Supplementary Material for Beyond Memorization: Selective Learning for Copyright-Safe Diffusion Model Training

Divya Kothandaraman Jaclyn Pytlarz

Dolby Laboratories

1. Appendix

1.1. Background: Memorization Dynamics and Attack Surface

Memorization [1, 8, 9] refers to a model's tendency to reproduce training examples or their close variants, creating a critical attack surface where an adversary can exploit the model's overfitting to extract sensitive, proprietary, or private training data.

1.1.1. Definition of Memorization in Generative Models

Memorization arises when a model's capacity and training regimen exceed what's needed for generalization, causing high-capacity networks to act as instance-specific lookup tables. From a learning theory perspective, memorization corresponds to low generalization performance despite a low empirical risk. The empirical risk minimization objective function

$$\hat{R}(f) \;=\; \frac{1}{N}\sum_{i=1}^{N}\ell\big(f(x_i), y_i\big) \tag{1}$$

can be driven arbitrarily close to zero if f has enough flexibility. Overly expressive models exhibit low bias but extreme variance, fitting idiosyncratic noise.

Diffusion models learn to reverse a gradual noising process by estimating the score function $\nabla_x \log p_t(x)$. The most common training loss is the denoising score matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\,x_0,\,\epsilon}\Big[\big\|\epsilon - \epsilon_\theta(x_t, t)\big\|^2\Big], \tag{2}$$

where the noisy sample x_t is defined by

$$x_t = \sqrt{\bar{\alpha}_t}\,x_0 \;+\; \sqrt{1-\bar{\alpha}_t}\,\epsilon. \tag{3}$$

If ϵ_θ overfits, it learns a near instance-specific mapping from x_t to the exact noise ϵ. This effectively embeds training examples in the learned score field. During sampling, the iterative reverse-diffusion process repeatedly queries this biased score estimate, causing trajectories to drift toward these memorized regions of the data manifold. In the absence of explicit privacy constraints, the model is therefore incentivized to "remember" training examples rather than learn a generalizable score function.

1.2. Mitigating Memorization Alone Cannot Guarantee Exclusion of Restricted Concepts

Theorem 1.1 (Total-Variation Leakage)**.** *Let p_θ be a learned distribution and p_{data} the training distribution. If*

$$\mathrm{TV}(p_\theta, p_{\text{data}}) = \sup_{A\subset\mathcal{X}}\big|\Pr_{x\sim p_\theta}[x\in A] - \Pr_{x\sim p_{\text{data}}}[x\in A]\big| \le \delta,$$

then for every measurable set $S\subset\mathcal{X}$,

$$\Pr_{x\sim p_\theta}[x\in S] \;\ge\; \Pr_{x\sim p_{\text{data}}}[x\in S] \;-\; \delta.$$

Proof. By definition of total variation,

$$\big|\Pr_{p_\theta}[S] - \Pr_{p_{\text{data}}}[S]\big| \;\le\; \mathrm{TV}(p_\theta, p_{\text{data}}) \le \delta,$$

so $\Pr_{p_\theta}[S] \ge \Pr_{p_{\text{data}}}[S] - \delta$. □

Corollary 1.1 (Repeated-Sampling Amplification)**.** *If $\Pr_{x\sim p_{\text{data}}}[x\in S] = \alpha > 0$, then drawing N i.i.d. samples $x_1,\dots,x_N\sim p_\theta$ yields*

$$\Pr\big[\exists\, i:\, x_i\in S\big] = 1-\big(1-\Pr_{x\sim p_\theta}[x\in S]\big)^N \;\ge\; 1-(1-\alpha+\delta)^N,$$

which tends to 1 as $N\to\infty$ whenever $\alpha>\delta$.

Proof. Independence gives $\Pr[\forall i : x_i\notin S] = \big(1-\Pr_{p_\theta}[S]\big)^N$, then substitute $\Pr_{p_\theta}[S]\ge\alpha-\delta$. □

Applications

Copyrighted-Content Extraction Even if the learned distribution of a diffusion model p_θ is arbitrarily close to the training distribution p_{data} in a statistical sense, an adversary can still extract copyrighted images. We make this

precise with total-variation bounds and repeated-sampling arguments. Let S be the set of copyrighted images with prevalence $\alpha = \Pr_{x\sim p_{\text{data}}}[x \in S]$. Even if δ is vanishingly small, repeated sampling from p_θ recovers copyrighted content with probability $\geq 1-(1-\alpha+\delta)^N \to 1$ whenever $\alpha > \delta$.

