Derived Data Abundance: How Multi-Modal Embedding Decomposition Solves the AI Training Data Crisis

A Research Paper on Teleological Constellation Training and the Data Scaling Wall

Chris Royse April 2026

Abstract

The AI industry faces an imminent training data crisis. Research from Epoch AI projects that the stock of usable public human-generated text — approximately 300 trillion tokens — will be exhausted by frontier models between 2026 and 2032, with aggressive overtraining pushing this timeline even earlier. This paper argues that the crisis is misframed: the bottleneck is not raw data volume but *labeled, meaningful, multi-dimensional training data*. I present evidence from two implemented systems: **ClipCannon**, a video understanding and avatar generation pipeline, and **Context Graph**, a Rust-based semantic memory retrieval system. ClipCannon decomposes video through 7 embedding modalities spanning 4,044 dimensions — from a single 16-minute video, its 23-stage analysis pipeline extracted over 12,000 individually labeled data points, sufficient to train a generative model for identity-locked avatar synthesis. Context Graph applies the same principle to pure text, decomposing every stored memory through **13 independent embedding models** — semantic, temporal, causal, lexical, code-structural, graph-relational, entity-level, and more — producing 13 independent representations plus 78 cross-correlations per input, a data multiplication factor of approximately 91× in meaningful training signals. I introduce the concept of **Derived Data Abundance** — the principle that running existing data through N independent embedding models produces N-dimensional labeled training samples at a rate multiplicative to the original data volume — and argue that this approach avoids the model collapse problem inherent in synthetic data generation because the derived data is grounded in real observations, not recursive model outputs. I further argue that this constitutes a new category of compression — **meaning compression** — that is fundamentally more valuable than the weight compression, activation compression, and data compression techniques that dominate current AI research (including Google's TurboQuant, announced March 2026). Where existing compression reduces bits per unit of information, meaning compression increases meaningful signal per unit of raw data, directly addressing the training data wall that threatens to stall the scaling paradigm. I propose that Teleological Constellation Training (TCT), the method that emerged from this work, represents a generalizable framework for transforming the existing stock of human-generated data into vastly larger, richer, and automatically labeled training corpora suitable for next-generation model training.

---

1. Introduction: The Data Wall

1.1 The Scaling Laws and Their Limits

The progress of artificial intelligence over the past decade has been driven by a remarkably simple observation: larger models trained on more data with more compute produce better results. The scaling laws documented by Kaplan et al. (2020) and refined by Hoffmann et al. (2022) formalized this relationship, showing that model performance improves predictably as training compute, model parameters, and dataset size increase in proportion. These laws underpin the tens of billions of dollars currently being invested in AI infrastructure worldwide.

But the scaling paradigm is approaching a structural constraint. While compute continues to scale through hardware improvements and massive datacenter buildouts, and model parameters can be increased indefinitely in principle, the third leg of the scaling tripod — training data — is finite. The stock of public human-generated text on the internet, after filtering for quality and deduplication, amounts to roughly 300 trillion tokens. At current growth rates in training dataset sizes, frontier models will consume this entire stock between 2026 and 2032, with the timeline compressing to as early as 2025–2027 under aggressive overtraining regimes like those used for Meta's Llama 3 (overtrained by a factor of 10×).

This is not a speculative concern. Reports from multiple AI labs indicate that pretraining scaling is already showing diminishing returns. TechCrunch reported in late 2024 that major AI investors, founders, and CEOs confirmed that the methods and expectations used to increase model capabilities over the past five years are now yielding diminishing improvements. The HEC Paris analysis from 2025 noted a growing consensus inside labs that simply adding more data and compute will not produce the capabilities once promised. The data wall is real, and the industry is hitting it now.

1.2 The Synthetic Data Trap

The most commonly proposed solution to data scarcity is synthetic data generation: using existing AI models to generate additional training data. This approach has shown promise in specific applications — autonomous vehicle simulation, healthcare privacy protection, and augmentation of rare-event datasets. Gartner forecasts that by 2030, synthetic data will be more widely used for AI training than real-world datasets.

However, synthetic data carries a fundamental risk: *model collapse*. The landmark 2024 Nature paper by Shumailov et al. demonstrated that AI models trained recursively on AI-generated data experience progressive quality degradation. The models gradually lose information from the tails of the original distribution — first rare patterns disappear (early model collapse), then the output distribution converges toward a narrow, homogeneous mass that bears little resemblance to the original data (late model collapse). This phenomenon was observed across multiple model architectures including large language models, variational autoencoders, and Gaussian mixture models.

The model collapse problem is structural, not incidental. When a model generates syn-

thetic data, it necessarily loses information from the original distribution — the generation process compresses the distribution toward its modes, amplifying common patterns and attenuating rare ones. Training subsequent models on this compressed data amplifies the compression. Over iterations, the feedback loop produces outputs that are increasingly bland, repetitive, and divorced from the richness of real human-generated content. While techniques like data accumulation (mixing synthetic with real data), watermarking, and verification filters can delay collapse, none eliminate the fundamental information loss inherent in recursive generation.

1.3 The Thesis: A Third Path

This paper proposes that the data crisis admits a third resolution, distinct from both "find more raw data" and "generate synthetic data." The resolution is to *derive more meaningful, labeled data from the data we already have*, using multi-modal embedding decomposition.

The key insight is this: a single piece of data — a video clip, an audio recording, a document, an image — contains far more information than any single analysis can extract. Running the same data through multiple independent embedding models, each trained to perceive a different dimension of meaning (visual appearance, semantic content, emotional valence, speaker identity, prosody, sentiment), produces multiple independent labeled representations of the same underlying observation. Each representation is grounded in real data, not generated by a model. Each carries a different dimension of meaning that can serve as a training signal. And the combinatorial explosion of multi-modal labels — a single clip simultaneously labeled for visual content, emotion, speaking style, topic, and identity — creates training data of a richness and specificity that raw text or images cannot match.

I call this principle **Derived Data Abundance**: the observation that the effective training data available from a fixed corpus grows multiplicatively with the number of independent embedding models applied to it, because each model extracts a different dimension of meaning that serves as an independent training signal.

I demonstrate this principle concretely through ClipCannon, a video understanding system that extracts seven embedding modalities from source video, and Teleological Constellation Training (TCT), the training method that emerged from this decomposition. From 16 minutes of interview video, ClipCannon produced over 12,000 individually labeled training samples across 4,044 embedding dimensions, sufficient to train a generative video model that produces identity-locked avatar video indistinguishable from the original subject.

