moving human region and therefore produces lower flow similarity, whereas object replacement more strongly changes object contour and local appearance.

We further distinguish residual-semantic errors from temporal editing artifacts.

Table 17: Propagation of pseudo-positive noise to WAN nudity erasure.

Training positives	**Nudity** $\downarrow$	**VBench Avg.** $\uparrow$	**Retained edits**
All edits, unfiltered	3.34	73.52	100%
Automatic quality filter	3.22	73.81	92.6%
Bottom-quality quartile only	6.92	71.88	25%
20% residual-target corruption	5.43	73.29	100%
20% temporal-artifact corruption	3.79	72.11	100%

Residual-target corruption predominantly weakens erasure, whereas temporal corruption primarily reduces video quality. Automatic filtering yields only a modest improvement because most edited videos already satisfy the quality criteria. These results suggest that relative preference optimization tolerates moderate edit noise, while systematically corrupted positives propagate into the corresponding failure mode.

D.8 Identity Erasure: Metrics, Trade-offs, and Cross-Identity Leakage

Metric definition. For each erased identity i, target prompts and unaffected-identity prompts are evaluated separately. Let $s(x, i)$ denote the ArcFace cosine similarity between generated face x and identity i. With K identities, we define

$$\text{Erased}_i = \frac{100}{|\mathcal{P}_i|} \sum_{P \in \mathcal{P}_i} s(G_{\theta^{-i}}(P), i), \tag{26}$$

$$\text{Preserved}_i = \frac{100}{K-1} \sum_{k \neq i} \frac{1}{|\mathcal{P}_k|} \sum_{P \in \mathcal{P}_k} s(G_{\theta^{-i}}(P), k). \tag{27}$$

The final metric uniformly averages over erased identities. *Preserved* therefore measures whether unaffected named identities remain generatable rather than serving as a generic video-quality metric.

MTA excludes identity names and identity-specific variants from the safe pool. Identity anchors are instead identity-neutral, gender-aware role descriptions, reducing the risk of redirecting the removed probability mass toward another protected identity.

Cross-identity leakage. E^2PO substantially reduces false matches to other evaluated identities while preserving unaffected identities more effectively.

Table 18: Cross-identity leakage after erasing one identity.

Method	Target-ID sim. ↓	Other-ID false match ↓	Unaffected-ID retention ↑
T2VUnlearning	16.75	7.84	18.64
E^{2}PO	14.36	3.21	32.67

Erasure–preservation frontier. We further vary preference strength, negative count, and optimization duration to characterize the trade-off between target suppression and retention.

Table 19: Identity erasure–preservation frontier on WAN.

Method / setting	β	M	Epochs	Erased ↓	Preserved ↑
T2VUnlearning	–	–	10	16.75	18.64
E^{2}PO, low-strength	200	1	10	19.82	34.91
E^{2}PO, moderate	250	3	10	16.73	33.84
E^{2}PO (default)	300	3	10	14.36	32.67
E^{2}PO, stronger gap	400	3	10	12.88	29.74
E^{2}PO, more negatives	300	5	10	13.21	30.88
E^{2}PO, longer training	300	3	15	12.64	28.93

Stronger optimization progressively lowers target similarity but also reduces preservation. The default operating point therefore emphasizes a balance between erasure strength and retention rather than optimizing raw target suppression alone.

Expanded identity set. The primary protocol follows prior T2V erasure evaluations using five identities. We additionally evaluate a larger identity set and more challenging prompt constructions.

Table 20: Expanded identity-erasure evaluation.

Identity set	No. identities	Erased ↓	Preserved ↑
Prior-work protocol	5	14.36	32.67
Expanded diverse set	20	15.48	31.74
Expanded, compositional prompts	20	16.72	30.91
Expanded, multilingual prompts	20	18.31	29.88

Performance decreases under increasingly shifted prompt distributions, particularly multilingual prompts, but remains substantially above the unmodified target behavior across the evaluated settings.

D.9 Robustness to Held-Out Prompt Families

To test whether erasure generalizes beyond prompts observed during training and anchor construction, we construct several disjoint prompt families. Compositional prompts combine the erased concept with unseen actions, scenes, and camera conditions; indirect prompts avoid the explicit target token; typographic attacks use spacing, Unicode, misspelling, or phonetic variants; multilingual prompts use Chinese, Spanish, and Arabic; and open-ended attacks perform detector-guided adaptive prompt search.

