# Art School for Machines:

Process-Based World Model Knowledge Distillation as an Architectural Solution to Copyright Liability in Generative AI

Travis Gilly
Executive Director, Real Safety AI Foundation
t.gilly@ai-literacy-labs.org

March 2026

# ABSTRACT

Contemporary generative AI systems for visual media are trained on datasets of completed works, creating structural copyright liability that no post-hoc mitigation can fully resolve. This paper proposes an alternative training architecture that eliminates copyright concerns at the pipeline's foundation rather than at its output. The proposed system uses a world model trained exclusively on recordings of artistic creation processes (technique demonstrations, instructional footage, studio recordings) to develop causal understanding of how visual art is physically produced. This process knowledge is then transferred via knowledge distillation into a text embedding model, which maps natural language prompts to technique-physics representations rather than aesthetic-pattern representations derived from copyrighted works. A downstream generative model conditioned on these embeddings produces visual outputs grounded in understanding of artistic method rather than statistical reproduction of existing works. The paper analyzes the copyright chain at each pipeline stage, demonstrating that no copyrightable expression enters the system at any point. The architecture is analogous to art education: the world model is the professor demonstrating technique, the embedding model is the student internalizing principles, and the generative model is the graduate creating original work. All component technologies (world models, knowledge distillation, embedding space alignment, generative architectures) exist independently in current literature; the contribution is their novel combination for the specific purpose of producing copyright-clean generative AI at the architectural level.



*Keywords: generative AI, copyright, world models, knowledge distillation, ethical AI, visual media, embedding alignment, artistic process*

# 1. INTRODUCTION

The relationship between generative artificial intelligence and copyright law has become one of the most contentious intersections of technology and policy in the 2020s. Lawsuits from artists, publishers, and content creators against AI companies have proliferated, with the central complaint being consistent: generative AI systems are trained on copyrighted works without consent, compensation, or attribution, and their outputs compete directly with the creators whose work made the systems possible.

The legal and ethical debates have largely focused on two questions: whether training on copyrighted works constitutes fair use, and whether AI-generated outputs that resemble copyrighted works constitute infringement. These are important questions, but they share a common assumption that this paper challenges. Both questions accept the premise that generative AI systems must be trained on finished creative works. They differ only on whether such training should be permitted and under what conditions.

This paper proposes an alternative: a training architecture in which no copyrighted finished work enters the pipeline at any stage. Rather than training a generative model on what art looks like (the pixel patterns of completed works), the proposed system trains on how art is made (the physical processes, techniques, and causal relationships involved in artistic creation). The distinction is not merely semantic. It represents a fundamentally different approach to what a generative model learns, how it represents that knowledge, and what legal and ethical status its outputs carry.

The proposed architecture draws on four independently established areas of machine learning research: world models, knowledge distillation, embedding space alignment, and generative output architectures. Each component is well-documented in existing literature. The novelty lies in their specific combination and the purpose to which they are directed.

The timing of this proposal is not incidental. In March 2026, Netflix acquired InterPositive, an AI filmmaking company founded by Ben Affleck, for a reported sum of up to $600 million (Bloomberg, 2026). InterPositive's approach trains AI models on a production's own dailies for post-production applications, explicitly avoiding the use of scraped content. The acquisition validates market demand for ethical AI tools in visual media production. However, InterPositive addresses only post-production modification of existing footage; it does not solve the copyright problem for generative creation from scratch. The architecture proposed in this paper addresses that remaining gap.

# 2. BACKGROUND AND RELATED WORK

## 2.1 The Training Data Problem in Generative AI

Current generative AI systems for visual media, including diffusion models (Rombach et al., 2022), generative adversarial networks (Goodfellow et al., 2014), and autoregressive transformers (Ramesh et al., 2021), share a common training paradigm. They learn statistical distributions from large datasets of completed visual works, then generate new outputs by sampling from or reversing the learned distributions.

This paradigm creates what might be called a structural copyright dependency. 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. The finished work is the training data. The creative expression is the raw material. This is not a design choice that can be easily modified; it is the fundamental mechanism by which these systems operate.