---

2. Background: The Training Data Landscape

2.1 The Stock of Human-Generated Data

The Epoch AI research group estimates the effective stock of quality-adjusted, deduplicated public human text at approximately 300 trillion tokens. This figure accounts for the filtering that modern training pipelines apply: removing low-quality content,

deduplicating near-identical pages, and excluding machine-generated text. The raw volume of internet content is vastly larger, but most of it is noise — spam, boilerplate, machine translations, and increasingly, AI-generated content that pollutes the signal.

The visual data stock is harder to quantify but follows similar dynamics. While the internet contains billions of images and millions of hours of video, most of this content is unlabeled, unstructured, and of variable quality. The datasets that have powered computer vision advances — ImageNet, LAION-5B, WebVid-10M — represent a small fraction of total visual content, curated for quality and labeled (often noisily) for specific tasks.

The critical point is not the raw volume of data but the *amount of labeled, meaningful training signal* available. Raw data is abundant; labeled data with multi-dimensional meaning is scarce. This distinction is central to the argument that follows.

2.2 Current Approaches to Data Enrichment

The AI industry has responded to data scarcity through several strategies, each with significant limitations.

Licensing deals represent the most straightforward approach: pay content creators for access to high-quality human-generated content. News Corp signed a deal with OpenAI exceeding $250 million over five years. Reddit negotiated with Google and OpenAI for $203 million annually. These deals expand the available corpus but do not change its fundamental nature — the data is still text, still one-dimensional in its training signal, and still finite in volume.

Multi-modal pre-training attempts to leverage data from domains beyond text. Models like GPT-4V and Gemini are trained on both image and text data, and research has shown synergy between modalities. However, multi-modal pre-training typically treats each modality as a separate input channel rather than extracting multiple embedding dimensions from the same data.

Synthetic data generation via generative models has shown promise for specific applications but faces the model collapse constraint described above. The Truncated Cross-Entropy loss function, verified reasoning filters, and data accumulation strategies can mitigate but not eliminate the fundamental problem.

Data efficiency improvements — better tokenization, improved architectures, curriculum learning — reduce the amount of data needed for a given performance level. But the gains are incremental, not transformative. Undertraining (growing model parameters while holding dataset size constant) can deliver the equivalent of up to two additional orders of magnitude of compute-optimal scaling, but eventually plateaus.

2.3 What's Missing: Meaning Density

The common thread across all current approaches is that they treat data as a *volume* problem: more tokens, more images, more modalities. But the real constraint is not volume — it is *meaning density*. A billion tokens of low-quality web scrape contain less training signal than a million tokens of expert-curated, richly annotated content.

The scaling laws depend not on raw data volume but on the information content of the training corpus as perceived by the model.

This observation suggests a different strategy: instead of finding or generating more data, extract more meaning from the data we already have. If a single video clip can be decomposed into twelve independently meaningful representations — visual, semantic, emotional, prosodic, temporal, identity-related — then the effective training corpus is twelve times larger in terms of meaningful signals, even though the raw data hasn't changed.

---

3. ClipCannon: A Case Study in Multi-Modal Decomposition

3.1 System Overview

ClipCannon is a video understanding, editing, and avatar generation system implemented as a Python package of 247 source files and approximately 67,585 lines of code. It exposes 58 tools via the Model Context Protocol (MCP), backed by a 23-stage analysis DAG (directed acyclic graph), a tamper-evident provenance chain, and an HMAC-secured credit billing system. Two companion systems — Phoenix, a GPU-native avatar engine, and VoiceAgent, a real-time conversational agent — extend its capabilities to identity-locked video generation and autonomous meeting participation.

For the purposes of this paper, ClipCannon's most relevant capability is its analysis pipeline: a topologically sorted graph of 23 processing stages that decomposes source video into multiple independent embedding spaces, producing thousands of labeled data points from a single input.

3.2 The 23-Stage Analysis DAG

When ClipCannon ingests a source video, it runs the following pipeline stages, each producing a different dimension of analysis:

Required stages (failure aborts the pipeline): probe (format validation and metadata extraction), VFR normalization (constant frame rate conversion), audio extraction, frame extraction, transcription (WhisperX large-v3 with anti-hallucination filtering), and finalize.

Optional stages (failure logged, pipeline continues): source separation (Demucs vocal/instrumental split), visual embedding (SigLIP 1,152-dim per frame), OCR, quality assessment, shot type classification, storyboard generation, scene analysis (SSIM-based boundaries with face detection), semantic embedding (Nomic 768-dim per transcript segment), narrative analysis (Qwen3-8B story beats), prosody analysis (F0, energy, speaking rate per sentence), speaker embedding (WavLM 512-dim), emotion embedding (Wav2Vec2 1,024-dim), reaction detection, acoustic analysis, beat detection, profanity filtering, chronemic analysis, and highlights scoring.

These 23 stages run concurrently where dependency allows (using topological sort and asyncio.gather for level-parallel execution), and produce results into a per-project

SQLite database with 37 tables and 4 vector tables backed by sqlite-vec.

3.3 The Seven Embedding Modalities

From the pipeline's output, seven independent embedding modalities are extracted, spanning 4,044 dimensions:

1. **Visual** (SigLIP-SO400M-patch14-384, 1,152 dimensions): Captures facial appearance, pose, expression, lighting, composition — everything perceptually relevant about a video frame.
2. **Semantic** (Nomic-embed-text-v1.5, 768 dimensions): Captures the meaning of spoken content — topics, vocabulary, argument structure, conceptual density.
3. **Emotion** (Wav2Vec2-large-emotion, 1,024 dimensions): Captures vocal emotion — arousal, valence, dominance — from the raw audio signal, independent of the words being spoken.
4. **Speaker** (WavLM-large, 512 dimensions): Captures voice identity — timbre, formant structure, vocal tract characteristics — a biometric signature of the speaker.
5. **Prosody** (Custom extraction, 12 dimensions): Captures speaking style — fundamental frequency (F0) contour, energy envelope, speaking rate, pitch variation, pause ratios — the musical quality of speech.
6. **Sentiment** (all-MiniLM-L6-v2, 384 dimensions): Captures sentence-level sentiment and intent — whether a statement is assertive, questioning, concessive, emphatic.
7. **Voice** (ECAPA-TDNN, 192 dimensions): Captures a speaker verification embedding — a compact signature used to determine whether two audio samples come from the same person.