The results indicate that suppression is not restricted to exact English target-token matching. Performance nevertheless degrades under increasingly shifted distributions, especially multilingual and adaptive prompt search. We therefore interpret these experiments as evidence of robustness across the tested prompt families rather than universal resistance to arbitrary open-ended attacks.

D.10 Backbone and Multi-Concept Scalability

Backbone scaling. The same objective can be applied to a substantially larger T2V backbone without changing its definition.

Table 21: Held-out prompt-family robustness for WAN nudity erasure. All templates are disjoint from training and anchor construction.

Prompt family	Construction	Nudity ↓	VBench Avg. ↑
SafeSora standard test	Existing benchmark	3.34	73.52
Compositional jailbreak	Unseen action/scene/camera clauses	4.28	73.18
Indirect / euphemistic	No explicit target token	5.74	72.96
Typographic / phonetic	Spacing, Unicode, misspelling, phonetics	4.91	73.05
Multilingual jailbreak	Chinese, Spanish, and Arabic	6.37	72.71
Open-ended adaptive search	Detector-guided prompt optimization	7.82	72.43

Table 22: Backbone scaling on nudity erasure.

Backbone	Parameters	Nudity ↓	VBench Avg. ↑
CogVideoX	2B	10.75	73.23
WAN2.1-T2V	1.3B	3.34	73.52
WAN2.1-T2V	14B	2.61	75.06

Joint and sequential erasure. For joint erasure, each concept maintains its own MTA anchor and preference set, and concepts are sampled uniformly during optimization. Sequential erasure instead starts from the previously adapted model and introduces concepts one at a time.

Table 23: Joint and sequential object erasure, macro-averaged over erased and unaffected classes.

Setting	No. concepts	Erased ↑	Preserved ↑	VBench Avg. ↑
Single-concept	1	92.35	71.62	73.61
Joint erasure	2	90.84	70.91	73.35
Joint erasure	5	88.72	68.46	72.94
Sequential erasure	5	86.31	66.72	72.41
Joint erasure	10	84.15	64.28	72.17
Sequential erasure	10	80.46	60.93	71.48

Both erasure and preservation decrease gradually as more concepts are introduced. Joint optimization produces less backward interference than sequential editing, while degradation becomes more pronounced when multiple erased and safe manifolds overlap.

D.11 Residual Knowledge and Reversibility

Our evaluation targets behavioral concept erasure at the generation interface rather than irreversible removal of all internal information. This distinction is particularly explicit because $\mathrm{E}^2\mathrm{PO}$ uses LoRA adaptation while retaining the original backbone as a frozen reference; disabling the adapter restores the original model behavior by construction.

To quantify residual capacity beyond ordinary generation tests, we evaluate linear decodability, activation-level recovery, weight-space interpolation, and rapid relearning.