Attempts to mitigate the resulting copyright concerns have followed several strategies, none of which address the structural dependency itself. Output filtering compares generated images against databases of known copyrighted works and rejects outputs that exceed similarity thresholds. This approach is reactive, computationally expensive, and unable to catch all forms of similarity, particularly stylistic similarity that does not correspond to any single copyrighted work. Prompt engineering and genericization modify user prompts to reduce the specificity of generated outputs, steering the model away from producing recognizable copies of copyrighted works. This treats the symptom rather than the cause and degrades the utility of the system. Licensed training data approaches, such as those employed by Adobe Firefly and Getty Images' generative tools, use datasets where creators have provided consent or compensation. This is ethically and legally superior to unconsented scraping but does not change the fundamental architecture: the model still learns from finished works and carries whatever copyright implications that entails. Production-specific training, as implemented by InterPositive, trains on footage that belongs to the production itself, limiting the model to post-production tasks on a specific project's own material. This approach is elegant for its use case but fundamentally limited to modification rather than creation.

### 2.2 World Models

World models are AI systems that learn internal representations of how environments operate by observing sensory data, typically video. Unlike classifiers or generators, world models are optimized for understanding causal dynamics: what happens next given what has happened so far, and why.

Recent work in world models has accelerated dramatically. World Labs, founded by Fei-Fei Li, raised $1 billion in early 2026 to develop "spatial intelligence" systems that understand geometry, physics, and dynamics (World Labs, 2026). Yann LeCun's Joint Embedding Predictive Architecture (JEPA) predicts abstract representations rather than raw pixels, learning the structure of environments at a conceptual level (LeCun, 2022). Odyssey, with Pixar co-founder Ed Catmull on its board, has built proprietary camera systems to capture real-world environments, explicitly constructing its own training data rather than relying on scraped sources.

The critical insight for the present proposal is that world models learn verbs, not nouns. A world model trained on video of painting processes does not learn "what a Monet looks like." It learns "what happens when a loaded brush contacts canvas at this angle with this pressure." The former is a copyrightable expression. The latter is physics.

### 2.3 Knowledge Distillation

Knowledge distillation, formalized by Hinton et al. (2015), is a technique for transferring learned representations from a large "teacher" model to a smaller "student" model. The teacher generates

"soft targets," probability distributions that encode not just correct answers but the relational structure between concepts. The student learns from these soft targets, acquiring the teacher's understanding in a more compact form.

Of particular relevance is TWIST (Yamada et al., 2023), which demonstrated teacher-student world model distillation for sim-to-real transfer in robotics. In TWIST, a teacher world model trained on privileged state information supervises a student world model that operates on visual observations. This demonstrates the viability of transferring causal understanding from a world model to a downstream model that operates in a different modality. More recently, GigaBrain-0.5M (2026) demonstrated a vision-language-action model that learns from world model-based reinforcement learning, using the world model's spatiotemporal reasoning as a foundation for downstream task performance.

### 2.4 Embedding Space Alignment

Text embedding models map natural language inputs to high-dimensional vector representations. In current generative AI systems (e.g., CLIP; Radford et al., 2021), these embeddings are aligned with visual representations learned from finished artworks. When a user types "impressionist landscape," the embedding maps to a region of vector space that encodes statistical patterns of what impressionist landscapes look like, derived from training on actual impressionist paintings. The alignment between text embeddings and visual representations is not fixed by the architecture; it is determined by training data and objectives. An embedding model could, in principle, be aligned with any set of representations, including those derived from process knowledge rather than finished-work aesthetics.

## 3. PROPOSED ARCHITECTURE

The proposed system, referred to here as Process-Based World Model Knowledge Distillation (PBWMKD), comprises four primary components arranged in a sequential pipeline.