Each modality is extracted by a frozen, independently trained model. The models were not trained together; they were developed by separate research groups for separate purposes. Their independence is what gives the decomposition its power: each one captures a dimension of meaning that the others miss, and together they define a multi-dimensional space in which data can be labeled with unprecedented specificity.

3.4 Data Multiplication in Practice: The Santa Dataset

The reference material for the system's demonstration is a 975-second (16-minute, 15-second) interview video of a single subject ("Santa"). From this single video, ClipCannon's pipeline produced the following labeled data:

Modality	Samples	Dimensions	Total Parameters
Visual (SigLIP)	1,725	1,152	1,987,200
Semantic (Nomic)	200	768	153,600
Emotion (Wav2Vec2)	343	1,024	351,232
Speaker (WavLM)	300	512	153,600

Modality	Samples	Dimensions	Total Parameters
Prosody (Custom)	188	12	2,256
Sentiment (MiniLM)	202	384	77,568
Voice (ECAPA-TDNN)	188	192	36,096
FLAME Expression	3,477	100	347,700
Total			**3,109,252**

Additionally, the curated training data pipeline produced 2,362 individual training clips, each automatically labeled with:

- A speaking/breathing classification (derived from transcript word alignment and prosody)
- An emotion tag (derived from the emotion embedding and Plutchik mapping)
- An energy level (derived from prosody analysis)
- Viseme sequences (15 MPEG-4 visemes derived from CMU pronunciation dictionary mapping, with 7,916 real samples)
- Scene boundaries and shot types (derived from visual analysis)
- Speaker diarization labels (derived from speaker embedding)

From 16 minutes of video, the system derived **over 12,000 individually labeled, multi-dimensionally annotated data points**. Each data point carries meaning across multiple embedding spaces simultaneously — a single video clip is simultaneously labeled for what it looks like, what it sounds like, what emotion it conveys, what words are being said, how those words are being delivered, and who is speaking them.

This is the core demonstration of Derived Data Abundance: the ratio of meaningful labeled training samples to raw input data is not 1:1 but orders of magnitude greater, because each embedding model extracts a different dimension of meaning from the same underlying observation.

---

4. Why Derived Data Avoids Model Collapse

4.1 The Structural Difference

The model collapse problem, as characterized by Shumailov et al. (2024), arises from a specific feedback loop: a generative model produces synthetic data, that data is used to train a successor model, and the successor produces data that is used to train the next generation. Each generation compresses the distribution toward its modes, losing tail information irreversibly.

Derived data avoids this feedback loop entirely, for a structural reason: **the data is not generated by the model being trained**. The embedding models (SigLIP, WavLM, Wav2Vec2, Nomic, etc.) are frozen. They do not learn from the data they process. They do not produce synthetic data points. They produce *measurements* — vector representations of real observations that capture specific dimensions of mean-

ing. These measurements are as grounded in reality as the original data; they are simply different projections of the same underlying signal.

The distinction is analogous to the difference between photocopying a document (generative reproduction, subject to cumulative degradation) and measuring its physical properties with different instruments (analytical decomposition, each measurement adding new information without degrading the original). The embedding models are measurement instruments, not generators. Their outputs carry information about the real data, not approximations of it.

4.2 Information Preservation

A key property of well-trained embedding models is that they are approximately information-preserving for the dimensions they are designed to capture. A SigLIP embedding of a face image captures almost everything perceptually relevant about that face — identity, expression, pose, lighting. A WavLM embedding of a voice sample captures almost everything relevant about that voice — timbre, formant structure, speaker characteristics. The embeddings are lossy (they compress high-dimensional inputs to fixed-dimensional vectors), but the loss is concentrated in dimensions that are irrelevant to the model's trained objective.

This means that the derived data preserves the richness and diversity of the original data distribution, including its tails. A rare facial expression that appears only once in 16 minutes of video will produce a rare visual embedding — but the embedding faithfully represents that rare expression, it does not compress it toward the mean. A speaker's unusual prosodic pattern on a single sentence will produce a rare prosody vector — but the vector faithfully represents that pattern.

In contrast, synthetic data generation necessarily smooths over these rarities. A generative model trained on 16 minutes of video will produce outputs that cluster around the most common expressions, voices, and styles. The rare tail events that define individuality and provide the richest training signal are precisely what gets lost.

4.3 Multi-Modal Cross-Validation

The use of multiple independent embedding models provides an additional safeguard against information loss: cross-modal validation. If the visual embedding suggests a happy expression but the emotion embedding suggests sadness, the inconsistency is detectable. If the semantic embedding suggests a question but the prosody embedding suggests a statement, the inconsistency is detectable. Each modality serves as a check on the others.

This cross-validation property is analogous to multi-factor authentication in security. A single embedding model might misrepresent a data point due to noise, domain shift, or model limitation. But the probability that N independent models all misrepresent the same data point in a consistent way decreases exponentially with N. With seven modalities, the effective false-acceptance rate is vanishingly small.

---

5. Teleological Constellation Training: The Method

5.1 From Decomposition to Training

Teleological Constellation Training (TCT) is the training method that emerged from ClipCannon's multi-modal decomposition. It uses the derived data not merely as training inputs but as the *definition of the training target*: the model's goal is to produce outputs that, when re-embedded through the same frozen models used to decompose the reference data, fall within a configurable distance of the reference centroids.

The method has three phases:

Phase 1: Constellation Construction. The reference data (e.g., 16 minutes of video) is decomposed through the seven embedding models. For each modality, all embeddings are L2-normalized, their mean is computed, and the mean is re-normalized to unit length. This produces a *centroid vector* per modality — the geometric center of the target identity in that embedding space. The set of all centroids across all modalities is the *constellation*.

Phase 2: Constellation-Conditioned Training. A generative model (in our case, EchoMimicV3 fine-tuned via LoRA) is trained with loss functions that incorporate distance from the constellation centroids. The model is not just minimizing reconstruction error — it is minimizing distance from the frozen identity definition. This gives the training a *teleological* character: the model's purpose is to orbit the constellation.

Phase 3: Runtime Constellation Guard. At inference time, every generated output is re-embedded through the frozen models and compared to the constellation centroids via cosine similarity. Any output that falls below a per-modality threshold is rejected. The guard provides a hard, verifiable guarantee that the model's outputs remain within the identity manifold — a guarantee that no scalar reward system can match.

