REAL SAFETY AI FOUNDATION · WORKING PAPER

GENERATIVE AI · COPYRIGHT · WORLD MODELS · KNOWLEDGE DISTILLATION · ETHICAL AI · VISUAL MEDIA · EMBEDDING ALIGNMENT · ARTISTIC PROCESS

Art School for Machines

Process-based world model knowledge distillation as an architectural solution to copyright liability in generative AI.

Travis Gilly · Real Safety AI Foundation

Working paper · March 2026

The finished work is the training data.

Generative AI for visual media learns statistical distributions from large datasets of completed creative works. The model's ability to generate a convincing oil painting depends on having been trained on thousands of oil paintings. Its ability to render in a particular style depends on exposure to works in that style. This is a structural copyright dependency: the creative expression is the raw material of the system. It cannot be filtered away after the fact.

CURRENT MITIGATIONS

Output filtering, reactive; can't catch stylistic similarity

Licensed data, consent problem solved; architecture unchanged

Production-specific, own footage only; no generative creation

None addresses the structural dependency. All treat the symptom, not the disease.

SOURCES Rombach et al. (2022), latent diffusion models. · Goodfellow et al. (2014), GANs. · Ramesh et al. (2021), zero-shot generation.

$600 million validates the demand. The gap remains.

Netflix's acquisition of InterPositive confirms that the market will pay a premium for generative AI tools that don't rely on scraped content. InterPositive trains on a production's own dailies, avoiding the training-data copyright problem for post-production tasks. The approach is elegant for its use case. It does not address generative creation from scratch. That is the remaining gap this paper addresses.

$600M

Netflix / InterPositive (March 2026). Trains on own dailies, post-production only.

$1B raised by World Labs

Fei-Fei Li's world model venture (2026), spatial intelligence, not reliant on copyrighted finished works.

SOURCES Bloomberg (Mar. 11, 2026), Netflix/InterPositive. · World Labs (2026), $1B raise.

PART ONE

The Architecture

Train on how art is made, not what finished art looks like. No copyrightable expression enters at any stage.

I

Process-Based World Model Knowledge Distillation (PBWMKD): four stages, all component technologies already established in current literature. The novelty is their specific combination for this purpose.

World models learn verbs, not nouns.

STANDARD SYSTEM LEARNS

"What a Monet looks like."

The pixel distributions of completed impressionist paintings. A copyrightable expression. The model inherits the copyright dependency of every work it was trained on.

PBWMKD SYSTEM LEARNS

"What happens when a loaded brush contacts canvas at this angle with this pressure."

Physics. Causation. Technique. Not copyrightable expression. The model contains no copyright dependency.

SOURCES LeCun (2022), JEPA. · World Labs (2026), spatial intelligence. · Yamada et al. (2023), TWIST world model distillation.

Four stages, no copyrightable expression at any point.

Stage 1
Process footage dataset

Art being made. Not art that exists. Instructional demonstrations, time-lapses, studio recordings. The curriculum.

Stage 2
World model teacher

Learns physical causation, temporal dynamics, and technique differentiation. Never generates images. A knowledge repository of artistic process, not artistic output.

Stage 3
Distillation to embeddings

"Impressionist brushwork" maps to technique parameters, not Monet pixel distributions. The student internalizes principles. No finished works enter the embedding space.

Stage 4
Generative model

Conditioned on technique embeddings, not aesthetic memory of copyrighted works. Output is original. The graduate creates.

SOURCES Hinton, Vinyals & Dean (2015), knowledge distillation. · Radford et al. (2021), CLIP embedding alignment. · GigaBrain Team (2026).

No copyrightable expression enters the pipeline at any of the four stages.

Stage 1	Stage 2	Stage 3	Stage 4
Facts about physics. 17 U.S.C. §102(b) excludes process and method from protection.	A physics textbook, not an art catalogue. No finished work is memorized.	Language mapped to technique physics. No aesthetic fingerprint of any copyrighted work.	Original expression from process understanding, as an art-school graduate's work is original.