### 3.1 Stage 1: Process Footage Dataset

The foundation of the system is a curated dataset of video recordings depicting artistic creation processes. This dataset includes instructional recordings across visual media (oil painting, watercolor, acrylic, digital painting, charcoal, pencil, ink, pastel, sculpture, ceramics, glasswork, metalwork, textile arts, printmaking, and others); tutorial and demonstration videos showing specific technique execution; time-lapse recordings of complete creation processes from blank surface to finished work; behind-the-scenes production footage showing cinematographic, lighting, and compositional techniques; animation technique demonstrations (traditional frame-by-frame, rotoscoping, stop-motion, digital rigging, motion capture); and controlled recordings produced in dedicated studios where artists demonstrate specific techniques under documented, reproducible conditions.

The critical constraint is that the dataset consists of process footage, not product databases. The system observes art being made. It does not ingest databases of completed art. A recording of an art class in which an instructor demonstrates impasto technique is fundamentally different, both legally and informationally, from a JPEG of a painting that used impasto technique.

### 3.2 Stage 2: World Model Teacher

A world model is trained on the process footage dataset with optimization objectives focused on causal understanding rather than visual reproduction. The world model learns physical causation (the relationship between tool properties, medium properties, surface properties, and resulting visual properties); temporal dynamics (how visual properties evolve during and after creation); compositional mechanics (how spatial arrangement, proportion, balance, rhythm, and focal management operate as decisions with predictable visual consequences); light interaction (how illumination and surface reflectance interact, learned from observation during creation); and technique differentiation (the ability to distinguish artistic techniques based on their causal properties rather than their visual outputs).

The world model does not generate images or video at any point. It functions exclusively as a knowledge repository of artistic process understanding; a professor who demonstrates but does not produce artwork for exhibition.

### 3.3 Stage 3: Knowledge Distillation to Text Embeddings

The world model's process understanding is transferred to a text embedding model through knowledge distillation. The world model generates soft target representations encoding its causal understanding. The text embedding model is trained to align natural language strings with these soft targets. Through iterative optimization, the embedding space develops a structure in which concepts are organized by process similarity rather than visual output similarity.

After distillation, the text string "impressionist brushwork" maps not to a cluster of pixel distributions extracted from Monet, Renoir, and Pissarro paintings, but to a representation encoding: short, visible individual strokes; thick paint application preserving stroke texture; color mixing achieved through optical blending at viewing distance rather than physical mixing on palette; varied brushwork direction following form contours. These are technique descriptors, not aesthetic fingerprints.

### 3.4 Stage 4: Generative Output Model

A generative model receives the technique-grounded embeddings and produces visual outputs. The generative architecture is not specified by this proposal; diffusion models, GANs, autoregressive transformers, flow-matching models, or any future architecture capable of image or video generation from conditioning signals may be used. The generative model's training objective is to produce outputs that are physically consistent with the technique-based embeddings it receives. Because these embeddings encode process understanding rather than aesthetic patterns, the generative model creates from method knowledge rather than from memorization.

### 3.5 Optional: World Model Feedback Loop

An optional refinement stage uses the world model to evaluate generated outputs for physical consistency. If the generative model produces an image conditioned on "watercolor on cold-pressed paper" that exhibits properties inconsistent with actual watercolor physics, the world

model flags the inconsistency. This feedback can be used to refine the generative model iteratively, improving its physical fidelity over time.

## 4. COPYRIGHT CHAIN ANALYSIS

The central claim of this paper is that the proposed architecture eliminates copyright liability at the structural level. This section examines the copyright status of each pipeline stage.

### 4.1 Stage 1: Process Footage

Video recordings of artistic processes depict factual physical events: a brush moving across a canvas, pigment mixing with solvent, clay being shaped on a wheel. While a recording itself may carry thin copyright as an audiovisual work, the techniques and physical processes it depicts are not copyrightable. One cannot copyright "how to blend oil paints" or "the physical behavior of watercolor on absorbent paper." These are facts about the physical world, and facts are explicitly excluded from copyright protection (17 U.S.C. Section 102(b): "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery").

The use of instructional and tutorial content as training data is further protected by the educational nature of the recordings. Art classes exist to teach technique; technique is, by definition, transferable knowledge rather than copyrightable expression. The proposed system can address recording copyrights through licensing of instructional content, production of original process recordings on controlled soundstages, or use of recordings where the copyright holder has consented to AI training use.