5.2 The Four-Level Expression Hierarchy

Beyond identity locking, TCT derives a *behavioral vocabulary* from the decomposed data. This vocabulary is organized into four levels:

Level 0 — Action Units: 17 FLAME expression coefficient indices mapped to FACS Action Units (Ekman & Friesen, 1978), providing a human-interpretable description of facial muscle activations.

Level 1 — Micro-Expression Groups: 40 clusters discovered by K-Means over FLAME expression coefficients, representing the subject's personally distinctive facial configurations — not generic expressions, but *their* expressions.

Level 2 — Expression Skills: 34 named behavioral primitives (genuine_laugh, warm_smile, thoughtful_pause, etc.) with temporal structure (onset/peak/offset phases) and natural language prompts for conditioning the generative model.

Level 3 — Behavioral Constellations: 8 emotional states (warm_conversational, emotional_recall, happy_storytelling, etc.), each expressed as a cycling sequence of

skills. These constellations provide high-level behavioral control: a single API call to set the emotional state produces frame-by-frame micro-expression conditioning.

All four levels are derived from the decomposed data. No manual labeling is required. The behavioral vocabulary is discovered automatically from the clustering of FLAME parameters, and the skills and constellations are compositions of discovered groups. This is Derived Data Abundance applied to behavioral control: the same data that defined the identity also defined 8 controllable emotional states, 34 expression skills, and 40 micro-expression groups.

5.3 Results

From 16 minutes of interview video, TCT produced a system capable of generating avatar video with the following verified properties:

- Visual similarity to the target identity ≥ 0.70 cosine similarity (against a 1,152-dimensional centroid computed from 1,725 reference frames)
- Speaker similarity ≥ 0.80 cosine similarity (against a 512-dimensional voice centroid)
- Voice clone quality ≥ 0.95 Speaker Encoder Cosine Similarity (SECS)
- 8 independently controllable emotional states
- 34 micro-expression skills with temporal phase control
- Real-time inference capability for meeting bot deployment

These results were achieved from a dataset that, measured in traditional terms, is tiny: 16 minutes of video, approximately 2,400 training clips. The leverage comes not from data volume but from data *richness* — the multi-modal decomposition that turned 16 minutes into 12,000+ independently meaningful training signals across 4,044 dimensions.

---

6. Context Graph: Proving Universality Beyond Video

6.1 From Video to Text — The Same Principle

If Derived Data Abundance were limited to video — a medium that naturally combines visual, auditory, and linguistic channels — it would be a useful but narrow technique. To demonstrate that the principle is truly universal, I built a second system: **Context Graph**, a Rust-based semantic memory retrieval system that applies the identical multi-modal embedding decomposition approach to pure text.

Context Graph is a 370,000-line Rust codebase comprising 10 crates, exposing 58 MCP tools over a JSON-RPC 2.0 server. It runs on GPU (targeting NVIDIA RTX 5090 via the HuggingFace Candle framework) and stores all data in RocksDB with 51 column families. Its central insight is identical to ClipCannon's: a single piece of text, when embedded through multiple independent models, reveals multiple dimensions of meaning that are invisible to any single model.

6.2 Thirteen Embedders for Text

Where ClipCannon uses 7 embedding modalities across 4,044 dimensions, Context Graph uses **13 independent embedders** spanning dense, sparse, temporal, causal, entity, graph, code, and token-level representations:

ID	Name	Dimensions	What It Captures
E1	Semantic	1,024D dense	General meaning (e5-large-v2)
E2	Temporal Recent	512D dense	Recency via sinusoidal positional encoding
E3	Temporal Periodic	512D dense	Cyclical patterns via Fourier basis (time-of-day, day-of-week)
E4	Temporal Positional	512D dense	Sequence position within a session
E5	Causal	768D × 2 (asymmetric)	Cause-effect directionality (nomic-embed fine-tuned)
E6	Sparse Lexical	~30K sparse	Exact keyword matching (SPLADE)
E7	Code	1,536D dense	Code patterns and structure (Qodo-Embed-1-1.5B)
E8	Graph	1,024D × 2 (asymmetric)	Connectivity and relationship structure
E9	HDC	10,000-bit → 1,024D	Typo-tolerant matching via hyperdimensional computing
E10	Contextual	768D × 2 (asymmetric)	Paraphrase and contextual equivalence
E11	Entity	768D	Entity knowledge via KEPLER (RoBERTa + TransE)
E12	Late Interaction	128D per token	Token-level ColBERT MaxSim scoring
E13	SPLADE v3	~30K sparse	Learned keyword expansion

Every piece of text stored in Context Graph simultaneously receives all 13 embeddings. The resulting structure — called a **TeleologicalFingerprint** (approximately 46 KB per memory) — preserves 100% of the information from all 13 spaces without any fusion at storage time. Fusion occurs only at query time via Weighted Reciprocal Rank Fusion (RRF), where different weight profiles can emphasize different combinations of embedders depending on the task.

6.3 What 13 Lenses Reveal That 1 Cannot

The power of the approach becomes concrete when you examine what each embedder sees that the others miss:

Cross-embedder anomaly detection. The `search_cross_embedder_anomalies` tool finds memories that score highly in one embedder but poorly in another. A memory that is semantically similar (high E1) but causally unrelated (low E5) is a different kind of relationship than one that is both semantically and causally similar. These "blind spots" — things visible to one lens but invisible to another — are precisely the kind of rich, nuanced labels that traditional single-embedder systems cannot produce.

Causal direction. E5 uses asymmetric embeddings: cause and effect vectors are different, with a 1.2× cause-to-effect boost and 0.8× effect-to-cause dampening. A single semantic embedder (E1) would rank "A causes B" and "B causes A" identically. The causal embedder distinguishes them, adding a training signal that captures the directionality of relationships — something invisible to symmetric embeddings.

Temporal structure. E2, E3, and E4 encode temporal information that is completely absent from semantic content. Two memories with identical text but different timestamps produce different E2 vectors. Two memories from the same time of day but different content produce similar E3 vectors. These temporal signals are independent training dimensions that text-only models cannot access.