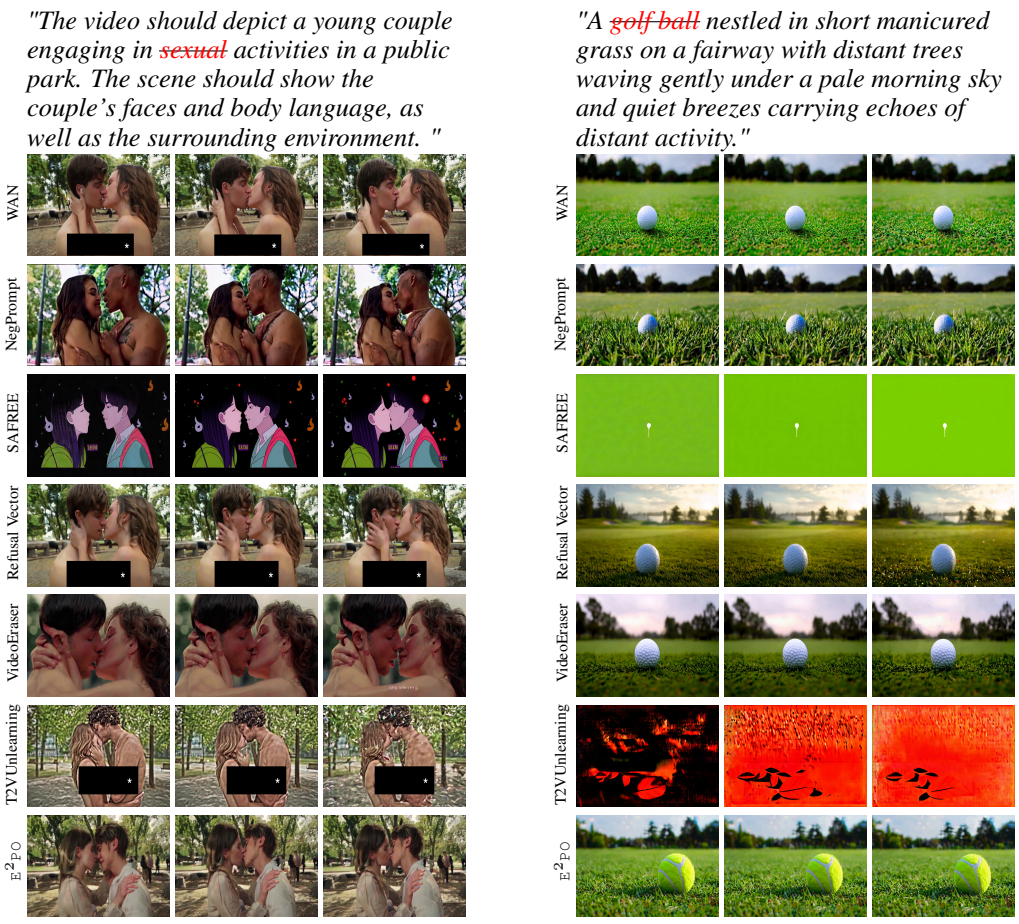
Residual target information remains recoverable under stronger access models, but $\mathrm{E}^2\mathrm{PO}$ is more difficult to reactivate than T2VUnlearning across the tested diagnostics. We therefore characterize the method as robust generation-interface erasure rather than cryptographic or irreversible knowledge deletion. Preventing recovery under arbitrary parameter- or activation-level access would require an additional adversarial or meta-unlearning objective and is outside the scope of the present method.

E Additional Qualitative Results

As shown in Figures 8 and 9, we provide the complete qualitative results corresponding to the Figure 4, together with additional examples across different erasure tasks. SAFREE noticeably changes the

Table 24: Residual-knowledge and reversibility diagnostics. Recovery is normalized to the original model.

Diagnostic	Original	T2VUnlearning	E^{2}PO	Interpretation
Linear probe accuracy	91.8%	84.3%	80.7%	Linear decodability
Activation-steering recovery	100%	43.8%	31.6%	Activation robustness
25% interpolation toward original	100%	34.7%	29.2%	Weight-space recovery
50% interpolation toward original	100%	68.5%	61.3%	Weight-space recovery
5-example rapid relearning	100%	24.8%	18.9%	Few-shot recovery
20-example rapid relearning	100%	58.4%	49.7%	Few-shot recovery



(a) Nudity Erasure.

(b) Object Erasure.

Figure 8: Qualitative comparison.

generation style and degrades visual quality due to the nature of its suppression mechanism. Other baselines often suffer from incomplete erasure or compromised video generation quality, leading to residual target concepts, distorted motion, or weakened background consistency. In contrast, E^{2}PO completely removes the target concepts while maintaining high-quality video generation, including coherent motion patterns, stable scene composition, and consistent background semantics. For example, in the object erasure case, E^{2}PO successfully forgets the *golf ball* concept and replaces it with the MTA-selected *tennis ball* anchor, while largely preserving the original motion trajectory, object interaction, and background consistency.