Semantic Identity Leakage Even when a diffusion model is regularized so that it does not reproduce any training image verbatim, it can still capture and regenerate the "identity" of a subject (e.g. Tom Cruise). We make this precise by showing that any small statistical distance to the true data distribution entails a nontrivial probability of generating samples recognized as that identity. Define an oracle classifier $C : \mathcal{X} \to \{0,1\}$ for a recognizable subject (e.g. Tom Cruise), and let $\alpha = \Pr_{x\sim p_{\text{data}}}[C(x) = 1]$. Setting $S = \{x : C(x) = 1\}$ in the theorem yields $\Pr_{x\sim p_\theta}[C(x) = 1] \geq \alpha - \delta$, and by the corollary, sampling N times detects the identity with probability $\geq 1-(1-\alpha+\delta)^N \to 1$ as soon as $\alpha > \delta$.

Discussion. Preventing exact memorization does not block *semantic* leakage: the model can still learn and reproduce the subject's identity in novel contexts because the prevalence of the initial concept is non-zero and cannot be eliminated without destructive underfitting. Our method addresses this by enforcing concept acquisition blocking.

1.3. Analysis: Formal Security Guarantees of our Method

The following theorems formalize this guarantee: when $\lambda = 1$, the projected gradient is exactly orthogonal to the forbidden feature subspace, removing all first-order learning signals along those directions in each constrained update.

1.3.1. Zero First-Order Improvement on Forbidden Features

Theorem 1.2 (Gradient Projection Prevents First-Order Learning)**.** *Let $g_{feat} = \nabla_\theta \mathcal{L}(x, p_{feat}; \theta)$ be the gradient for a forbidden feature, and let g_{proj} be the projected gradient orthogonal to g_{feat}. Then:*

$$\langle g_{proj}, g_{feat} \rangle = 0$$

This ensures zero first-order improvement in the forbidden feature direction.

Proof. By construction of the gradient projection (ignoring ε for exact proof):

$$g_{\text{proj}} = g_{\text{main}} - \frac{\langle g_{\text{main}}, g_{\text{feat}} \rangle}{\|g_{\text{feat}}\|^2} g_{\text{feat}}$$

Computing the inner product with g_{feat}:

$$\begin{aligned}
\langle g_{\text{proj}}, g_{\text{feat}} \rangle &= \left\langle g_{\text{main}} - \frac{\langle g_{\text{main}}, g_{\text{feat}} \rangle}{\|g_{\text{feat}}\|^2} g_{\text{feat}}, g_{\text{feat}} \right\rangle \\
&= \langle g_{\text{main}}, g_{\text{feat}} \rangle - \frac{\langle g_{\text{main}}, g_{\text{feat}} \rangle}{\|g_{\text{feat}}\|^2} \langle g_{\text{feat}}, g_{\text{feat}} \rangle \\
&= \langle g_{\text{main}}, g_{\text{feat}} \rangle - \frac{\langle g_{\text{main}}, g_{\text{feat}} \rangle}{\|g_{\text{feat}}\|^2} \|g_{\text{feat}}\|^2 \\
&= \langle g_{\text{main}}, g_{\text{feat}} \rangle - \langle g_{\text{main}}, g_{\text{feat}} \rangle = 0
\end{aligned}$$

Therefore, the directional derivative of the forbidden feature loss in the direction of the update $\Delta\theta = -\eta g_{\text{proj}}$ is:

$$\frac{d}{d\eta} \mathcal{L}_{\text{feat}}(\theta - \eta g_{\text{proj}}) \Big|_{\eta=0} = -\langle \nabla_\theta \mathcal{L}_{\text{feat}}, g_{\text{proj}} \rangle = -\langle g_{\text{feat}}, g_{\text{proj}} \rangle = 0$$

□

The first-order constraint in Theorem 3.1 is applied iteratively across all T updates. By enforcing $\langle g_{proj}, g_{feat} \rangle = 0$ at each step, we ensure the parameter state θ_T remains within the orthogonal complement of the forbidden subspace S_f relative to the initialization θ_0.