### 4.2 Stage 2: World Model Representations

The world model's internal representations encode physical causation. "Brush pressure of X on surface with texture Y produces mark with property Z" is a statement about physics, not a creative expression. The world model does not memorize any finished artwork; it learns the mechanical and physical relationships that underlie artistic creation. Its representations are more analogous to a physics textbook than to an art catalogue.

### 4.3 Stage 3: Text Embeddings

The distilled embeddings map language to technique physics. These representations contain no information derived from copyrighted finished works. The embedding for "impressionist brushwork" encodes stroke mechanics, not Monet's water lilies. The copyright chain remains clean.

### 4.4 Stage 4: Generated Outputs

Outputs are generated from technique-physics embeddings with no copyrighted expression in the conditioning signal. The generative model has never seen a copyrighted finished work at any stage

of the pipeline. Its outputs are original expressions derived from process understanding, precisely as an art school graduate's paintings are original expressions derived from technique education.

### 4.5 The Art School Analogy

The entire pipeline mirrors the process of art education. The process footage dataset is the curriculum: demonstrations of technique performed by skilled practitioners. The world model is the professor: absorbing thousands of hours of technique demonstration and developing deep understanding of how visual art is physically created. The knowledge distillation pipeline is the teaching process: the professor's understanding is conveyed to the student in a structured, digestible form. The text embedding model is the student: internalizing the professor's knowledge as principles and relationships, not as memories of specific artworks. The generative model is the graduate: creating original work informed by technique knowledge, with no obligation to any specific existing artwork.

No one sues art schools for producing graduates who paint well. The graduates' works are original even though their techniques were learned from others. The proposed system operates on precisely the same principle, implemented computationally.

## 5. COMPARISON WITH EXISTING APPROACHES

Standard generative AI systems (Stable Diffusion, Midjourney, DALL-E) train directly on finished works. The proposed system trains on process footage. Standard systems learn aesthetic patterns; the proposed system learns physical causation. Standard systems carry structural copyright liability; the proposed system eliminates it at the foundation.

Licensed training data approaches (Adobe Firefly, Getty Generative) obtain consent for training on finished works. This addresses the consent problem but not the architectural dependency on finished works. The proposed system does not require consent from finished-work creators because it does not use their works.

Production-specific training (InterPositive/Netflix) trains on a production's own dailies for post-production tasks. It solves "don't use other people's footage" but does not address generative creation. The proposed system solves "don't use anyone's finished works for generation." These are complementary rather than competing approaches.

Output filtering and prompt rewriting attempt to prevent copyrighted outputs after training on copyrighted data. They treat symptoms. The proposed system prevents copyrighted inputs from entering the training pipeline. It eliminates the disease.

## 6. IMPLEMENTATION CONSIDERATIONS

The largest practical challenge is assembling a sufficiently comprehensive process footage dataset. However, the volume of freely available instructional art content is enormous. Platforms hosting art tutorials collectively contain hundreds of thousands of hours of process demonstrations across virtually every visual medium. A hybrid approach, combining licensed existing instructional content with purpose-produced studio recordings, is likely optimal. The proposed system's data

requirements may actually be lower than those of standard generative AI systems, because process footage is informationally denser than finished-work datasets. One hour of video showing an artist painting from start to finish encodes far more about technique than a thousand static images of finished paintings.

The world model component requires an architecture capable of learning causal dynamics from video. Several existing architectures are suitable, including video prediction models, spatiotemporal transformers, and physics-informed neural networks. Cross-modal knowledge distillation from video-trained world model to text embedding model is more complex than same-modal distillation but has been demonstrated in multiple contexts. The generative model must be trained to produce outputs conditioned on technique-physics embeddings, which may require modified training objectives that reward physical consistency.

The proposed system is domain-agnostic. While this paper focuses on visual art and filmmaking, the same architecture applies to any visual media domain: illustration, animation, architectural visualization, game asset creation, fashion design, product design, and others. Each domain requires appropriate process footage, but the pipeline architecture remains constant.

