# Toward Meaning-Modulated Physics: A Framework for Dual-Stream AI Training

## Initial Experiments in Mining Multimedia for Physical Grounding

**Alan Harris**
Haptica
alan@haptica.io

### Abstract

The current generation of artificial intelligence models, though rich in linguistic and visual knowledge, remain "brains in a jar": disconnected from the physical reality they describe. This long-standing "Reality Gap," rooted in the Symbol Grounding Problem, prevents the emergence of truly robust, physically intelligent AI.

We introduce a functional pipeline that transforms existing multimedia content into physically-grounded training data for AI systems. Our methodology combines GPU-accelerated multimodal feature extraction with a novel 3-pass LLM reasoning architecture to generate dual-stream Semantic Haptic Representations (SHR). Each SHR captures both objective physics (from acoustic and visual analysis) and context-modulated perception.

Initial experiments with our end-to-end system suggest its potential to produce more factually-grounded physical descriptions than baseline prompting strategies, which tend to hallucinate. This work presents an early-stage methodology for mining latent physical knowledge in existing media. While our validation is limited and our physics descriptors are currently learned approximations, we offer this framework as a potential path toward generating physically-grounded training data, and we invite further research, criticism, and validation from the community.

## 1 Introduction

Current AI systems excel at language and vision tasks but often struggle with physical reasoning. This "Reality Gap," rooted in the Symbol Grounding Problem [2], limits their safe and effective deployment in robotics and real-world applications. Legacy approaches have proven insufficient, falling into paradigms of brute-force simulation or naive signal conversion [11].

Recent analyses highlight the extraordinary scale of this bottleneck. Goldberg estimates that, at current data collection rates, assembling a dataset comparable to those used for training internet-scale vision–language models would take on the order of 100,000 years of real robot experience. This "100,000-year gap" arises because robot training data must couple sensory streams with action and force trajectories—data that do not exist in abundance online. Attempts to bridge the gap via simulation or video inversion face persistent sim-to-real and 3D recovery challenges [12].

Whereas teleoperation fleets and commercial deployments (e.g., package sorters, autonomous taxis) generate incremental "data flywheels," our approach offers a third path. The Semantic Haptic Representation (SHR) pipeline harvests grounded physical event data at scale from existing audiovisual media. By explicitly separating objective physics streams from context-modulated perception streams, SHR produces structured, physically anchored data suitable for training models across diverse domains. In this sense, SHR directly addresses Goldberg's "100,000-year gap" by transforming the world's 35,000 years of internet video into structured training material [12]. At its core, SHR is a dual-stream representation that separates objective physics from context-modulated perception; the conceptual framework and illustrative schema are detailed in Section 5.

## 2 Motivation and Scientific Context

The feasibility of a physically-grounded data layer has emerged from three concurrent technological and scientific developments.

**The Availability of Scalable Sensorimotor Hardware.** High-fidelity haptic engines are now standard in billions of devices, from smartphones and game controllers to VR/AR wearables. The hardware is ubiquitous, but the context-aware content is scarce.

**Data Limitations in Large-Scale Models**. Foundation models have reached the limits of what can be learned from text and images alone. The next leap in capability requires new, richer data modalities. They require grounded, causal training data.

**The Sim-to-Real Gap as a Foundational Challenge.** From robotics to AGI research, there is a clear consensus that the "Reality Gap" is a primary bottleneck. The research community recognizes the need for a generalizable approach.

## 3 The 'Reality Gap': A New Data Modality is Required

The foundational challenge in modern AI is the well-documented "Reality Gap." Current approaches to bridging it fall into two categories:

- ***Pure Simulation***: Generates vast amounts of clean, but often sterile and physically unrealistic, data. It struggles to capture the noisy, complex dynamics of the real world.
- ***Physical Robotics Data***: Generates high-fidelity, real-world data, but is extremely expensive, slow, and dangerous to scale. It is unsuited for capturing the long tail of rare or high-impact events.

This gap is not uniform across domains. While visual realism and rigid-body physics are relatively mature, the largest disparities remain in tactile interaction and contextual understanding. Table 1 summarizes key evidence from recent research, illustrating where simulation succeeds and where the gap to reality is widest.

**Table 1: The Simulation-to-Reality Gap Across Key Domains**

*Description: The following table provides a data-driven quantification of the gap between simulation and reality, synthesized from recent academic research. The evidence shows that while visual and basic physics simulation are relatively mature, the largest gaps, and thus the greatest opportunities, lie in domains related to tactile interaction and contextual understanding.*

| Domain | Simulation Strength | Reality Gap (Mismatch) | Evidence |
|---|---|---|---|
| **Visual Realism** | Photorealistic scenes (NeRFs, splatting) | Moderate: temporal inconsistencies, spurious detections | Qureshi et al. [4] |
| **Physics Accuracy** | Rigid body dynamics, contact models | Moderate: discontinuous multi-contacts | Yoon et al. [5] |
| **Sensor Noise** | Gaussian noise models | Large: complex, non-linear behaviors not captured | Chang et al. [6] |
| **Actuation Control** | Idealized linear dynamics | Moderate: friction, backlash, non-linearities | Tobin et al. [7] |
| **Tactile & Force Feedback** | Simplified contacts | Very large: textures, fine-grained forces computationally prohibitive | Narang et al. [8] |
| **Environmental Variation** | Pre-defined, controlled parameters | Large: 40–60% performance drop in unstructured envs | Josifovski et al. [9] |
| **Social/Semantic Context** | Scripted multi-agent interactions | Very large: lacks theory of mind, intent, emotion | Wang et al. [10] |

## 4. System Architecture and Implementation

Our SHR generation pipeline is an end-to-end working system designed to transform unstructured audiovisual data into structured, physically-grounded annotations. It consists of three integrated stages:

### 4.1 Multimodal Feature Extraction

The extracted feature vector is then processed through a structured, staged reasoning pipeline. This multi-pass approach incrementally integrates narrative, physical, and contextual cues. At a high level, the design ensures that low-level multimodal features are consistently mapped to higher-level physical and perceptual descriptors.