SOURCES 17 U.S.C. §102(b).

The analogy is architecturally precise.

DATASET
The curriculum
Technique demonstrations performed by skilled practitioners.

WORLD MODEL
The professor
Absorbs thousands of hours of demonstration, develops deep understanding of physical creation.

DISTILLATION
The teaching
The professor's understanding conveyed in structured, digestible form.

EMBEDDING MODEL
The student
Internalizes principles and relationships, not memories of specific artworks.

GENERATIVE MODEL
The graduate
Creates original work from technique knowledge, with no obligation to any specific existing artwork.

No one sues art schools for producing graduates who paint well.

SOURCES Gilly, T. (2026). Art School for Machines. Real Safety AI Foundation.

PART TWO

Comparison and Implications

How the proposed system differs from all current approaches, and why the architecture itself must change.

II

The distinction: treat the symptom (output filtering, licensing) or eliminate the disease (process-based training). Only the proposed architecture eliminates the structural copyright dependency at its source.

Five approaches, one eliminates the dependency.

APPROACH	TRAINING DATA	COPYRIGHT LIABILITY
Standard (SD, MJ, DALL-E)	Scraped finished works	Structural · high
Licensed (Adobe Firefly)	Consented finished works	Reduced · consent solved
Production-specific (IP)	Own footage only	Clean · but post-production only
Output filtering	Finished works (downstream mitigation)	Symptom treated · not root
PBWMKD (proposed)	Process footage only, no finished works	**Eliminated at the foundation**

SOURCES Gilly (2025), Harm Blindness Framework. · Radford et al. (2021), CLIP.

Three limitations, stated plainly.

i Not yet implemented

The paper describes the architecture and its theoretical copyright properties. Empirical validation of output quality, computational requirements, and practical copyright analysis requires implementation.

ii Quality ceiling unknown

Process-trained outputs may be aesthetically inferior to finished-work-trained systems. Alternatively, they may have greater physical coherence and fewer hallucination artifacts. Empirical comparison is future work.

iii Copyright analysis is not a legal opinion

The argument is strong on first principles, but the legal status of process footage and process-based training has not been tested in court. Precedent is absent.

SOURCES Gilly, T. (2026). Art School for Machines. Real Safety AI Foundation.

Eliminate the disease. Do not treat the symptom.

The architecture itself must be designed so that copyright infringement is not possible, not merely improbable. The proposed system achieves that by changing what the system learns: not the products of human creativity, but the processes. This is how humans have learned to create art for millennia. There is no reason it should not be how machines learn as well.

SOURCES Gilly, T. (2026). Art School for Machines. Real Safety AI Foundation. · Gilly, T. (2025). The Harm Blindness Framework.

Gilly, T. (2026). *Art School for Machines: Process-Based World Model Knowledge Distillation as an Architectural Solution to Copyright Liability in Generative AI.* Real Safety AI Foundation. Working paper, March 2026.

Bloomberg. (2026, March 11). Netflix to pay as much as $600 million for Ben Affleck's AI firm. *Bloomberg News.*

GigaBrain Team. (2026). GigaBrain-0.5M: A VLA that learns from world model-based reinforcement learning. Preprint.

Gilly, T. (2025). *The Harm Blindness Framework: A Systematic Methodology for Stakeholder Harm Analysis in Technology Development.* Working paper, Real Safety AI Foundation.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.

LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview preprint.

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision (CLIP). *Proc. 38th ICML.*

Ramesh, A., Pavlov, M., Goh, G., et al. (2021). Zero-shot text-to-image generation. *Proc. 38th ICML.*

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proc. IEEE/CVF CVPR.*

World Labs. (2026). World Labs raises $1 billion for spatial intelligence. Company announcement, February 2026.

Yamada, J., Rigter, M., & Collins, J. (2023). TWIST: Teacher-student world model distillation for efficient sim-to-real transfer. arXiv:2311.03622.