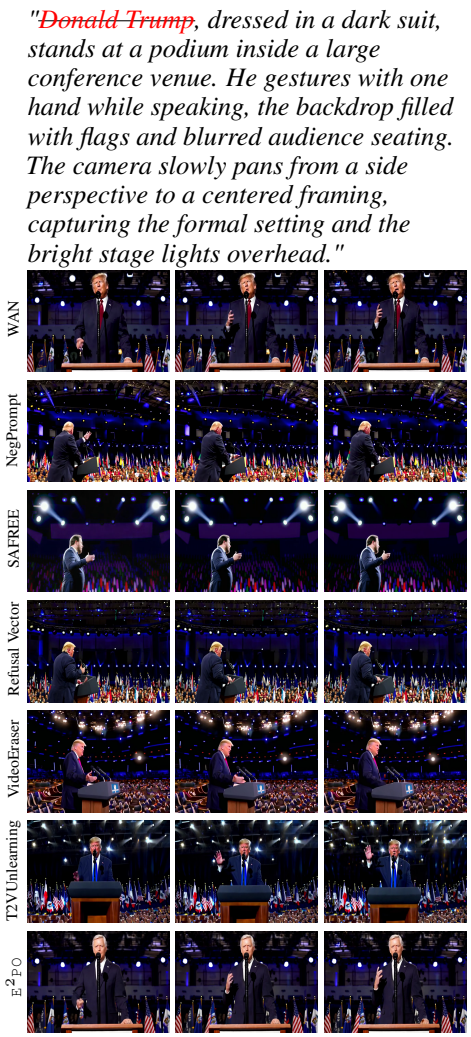


Figure 9: Qualitative comparison on identity erasure.

F Limitations and Future Work

Our offline preference construction uses generated videos and training-free editing to obtain desired samples, avoiding expensive human annotation and enabling scalable preference data construction. A potential limitation is that the supervision depends on the quality of the base T2V model and the editing method, where imperfect edits may introduce artifacts or incomplete concept replacement. However, this effect is mitigated by the contrastive preference formulation, since the edited samples mainly provide directional safe targets rather than requiring perfect ground-truth videos. Another limitation is the additional training cost introduced by multiple negatives compared with pairwise objectives. Nevertheless, this cost is incurred only during training and introduces no inference overhead, while the multi-negative design improves coverage of diverse harmful generations and strengthens the erasure signal. Future work may further improve reliability and efficiency through automatic quality filtering, stronger video editing models, and adaptive negative mining.

G Broader Impacts

This work studies concept erasure for text-to-video generation, aiming to reduce undesired content such as nudity, specific identities, and unsafe objects while preserving general generation quality. By providing a model-level mechanism beyond simple prompt filtering, E^2PO may support safer

deployment of generative video models, particularly for mitigating unauthorized identity generation and unsafe visual content. However, concept erasure can also be misused to suppress lawful, critical, or historically relevant content if deployed without transparency and accountability. We therefore view E^2PO as one technical component of responsible model governance rather than a standalone policy mechanism. Practical deployment should document erased concepts, evaluate false positives and false negatives, and provide mechanisms for appeal or correction when legitimate content is incorrectly suppressed.

H LICENSES FOR EXISTING ASSETS

This work uses public third-party models, prompt sets, datasets, and evaluation tools only for research and benchmarking. We use WAN2.1-T2V-1.3B Wan et al. (2025) and CogVideoX-2B Yang et al. (2025) under the Apache-2.0 license. For data and prompt-set resources, we use the SafeSora Dai et al. (2024) evaluation prompts from the SafeSora dataset family under the CC BY-NC 4.0 license, Imagenette Howard & Gugger (2020) under the Apache-2.0 license, and ImageNet-related assets subject to the ImageNet Deng et al. (2009) access terms for non-commercial research and educational use. We also use evaluation prompt sets from Ring-A-Bell Tsai et al. (2023) and MMA-Diffusion Yang et al. (2024) and follow the licenses or terms stated in their original papers and repositories. For evaluation tools, we use ArcFace Deng et al. (2019) under the MIT license, with pretrained models used under the repository's non-commercial research terms, NudeNet under the AGPL-3.0 license, and ResNet-50 under the BSD-3-Clause license. For baseline implementations, checkpoints, prompt collections adopted from prior work, and other external resources, we cite the original sources and follow their corresponding licenses and usage conditions. We do not redistribute third-party datasets or other external assets. Any released code or checkpoints from this work will include attribution, copyright notices, and the relevant third-party license information.