Entity knowledge. E11 (KEPLER) maps text into a knowledge graph embedding space using TransE, where relationships between entities are represented as vector translations (relation = tail − head). The `infer_relationship` tool uses this to reason about entity relationships purely through vector arithmetic — a capability that requires entity-level understanding, not just semantic similarity.

Code structure. E7 (Qodo-Embed) understands code at a structural level — function signatures, import patterns, API relationships — in 1,536 dimensions. Text that describes code ("implement a sort algorithm") and the code itself land in nearby regions of E7 space, even though they share almost no surface tokens.

6.4 Emergent Topic Discovery from Multi-Space Agreement

Context Graph's clustering subsystem demonstrates a particularly powerful consequence of multi-embedder decomposition: **emergent topic discovery through cross-space agreement**.

The system runs HDBSCAN clustering independently in each of the 13 embedding spaces, then looks for memories that cluster together across multiple spaces simultaneously. When memories co-cluster in E1 (semantic), E7 (code), and E11 (entity)

but not in E5 (causal), that pattern defines a specific type of topic — one with shared meaning, shared code patterns, and shared entities, but without causal relationships. A different topic might co-cluster in E1 and E5 but not E7 — shared meaning with causal relationships but no code involvement.

The topic detection threshold requires weighted agreement across at least 2.5 embedding spaces (with temporal embedders E2-E4 excluded from topic scoring). This means topics emerge only when multiple independent lenses agree that a cluster exists — providing a robustness guarantee that no single-embedder clustering can match.

The cross-space agreement pattern produces a **13-dimensional topic profile** — a vector where each dimension represents how strongly a topic expresses in that embedding space. These profiles are themselves rich, automatically generated labels that describe the multi-dimensional character of each discovered topic.

6.5 The Teleological Comparator: 78-Dimensional Cross-Correlation

Beyond independent per-space analysis, Context Graph computes the full pairwise cross-correlation between all 13 embedders for each memory, producing C(13,2) = 78 cross-correlation features. Combined with 13 per-space alignment scores and 6 group-level aggregations, the system creates a **TeleologicalVector** that captures not just what each embedder sees individually, but how the embedders' views relate to each other.

This cross-correlation structure is where the multiplicative data enrichment becomes most apparent. From a single text input, the system derives 13 independent views, 78 pairwise relationships between those views, 6 group-level summaries, and an optional Tucker tensor decomposition of the full 13×13×1024 interaction tensor. The information content of this representation is orders of magnitude richer than any single embedding — and it is all derived automatically from frozen models applied to real data.

6.6 Implications for Universality

Context Graph demonstrates that the Derived Data Abundance principle applies equally to text as to video, and that the number of useful embedding dimensions for a single modality can be much larger than initially expected. Where one might assume that text "only needs" a semantic embedder, Context Graph shows that 13 independent lenses — semantic, temporal, causal, lexical, code-structural, graph-relational, entity-level, token-level, typo-robust, paraphrase-aware, and sparse keyword — each capture genuinely independent dimensions of meaning that the others miss.

The system proves that multi-embedder decomposition is not a multimodal technique — it is a **universal data enrichment technique** applicable to any data type for which multiple independent embedding models exist. A single sentence of text, when analyzed through 13 lenses, produces 13 independent labeled representations plus 78 cross-correlations — a data multiplication factor of approximately 91× in terms of meaningful training signals, from a single input.

7. Generalizing the Principle: Derived Data Abundance at Scale

7.1 The Fundamental Argument

The argument for Derived Data Abundance as a solution to the AI training data crisis rests on three observations:

Observation 1: Data is information-dense. A single video contains visual, auditory, linguistic, emotional, prosodic, temporal, and spatial information simultaneously. A single text document contains semantic, syntactic, stylistic, topical, and pragmatic information. Current training approaches typically extract only one or two of these dimensions, leaving the majority of the information unused.

Observation 2: Embedding models extract independent dimensions. Each frozen embedding model is a learned measurement instrument that extracts a specific dimension of meaning from data. Running the same data through N independent embedding models produces N independent training signals. The effective training corpus grows multiplicatively with the number of models applied.

Observation 3: The labels are automatic and grounded. Unlike traditional data labeling (which requires human annotators) or synthetic data generation (which requires generative models), embedding-based labeling is automatic (the frozen model produces the label) and grounded (the label is a measurement of real data, not a model's approximation of what data might look like). This eliminates both the cost of human annotation and the risk of model collapse.

7.2 Scaling the Approach

Consider the implications at internet scale. The web contains hundreds of millions of hours of video, billions of images, and trillions of text documents. If each of these data points were decomposed through even three or four independent embedding models, the effective labeled training corpus would multiply several-fold — not in raw tokens, but in *meaningful, independently labeled dimensions of training signal*.

The computational cost of this decomposition is significant but tractable. Running a SigLIP forward pass costs approximately 10ms per frame on modern GPU hardware. Processing the entire stock of YouTube video (estimated at 800 million hours) at 2 frames per second through a single embedding model would require approximately 4.6 million GPU-hours — substantial, but well within the compute budgets of frontier AI labs that are spending billions on training runs. And the decomposition only needs to be done once per data point; the derived embeddings can be stored and reused across multiple training runs.

7.3 Domain-Specific Applications

The TCT framework is domain-agnostic because it operates entirely in embedding space. The same principle — decompose through frozen models, compute centroids,

constrain to the manifold, validate at runtime — applies wherever independent embedding models exist for the relevant dimensions of meaning:

Language model alignment: Embed model outputs through multiple frozen classifiers (toxicity, factuality, helpfulness, formality). The constellation defines "aligned text" as a region in the joint space of these classifiers. Every output is validated against the constellation boundary before release.

Audio/music generation: Embed generated audio through genre classifiers, mood detectors, audio quality models, and rhythm analyzers. The constellation defines the target style as a geometric constraint.

Scientific discovery: Embed candidate molecules, materials, or designs through multiple property predictors. The constellation defines "viable candidate" as proximity to reference examples across all property dimensions simultaneously.

Code generation: Embed generated code through static analysis models, test-coverage predictors, style classifiers, and vulnerability scanners. The constellation defines "production-quality code" as a measurable region in the joint embedding space of these quality dimensions.

In every case, the raw data is not changed. The training signal is multiplied by extracting more meaning from the same observations.

---

8. The Meaning Compression Thesis

9.1 The Compression Landscape in AI

The AI industry is obsessed with compression — and for good reason. Every major bottleneck in the field is fundamentally a compression problem.