## 7. LIMITATIONS AND FUTURE WORK

Several limitations warrant acknowledgment. First, the proposed system has not been implemented; this paper describes the architecture and its theoretical copyright properties. Empirical validation of output quality, computational requirements, and practical copyright analysis requires implementation.

Second, the quality ceiling of process-trained generative models relative to finished-work-trained models is unknown. It is possible that process-based training produces outputs that are physically accurate but aesthetically inferior to those produced by systems trained on millions of finished masterworks. Alternatively, process-based training may produce outputs with greater physical coherence and fewer of the hallucination artifacts common in current generative systems.

Third, the copyright analysis in this paper, while thorough, is not a legal opinion. The legal status of process footage, technique representations, and outputs generated from process-based training has not been tested in court. The argument is strong on first principles, but precedent is absent.

Future work should include proof-of-concept implementation using existing open-source world model and knowledge distillation frameworks, empirical comparison of output quality between process-trained and finished-work-trained generative models, legal review of the copyright chain analysis by intellectual property attorneys, and exploration of hybrid approaches that combine process-based training with consensually licensed finished-work refinement.

## 8. BROADER IMPLICATIONS

The proposed architecture has implications beyond copyright compliance. A generative AI system that understands how art is physically created rather than what existing art looks like may produce outputs with greater physical coherence, greater responsiveness to technique-specific prompts, and

greater utility for artists who want AI tools that understand their craft rather than tools that approximate their output.

The system also has implications for AI ethics more broadly. The Harm Blindness Framework (Gilly, 2025), a systematic stakeholder analysis methodology, asks "Who gets hurt?" at each stage of system design. Applied to the proposed architecture: artists are not hurt because their finished works are not used as training data; instructional content creators are compensated through licensing; art education is not devalued but rather validated as the paradigm for ethical AI training. The architecture does not replace artists; it builds a tool that understands their craft.

If generative AI is to be integrated into creative industries in a way that is sustainable, legal, and respectful of the human creators who make those industries possible, the training paradigm must change. Filtering outputs is not enough. Licensing inputs is better but insufficient. The architecture itself must be designed so that copyright infringement is not possible, not merely improbable. The system proposed here achieves that by changing what the system learns: not the products of human creativity, but the processes of human creativity.

## 9. CONCLUSION

This paper has proposed a generative AI training architecture that eliminates copyright liability at the structural level by training on artistic process rather than artistic product. Using a world model as a teacher that distills causal technique understanding into text embeddings for generative conditioning, the system produces visual outputs derived from understanding how art is made rather than memorization of what existing art looks like.

All component technologies exist in current literature. The contribution is their novel combination for a specific and urgent purpose: enabling generative AI that does not require training on copyrighted finished works, does not require post-hoc copyright mitigation, and does not require the creative industries to choose between AI capability and creator rights.

The art school analogy is not merely illustrative; it is architecturally precise. The system learns technique from demonstrations, internalizes principles through structured knowledge transfer, and creates original work from that understanding. This is how humans have learned to create art for millennia. There is no reason it should not be how machines learn as well.

## ACKNOWLEDGMENTS

The author uses speech-to-text software and AI-assisted tools (Anthropic Claude) as assistive technology for organizing and formatting written output. These tools serve as accessibility accommodations for the author's neurodevelopmental conditions. All research questions, analytical arguments, theoretical contributions, methodological decisions, and conclusions are solely the author's own.

## 10. REFERENCES

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

Gilly, T. (2025). The Harm Blindness Framework: A systematic methodology for stakeholder harm analysis in technology development. Working paper, Real Safety AI Foundation.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (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 preprint 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. Proceedings of the 38th International Conference on Machine Learning.

Ramesh, A., Pavlov, M., Goh, G., et al. (2021). Zero-shot text-to-image generation. Proceedings of the 38th International Conference on Machine Learning.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

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

Yamada, J., Pertsch, K., Haldar, A., & Lim, J. J. (2023). TWIST: Teacher-student world model distillation for efficient sim-to-real transfer. arXiv preprint arXiv:2311.03622.