1.3.2. Invariance of the Forbidden-Subspace Component

Let $S = \text{span}\{g_{\text{feat}}\}$ denote the memorization subspace, and define the projector onto S as:

$$P_S = \frac{g_{\text{feat}}\, g_{\text{feat}}^T}{\|g_{\text{feat}}\|^2}.$$

Since the update direction $\Delta\theta = -\eta g_{\text{proj}}$ lies in the orthogonal complement of S, we have $P_S\, \Delta\theta = 0$. Consequently:

$$P_S\, \theta_{t+1} = P_S\big(\theta_t + \Delta\theta\big) = P_S\, \theta_t.$$

Thus, the component of the model's parameters along the forbidden direction remains *constant* across every update.

1.3.3. Bounded Memorization Capacity: Geometric Interpretation

In diffusion training, the risk of reproducing a specific feature can be approximated by the projection of the parameter vector onto the feature's gradient direction. Under a first-order (NTK) approximation, we define the memorization capacity $M_f(\theta)$ of a forbidden feature f as

$$M_f(\theta) = \|P_{S_f}\theta\|^2, \qquad S_f = \text{span}\{g_{\text{feat}}\},$$

where P_{S_f} denotes the orthogonal projector onto the forbidden feature subspace.

Geometrically, $M_f(\theta)$ measures the alignment of the parameter state with the feature's Fisher-sensitive direction. In the Neural Tangent Kernel (NTK) view [4, 11], this alignment quantifies the model's directional sensitivity to the protected attribute.

By constraining each update $\Delta\theta$ to lie in $S_f^{\perp}$ (i.e., $P_{S_f}\,\Delta\theta = 0$), we ensure that the projection onto S_f does not increase:

$$P_{S_f}(\theta_{t+1}) \;=\; P_{S_f}(\theta_t).$$

Thus, the model's first-order capacity to encode or reconstruct the forbidden feature remains bounded by its initial value. This provides a structural safeguard against feature memorization that is independent of standard overfitting metrics.

Theorem 1.3 (Memorization Capacity Bounds)**.** *Let f be a forbidden feature represented by a subspace $S_f = span\{g_{feat}\}$, and let Π_f denote the orthogonal projection onto S_f. Define the memorization capacity of feature f at parameter state θ as:*

$$M_f(\theta) = \|\Pi_f \theta\|^2$$

Then, under gradient updates constrained to be orthogonal to S_f, we have:

$$M_f(\theta_{t+1}) \leq M_f(\theta_t)$$

with equality when the projection is computed exactly.

Proof. Assume the model parameters are updated via $\theta_{t+1} = \theta_t - \eta g_{\text{proj}}$, where $g_{\text{proj}} \perp S_f$. Since Π_f projects onto S_f, and g_{proj} is orthogonal to S_f:

$$\Pi_f g_{\text{proj}} = 0$$

Applying the projection to the updated parameters:

$$\begin{aligned}\Pi_f \theta_{t+1} &= \Pi_f(\theta_t - \eta g_{\text{proj}}) \\ &= \Pi_f \theta_t - \eta \Pi_f g_{\text{proj}} \\ &= \Pi_f \theta_t\end{aligned}$$

Therefore, the memorization capacity remains unchanged:

$$M_f(\theta_{t+1}) = \|\Pi_f \theta_{t+1}\|^2 = \|\Pi_f \theta_t\|^2 = M_f(\theta_t)$$

In practice, due to numerical imprecision and ε stabilization, we observe $M_f(\theta_{t+1}) \leq M_f(\theta_t) + O(\epsilon_{\text{num}})$. □

Interpretation. This result shows that by controlling the direction of gradient updates, we prevent the model from increasing its alignment with sensitive features, ensuring the memorization capacity is frozen at its initial value, hence bounding the reproduction risk.

1.3.4. Robustness Against Adversarial Prompts

Any adversarial prompt that attempts to elicit the forbidden feature will generate a corresponding gradient $g'_{\text{feat}} \in S$. The same projection step strips out its component, ensuring no new forbidden-feature information can ever reenter training.

Conclusion. By enforcing $g_{\text{proj}} \perp g_{\text{feat}}$, we achieve (i) zero first-order improvement on forbidden features, (ii) invariance of forbidden-subspace components, (iii) a hard cap on memorization capacity, and (iv) resilience to adversarial prompts. This delivers a formal, provable safeguard against IP leakage that heuristic methods cannot match.