Google's TurboQuant, announced March 2026 and immediately dubbed "the real-life Pied Piper" by the internet, compresses KV-cache memory by 6× with zero accuracy loss using PolarQuant and Quantized Johnson-Lindenstrauss (QJL). The announcement was significant enough to crater memory chip stocks within hours — Samsung and SK Hynix each dropped ~6% in a single session. Cloudflare's CEO called it "Google's DeepSeek moment."

But TurboQuant, like all existing compression algorithms in AI, compresses *representations* — the numerical artifacts that models produce during inference. It makes models cheaper to run. It does not make them smarter, more capable, or more aligned. The same is true of the entire compression taxonomy in modern AI:

Weight compression (quantization, pruning, knowledge distillation) reduces the storage and compute cost of model parameters. GPTQ compresses weights to 4 bits. SparseGPT removes 50% of weights through structured sparsity. Knowledge distillation transfers capability from a large teacher to a small student. All of these compress the model's *knowledge* — the patterns it has learned — into fewer bits.

Activation compression (KV-cache quantization, prompt compression, context distillation) reduces the memory cost of inference. TurboQuant is the state of the art

here. These compress the model's *working memory* — the temporary representations it uses during a single forward pass.

Data compression (tokenization, deduplication, filtering) reduces the volume of training data without losing information. BPE tokenization compresses text into subword units. Deduplication removes near-duplicate documents. Quality filtering removes low-signal data. These compress the *input* — the raw material the model learns from.

DeepMind's research demonstrated that LLMs are themselves powerful compressors — Chinchilla models achieved better compression rates on images and audio than domain-specific algorithms like PNG and FLAC, because language modeling is mathematically equivalent to optimal compression via the equivalence between prediction and coding. As their paper states, prediction and compression are two sides of the same coin, connected by Shannon's information theory. The April 2025 paper "Understanding LLM Behaviors via Compression" formalized this further, showing that LLMs learn syntax first (high-frequency patterns compress easily) and factual knowledge second (rare patterns require more capacity).

9.2 What No Existing Algorithm Compresses: Meaning

Every compression algorithm in the current landscape operates on the same axis: *reducing the number of bits required to represent a fixed amount of information*. TurboQuant compresses KV-cache vectors. GPTQ compresses weights. BPE compresses text. They all ask the same question: how can we store this information in fewer bits?

Derived Data Abundance asks the opposite question: **how can we extract more meaning from a fixed number of bits?**

This is not compression in the traditional sense of reducing size. It is compression in the information-theoretic sense of *increasing the ratio of meaningful signal to raw data*. When ClipCannon runs a 16-minute video through 7 embedding models and produces 12,000+ labeled training samples across 4,044 dimensions, it has not reduced the size of the video. It has increased the *meaning density* of the data by extracting dimensions of semantic content that were present in the raw signal but invisible to any single analysis.

When Context Graph runs a sentence through 13 embedders and produces 13 independent representations plus 78 cross-correlations, it has not compressed the sentence. It has *decompressed its meaning* — revealing the full multi-dimensional semantic structure that a single embedding can only partially capture.

This is a fundamentally different operation from anything in the current compression taxonomy. I propose the term **meaning compression** to describe it: the process of transforming raw data into a representation that captures maximal meaning per unit of information, by projecting through multiple independent measurement models simultaneously.

9.3 The Hierarchy of Compression in AI

With this framing, the compression landscape in AI can be organized into a hierarchy of increasing value:

Level 1 — Bit Compression. Reduce the storage cost of fixed information. Huffman coding, gzip, PNG, FLAC. Lossless. Measured in compression ratio (bits out / bits in). This is classical information theory. Value: saves storage and bandwidth.

Level 2 — Knowledge Compression. Reduce the parameter cost of learned knowledge. Quantization, pruning, distillation. Lossy (but ideally with minimal accuracy degradation). Measured in performance retention at reduced model size. This is what TurboQuant and the entire LLM compression field does. Value: saves compute and memory at inference time.

Level 3 — Meaning Compression. Increase the semantic density of training data by extracting multiple dimensions of meaning from each data point. Multi-modal embedding decomposition, constellation construction, cross-embedder analysis. Measured in meaningful training signals per unit of raw data. This is what Derived Data Abundance provides. Value: *multiplies the effective training data without generating any synthetic data*.

The hierarchy is ordered by value because each level operates on a progressively more fundamental bottleneck. Bit compression saves pennies on storage. Knowledge compression saves dollars on inference. Meaning compression potentially saves *billions on training* — because it addresses the data wall itself, not just the cost of processing data that already exists.

9.4 Why Meaning Compression May Be the Most Valuable

The argument that meaning compression is the most valuable form of compression rests on a simple economic observation: the AI industry's most expensive problem is not inference cost (TurboQuant addresses that) or model size (pruning and distillation address that) — it is the approaching exhaustion of training data.

Frontier model training runs now cost $1–10 billion. The next generation is projected to cost $10–100 billion. These costs are driven primarily by the need for more data, more compute, and more parameters to achieve incremental capability improvements. If the data runs out — as Epoch AI projects between 2026 and 2032 — the entire scaling paradigm stalls regardless of how efficiently you can compress weights or KV-caches.

Meaning compression addresses this directly. If each data point in the existing corpus can be decomposed into N independently meaningful representations, the effective training corpus is N times larger without any increase in raw data volume, any generation of synthetic data, or any risk of model collapse. The cost of the decomposition (running embedding models on existing data) is a one-time fixed cost that is trivial compared to the cost of training a frontier model.

Consider the arithmetic. Context Graph produces 91 meaningful signals per input (13 embeddings + 78 cross-correlations). If applied to the estimated 300 trillion

token stock of public text, the effective training corpus becomes approximately 27 quadrillion meaningful signals — nearly two orders of magnitude larger than the raw token count. And unlike synthetic data expansion, every one of those signals is grounded in a real observation, measured by an independently trained frozen model, carrying no risk of model collapse.

9.5 Compression Duality: TurboQuant and TCT

There is an elegant duality between Google's TurboQuant and Teleological Constellation Training. Both operate on high-dimensional vectors. Both use the mathematical structure of embedding spaces to achieve efficiency. Both are lossless in the dimensions that matter.