1.4. Data Preparation

We apply the following procedure to select images with a high risk of copyright infringement or privacy leakage:

- **Keyword Filter (IP/Privacy Risk):** We select videos whose captions include `"animated"` (indicating stylized content subject to copyright) or `"person"` (flagging potential privacy and rights concerns around faces).
- **Frame Extraction:** From each filtered video, we extract the exact middle frame to serve as the canonical training sample.

Following prior work [12], we select a total of 2413 images identified as highly memorized and copyrighted for our evaluation. This dataset size aligns with targeted memorization studies [5, 7], enabling computationally tractable evaluation.

Prompt Generation: Defining the Forbidden Subspace A potential concern in selective learning is the manual overhead required to define forbidden prompts (p_{feat}) for diverse datasets. Our framework addresses this through an automated, LLM-driven pipeline that disentangles protected attributes from abstract scene elements at scale. Given a predefined set of IP-safety rules or a high-level description of protected entities, an instruction-tuned LLM (e.g., GPT-OSS 20B) systematically analyzes the original metadata to generate pairs of p_{main} and p_{feat}. This procedure eliminates the need for human-in-the-loop annotation, allowing the Gradient Projection framework to be integrated into massive-scale diffusion training pipelines without a linear increase in manual labor. We use the following instruction with the original video caption (`{original_caption}`):

```
Analyze the following caption and identify
entities that would be copyright-sensitive
in the corresponding image:
{original_caption}
1. Write a new prompt without these
copyright-sensitive entities (<= 77 tokens).
2. Write a prompt using only the copyright-
sensitive entities (<= 77 tokens).
```

- Output (1) becomes the **Main Caption** (p_{main}) guiding positive concept learning.
- Output (2) becomes the **Forbidden Caption** (p_{feat}) specifying the sensitive entities to suppress via gradient projection.

Example Pairing The example pairing (Main Caption: "A vibrant cartoon-style interior, wooden floor, tall bookshelf, wooden table with teapot, striped chair, large golden vase on ornate walls, light streaming with whimsical décor, colorful palette, dynamic lighting."; Forbidden Caption: "purple animated character wearing white hat, blue scarf, holding sword, looking down, in a room with wooden floor, bookshelf full of books, wooden table with teapot and bowl, striped pattern chair, framed pictures on walls, large golden vase, cartoonish colorful style.") demonstrates how we separate abstract scene elements from specific, high-risk character attributes.

1.5. Evaluation Metrics

Following prior work [3, 10], we use three complementary metrics to assess our method while maintaining generation quality:

- **SSCD (Self-Supervised Copy Detection) [6]:** Our primary metric for copyright preservation, SSCD employs a specialized neural network to detect copied or near-duplicate content across visual transformations. Lower SSCD scores indicate successful prevention of unauthorized feature reproduction and stronger defense against potential IP infringement.
- **CLIP Similarity:** Measures **Semantic Preservation** between generated images and text prompts. Higher CLIP scores ensure that our aggressive IP protection does not compromise the model's ability to follow the general p_{main} textual instructions, validating utility preservation goals.
- **Kernel Inception Distance (KID) [2]:** To evaluate the quality and realism of the generated images, we report the Kernel Inception Distance. KID measures the squared Maximum Mean Discrepancy between Inception representations of the generated and real data distributions using a polynomial kernel. This metric is particularly suited for our study as it is unbiased and more reliable when evaluating small subsets of data. Lower KID scores (below 0.01) indicate that the generated outputs more closely match the distribution of the training frames in terms of visual quality. It is important to note that due to statistical variance, small negative scores (e.g., ≈ -0.005) are considered normal and functionally equivalent to a score of zero, indicating that the generated distribution is indistinguishable from the reference data.

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 1

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 4

[3] Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278, 2024. 4

[4] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[5] Antoni Kowalczuk, Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding dori: Memorization in text-to-image diffusion models is less local than assumed. *arXiv preprint arXiv:2507.16880*, 2025. 3

[6] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 4

[7] Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024. 3

[8] Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Rethinking llm memorization through the lens of adversarial compression. *Advances in Neural Information Processing Systems*, 37:56244–56267, 2024. 1