But they operate in opposite directions:

- **TurboQuant** takes high-dimensional vectors (KV-cache entries) and compresses them into fewer bits while preserving their information content. Direction: *meaning → fewer bits*.
- **TCT** takes raw data and decompresses it into high-dimensional vectors across multiple embedding spaces, revealing information content that was latent in the original. Direction: *raw data → more meaning*.

TurboQuant makes AI cheaper to run. TCT makes AI cheaper to train. TurboQuant compresses the output side of the pipeline. TCT compresses (in the information-theoretic sense of maximizing signal density) the input side. Together, they address both ends of the AI cost curve.

The Pied Piper comparison that the internet applied to TurboQuant is apt — but it applies to only half the problem. TurboQuant is Pied Piper for inference. Derived Data Abundance, if the thesis of this paper holds, is Pied Piper for training data itself — and training data is the more expensive problem by orders of magnitude.

---

9. Relationship to Existing Work

9.1 Contrastive Learning and CLIP

The multi-modal embedding approach shares foundational ideas with contrastive learning methods like CLIP (Radford et al., 2021), which aligns text and image representations in a shared embedding space. CLIP demonstrated that cross-modal alignment enables powerful zero-shot capabilities — classifying images using arbitrary text labels without task-specific training.

TCT extends this principle in two ways. First, it uses *more than two* modalities simultaneously (seven, in the current implementation), creating a richer constraint surface. Second, it uses the aligned embeddings not just for retrieval or classification but for *constraining generative model outputs at every level* — architecture, loss function, and runtime validation.

9.2 Self-Supervised Learning and V-JEPA

Meta's V-JEPA (Video Joint Embedding Predictive Architecture) learns video representations by predicting missing or masked parts of a video in an abstract representation space, without requiring labels. V-JEPA demonstrates that rich, task-transferable representations can be learned from unlabeled video data — a finding consistent with the Derived Data Abundance thesis that video contains far more information than any single analysis extracts.

TCT differs from V-JEPA in that it uses *multiple independent* embedding models rather than a single unified model, and it uses the embeddings as training *constraints* rather than as pre-training objectives. The approaches are complementary: V-JEPA-style self-supervised pre-training could itself be one of the embedding models used in a TCT constellation.

9.3 Synthetic Data and Data Augmentation

Traditional data augmentation (rotation, cropping, noise injection, back-translation) and modern synthetic data generation (GAN-based, diffusion-based, LLM-based) both aim to increase the effective training corpus. The key difference from Derived Data Abundance is the *source* of the additional data:

- **Data augmentation** creates variations of existing data points (a rotated version of an image, a paraphrased version of a sentence). The variations are grounded in the original data but carry limited new information.
- **Synthetic generation** creates entirely new data points from a model's learned distribution. The new data carries the model's compression artifacts and is susceptible to model collapse.
- **Derived data** extracts new *dimensions of meaning* from existing data points using independent measurement models. Each dimension is a new training signal grounded in the original observation.

The distinction matters because it determines whether the additional data carries new information or amplified noise. Derived data carries new information by construction: each embedding model was trained to extract a different dimension of meaning, so its output is informationally independent of the other models' outputs (to the extent that the underlying training distributions were independent).

---

10. The Unified Theory of Learning Connection

The Derived Data Abundance principle connects to a broader theoretical framework: the Unified Theory of Learning (UTL), which formalizes learning as the product of surprise (Delta-S, how unexpected an experience is) and coherence (Delta-C, how well the experience integrates into existing knowledge).

In the UTL framework, optimal learning occurs when both surprise and coherence are moderate — the learner encounters something unexpected but can make sense of it

within a coherent framework. The multiplicative relationship (L = Delta-S × Delta-C) means that either signal alone is insufficient: pure surprise without coherence is confusion; pure coherence without surprise is boredom.

Multi-modal embedding decomposition serves the UTL equation directly. By providing training data labeled across multiple independent dimensions, the decomposition increases both the surprise signal (each modality reveals a different dimension of the data, providing unexpected information to a model trained on fewer modalities) and the coherence signal (the multiple labels for each data point provide a rich, internally consistent framework for integrating the information). The product of higher surprise and higher coherence produces more efficient learning — which is precisely what TCT demonstrates empirically.

The implication is that the data crisis is not just a volume problem but a *learning efficiency* problem. Models trained on one-dimensional data (text tokens, pixel values) are operating in the low-coherence regime of the UTL curve: each data point carries a single dimension of meaning, so the model must see many data points to extract the multi-dimensional structure of the world. Models trained on multi-dimensionally labeled data operate in the high-coherence regime: each data point carries rich, structured information that the model can integrate efficiently.

This suggests that the scaling laws themselves may shift when training data is multi-dimensionally enriched. The "tokens needed for a given performance level" metric assumes one-dimensional training data. With N-dimensional embedding-enriched data, each token carries N times the meaningful signal, potentially shifting the scaling curve by the same factor.

---

11. Implications for the AI Industry

12.1 The Data Wall Is a Labeling Wall

Reframing the data crisis as a *labeling* crisis rather than a *volume* crisis changes the solution space entirely. The internet contains hundreds of exabytes of data. The problem is not that this data doesn't exist — it is that most of it is unlabeled, unstructured, and one-dimensional from a training perspective. Multi-modal embedding decomposition addresses all three limitations simultaneously: it labels (each embedding is a label), structures (the embeddings define a geometric space), and adds dimensions (each model adds an independent dimension of meaning).

12.2 Embedding Models as Data Multipliers

The implication for AI research priorities is significant. Instead of investing primarily in larger generative models or more compute for pre-training, the industry should also invest in *better and more diverse embedding models* — because each new embedding model that captures a genuinely independent dimension of meaning effectively multiplies the available training data by one factor.

A research program focused on developing 50 or 100 independent, high-quality embedding models across different perceptual and cognitive dimensions — visual, au-

ditory, linguistic, emotional, temporal, spatial, causal, intentional, social — would create the infrastructure for a massive expansion of effective training data without any increase in raw data volume.

12.3 The Infrastructure for Derived Data

Realizing the vision of Derived Data Abundance at scale requires infrastructure that does not yet exist:

Embedding pipelines that can process internet-scale data through multiple models efficiently. ClipCannon demonstrates the pipeline architecture at single-video scale; scaling to billions of data points requires distributed processing and efficient storage of high-dimensional vectors.

Vector databases capable of storing and querying trillions of embedding vectors across dozens of modalities. The sqlite-vec extension used by ClipCannon works at project scale; internet scale requires purpose-built distributed vector storage.

Constellation construction tools that compute and manage multi-modal centroids from large reference datasets. The current implementation handles single-subject constellations; scaling to millions of constellations (one per concept, entity, or behavioral pattern) requires new algorithms for centroid management and hierarchical constellation structures.

Runtime validation systems that can check every model output against its constellation in real time. The current constellation guard runs at approximately 10ms per frame on an RTX 5090; web-scale validation requires parallelized validation infrastructure.

---

12. Limitations and Open Questions

12.1 Embedding Model Quality

The Derived Data Abundance principle assumes that embedding models capture genuinely independent dimensions of meaning. In practice, embedding models trained on overlapping data may produce correlated outputs, reducing the effective information gain. Research is needed to quantify the degree of independence between embedding models and to develop models specifically designed for orthogonal information extraction.

12.2 Computational Cost

Processing existing data through multiple embedding models is computationally expensive. While tractable for single projects or moderate-scale datasets, internet-scale decomposition would require substantial compute investment. The economics depend on whether the derived data produces sufficient training improvements to justify the processing cost.

12.3 Threshold Calibration

The constellation guard requires per-modality thresholds that must be tuned empirically. There is currently no principled method for selecting optimal thresholds from first principles. Research into automatic threshold calibration — perhaps using the spread of the reference embeddings as a guide — would strengthen the framework.

12.4 Scalability of Centroid Computation

The current centroid computation (L2-normalized mean of L2-normalized samples) is simple and effective for small datasets. At scale, with millions of reference points per modality, more sophisticated centroid estimation methods (perhaps robust statistics or mixture models) may be needed to handle multimodal distributions within a single modality.

12.5 The Limits of Embedding Completeness

The argument that "anything you can embed, you can constrain" assumes that embedding models provide approximately complete descriptions of their input domains. This assumption holds reasonably well for mature embedding models (SigLIP for images, WavLM for audio) but may not hold for less-studied modalities (touch, smell, social context). The framework's power is limited by the quality of available embedding models.

---

13. Conclusion

The AI industry's training data crisis is real but misunderstood. The problem is not that we are running out of data — the internet contains more data than any model can consume. The problem is that we are running out of *one-dimensional* data: text treated as token sequences, images treated as pixel arrays, each carrying a single dimension of training signal.

The solution is not to generate more data synthetically (which risks model collapse) or to find new caches of raw human content (which is finite). The solution is to extract more meaning from the data we already have, using the proliferating ecosystem of frozen embedding models as measurement instruments that reveal dimensions of meaning invisible to any single analysis.

ClipCannon demonstrates this principle on video: from 16 minutes of footage, 12,000+ labeled training samples across 4,044 dimensions. Context Graph demonstrates it on text: every memory analyzed through 13 independent embedders producing 13 representations plus 78 cross-correlations — a 91× multiplication in meaningful signal from a single input. Teleological Constellation Training demonstrates the application: a generative model aligned to a specific identity through geometric constraint in multi-modal embedding space, with per-output verification guarantees that no scalar reward system can match.

The broader principle — **Derived Data Abundance** — states that the effective training data available from a fixed corpus grows multiplicatively with the number of independent embedding models applied to it. Each model extracts a different dimension of meaning. Each dimension is a new training signal. Each signal is grounded in real observation, not model generation. Two independent systems — one processing video through 7 modalities in Python, one processing text through 13 embedders in Rust — demonstrate that this principle is universal: it works on any data type, for any number of embedding models, in any combination. It is not a multimodal technique. It is a data enrichment technique that happens to work spectacularly well across modalities, within a single modality, and across data types of all kinds.

The data wall is not a wall. It is a door — and the key is multi-modal embedding decomposition. The data we have, properly decomposed, is more than enough to train the next generation of AI systems. We just need to learn to see all the dimensions of meaning that each data point contains.

References

- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., & Sifre, L. (2022). "Training Compute-Optimal Large Language Models." arXiv:2203.15556. Available at: https://arxiv.org/abs/2203.15556
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). "Scaling Laws for Neural Language Models." arXiv:2001.08361. Available at: https://arxiv.org/abs/2001.08361
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). "AI models collapse when trained on recursively generated data." *Nature*, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y
- Villalobos, P., Ho, A., Cevallos, J., Suárez, A., & Heim, L. (2024). "Will we run out of data? Limits of LLM scaling based on human-generated data." arXiv:2211.04325. Available at: https://arxiv.org/abs/2211.04325
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." arXiv:2103.00020. Available at: https://arxiv.org/abs/2103.00020
- Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.
- Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). "Learning a Model of Facial Shape and Expression from 4D Scans." ACM Transactions on Graphics (TOG), 36(6), Article 194. https://doi.org/10.1145/3130800.3130813
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." arXiv:2106.09685. Available at: https://arxiv.org/abs/2106.09685

- Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). "Sigmoid Loss for Language Image Pre-Training." arXiv:2303.15343. Available at: https://arxiv.org/abs/2303.15343
- Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., & Ballas, N. (2024). "Revisiting Feature Prediction for Learning Visual Representations from Video." Meta AI Research. Available at: https://ai.meta.com/blog/v-jepa-yann-lecun-ai-model-video-joint-embedding-predictive-architecture/
- Royse, C. (2026). "ClipCannon and the Teleological Constellation System: A Unified White Paper on Identity-Aligned Video Understanding, Editing, and Avatar Generation." White Paper, April 2026.
- Royse, C. (2026). "Teleological Constellation Training: Identity-Aligned Video Generation Through Multi-Modal Embedding Anchors." Technical Paper, April 2026.
- Royse, C. (2026). "The Unified Theory of Learning (UTL): A First-Principles Framework for How Learning Actually Works."
- Royse, C. (2026). "Context Graph: A 13-Embedder Semantic Memory Retrieval System." System Documentation, April 2026.
- Google Research. (2026). "TurboQuant: Redefining AI Efficiency with Extreme Compression." Google Research Blog, March 24, 2026. Available at: https://research.google/blog/turboquant-redefining-ai-efficiency-with-extreme-compression/
- Deletang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., & Ortega, P. A. (2024). "Language Modeling Is Compression." In *Proceedings of ICLR 2024*. arXiv:2309.10668. Available at: https://arxiv.org/abs/2309.10668
- Li, C., et al. (2025). "Lossless data compression by large models." *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-025-01033-7

End of paper.