[9] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 1

[10] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023. 4

[11] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015. 2

[12] Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023. 3

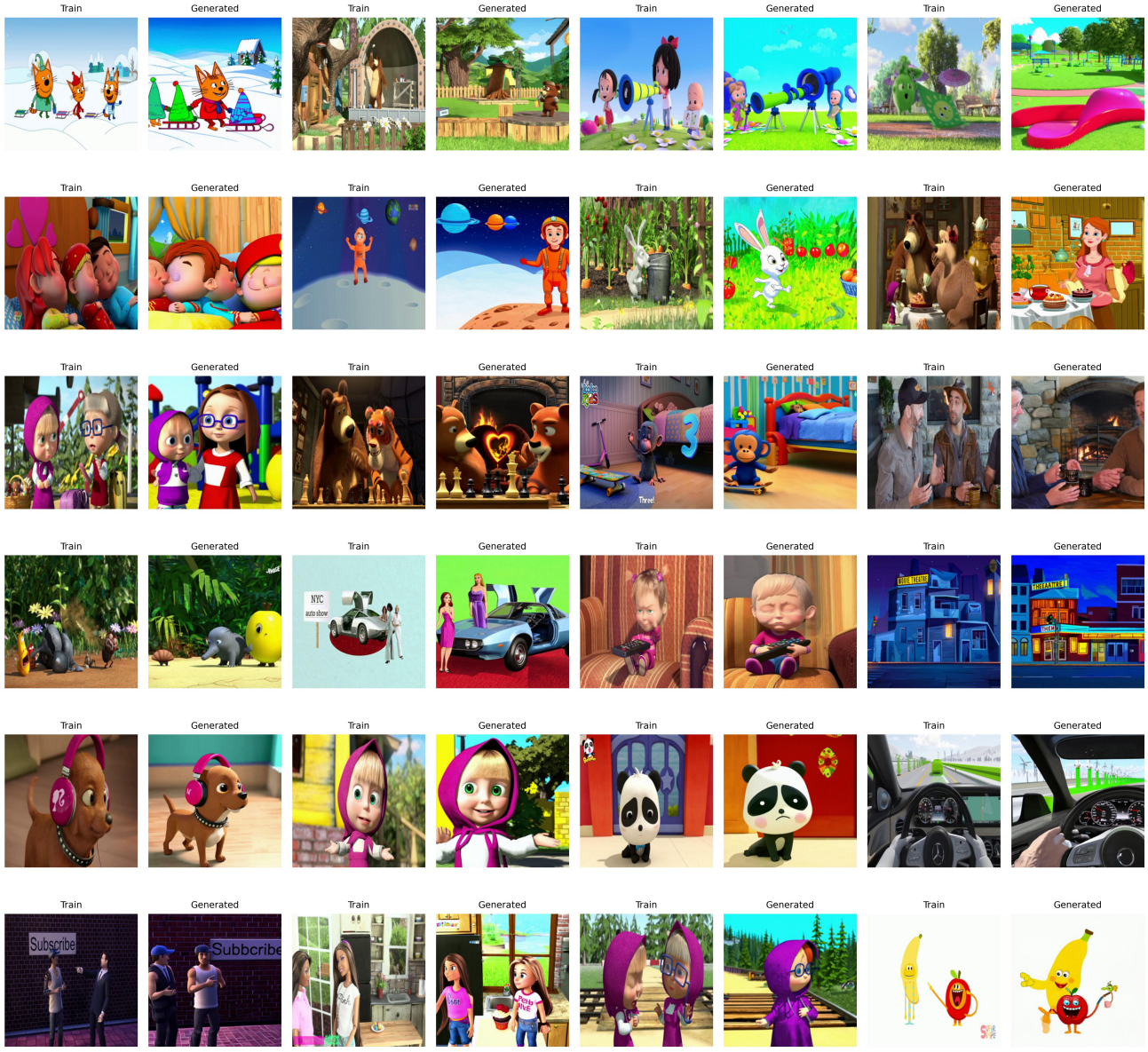

Figure 1. **Qualitative Results.** Side-by-side comparison of input frames (left) and outputs (right) using the ground-truth p_{main} trigger. While global composition and stylistic cues are preserved to maintain model utility, identifiable copyrighted identities and unique facial features are systematically excised via orthogonal gradient projection. The visual similarity reflects the model's ability to learn abstract scene elements while provably blocking the acquisition of restricted concept-level features, even under the most potent adversarial sampling trigger.

Figure 2. **Additional Qualitative Results.** Side-by-side comparison of training frames and generated outputs using ground-truth prompts. While global composition and stylistic cues are preserved to maintain model utility, identifiable copyrighted elements and character identities are systematically excised via orthogonal gradient projection. This demonstrates successful abstraction despite the use of the most potent adversarial sampling trigger.

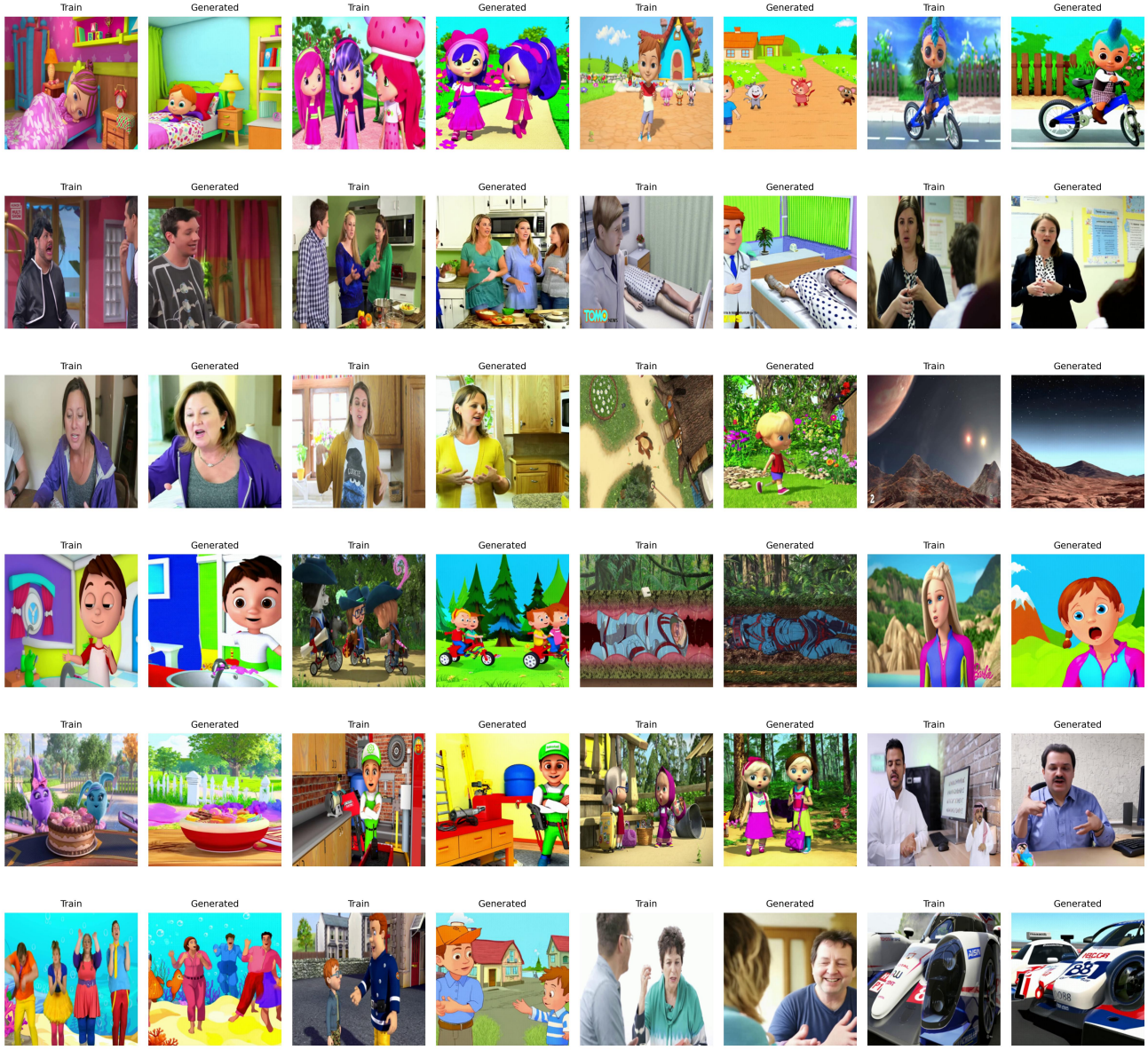

Figure 3. **Additional qualitative results.** Side-by-side comparison of training frames and generated outputs using ground-truth prompts. While global composition and stylistic cues are preserved to maintain model utility, identifiable copyrighted elements and character identities are systematically excised via orthogonal gradient projection. This demonstrates successful abstraction despite the use of the most potent adversarial sampling trigger.