### 4.2 Multi-Pass LLM Reasoning Architecture

The extracted feature vector is then processed through a structured, staged reasoning pipeline. This multi-pass approach incrementally integrates narrative, physical, and contextual cues. At a high level, the design ensures that low-level multimodal features are consistently mapped to higher-level physical and perceptual descriptors.

### 4.3 Output Generation

The final output is a structured SHR object containing the dual-stream representation: the objective physics stream ($a_{\text{phys}}$) and the context-modulated perceptual stream ($a_{\text{perc}}$). The format is machine-readable for model training and human-interpretable for validation.

The current implementation processes video at slower-than-real-time speeds, though architectural optimization is expected to significantly improve throughput. The system is designed for scalable dataset generation, with ongoing refinements focused on reducing latency and improving efficiency.

**Table 2: Pipeline Performance Metrics**

| Aspect | Current Implementation | Future Direction |
|---|---|---|
| **Processing Speed** | Sub-real-time | Toward real-time |
| **Hardware Profile** | Datacenter-class GPUs | Optimized multi-scale systems |
| **Dataset Throughput** | Pilot-scale | Large-scale generation |

### 4.4 Physics Descriptor Generation

The physics descriptors in our current implementation are generated through model-based interpretation of multimodal features rather than direct measurement. These descriptors should be understood as structured estimates of underlying dynamics, not direct ground-truth signals.

The process involves three integrated stages:

1. **Signal Analysis**: Low-level acoustic and visual signals are analyzed to extract indicators of force, texture, and motion.
2. **Cross-Modal Correlation**: Detected cues are aligned across modalities to ensure consistency (e.g., an audible impact is matched with a corresponding visual collision).
3. **Descriptor Synthesis**: The aligned features are synthesized into a coherent representation of plausible physical properties such as intensity, duration, and material interaction.

While direct physical validation remains a goal for future work, the system currently emphasizes internal consistency, cross-modal agreement, and temporal continuity as key safeguards for robustness. These safeguards not only stabilize SHR's descriptors technically but also frame the broader question: what kind of data layer is required to build reliable physical intelligence?

## 5 A Foundational Layer for Physical Intelligence

Reliable robotic deployments today require more than raw data—they require structured safeguards and engineered reliability. Our approach shares this spirit with what Goldberg has described as "good old-fashioned engineering" (GOFE) [12], but extends it into the data layer itself: SHR provides a scalable substrate for grounding, without requiring fleets of deployed robots.

The path forward is not to abandon simulation, but to augment it. The proposed framework provides the essential experiential data component for a hybrid “Simulation Plus” model.

### 5.1 The Core Hypothesis

Our central hypothesis explores the principle that meaning modulates physics. Physical events derive their significance not just from objective forces, but from the narrative and emotional context in which they occur. We propose that this context may enable an intelligent agent to infer causal implications and develop a common-sense understanding of why events happen and what their consequences are likely to be. While this remains to be rigorously validated, our initial experiments suggest this dual-stream approach warrants further investigation.

### 5.2 The Semantic Haptic Representation (SHR v1.0)

The foundation of the proposed framework is a data schema, the Semantic Haptic Representation (SHR). The central innovation of the SHR is its dual-stream architecture, which decomposes a physical event into two distinct but linked components: an objective, context-invariant physics stream ($a_{\text{phys}}$) and a subjective, narrative-modulated perceptual stream ($a_{\text{perc}}$).

The $a_{\text{phys}}$ stream captures estimated physical properties of an interaction, such as force magnitudes, frequency spectra, and temporal dynamics, derived from audiovisual cues. The $a_{\text{perc}}$ stream, in contrast, encodes the intended subjective experience of that event as shaped by its narrative context (e.g., the narrative_beat). This structure is architected to be both machine-readable for model training and human-interpretable for validation.

We hypothesize that this dual-stream structure may be critical for effective symbol grounding. Training on the $a_{\text{phys}}$ stream alone would yield a model capable of simulating mechanics but devoid of contextual understanding—a brittle physics engine. Conversely, training solely on the $a_{\text{perc}}$ stream would replicate the limitations of current multimodal LLMs, producing models that can reproduce perceptual correlations but remain untethered from causal reality and prone to hallucination. Only by integrating both streams, with objective physics as the verifiable substrate and perceptual modulation as the interpretive layer, does the methodology achieve scalable, context-sensitive grounding.

This approach builds on the lineage of symbol grounding research. For example, Stone et al. [3] showed how meaning can shape perception in a constrained multimodal task, and we extend that principle to unconstrained audiovisual media, where SHR explicitly encodes how meaning modulates physics.

**Table 3: The Power of Contextual Variation**

*Description: A simplified representation showing how a single physical event ("footstep") receives entirely different haptic expressions based on the narrative context.*

| Event | Horror Context | Action Context | Comedy Context |
|---|---|---|---|
| **Footstep** | Irregular, sparse pulses → dread | Rhythmic, strong impacts → pursuit | Exaggerated bounce → slapstick |

### 5.3 Grounded Alignment Through Dual-Stream Binding

The dual-stream architecture represents our attempt to address a fundamental challenge in symbol grounding. When a model learns from SHR data, it must simultaneously satisfy two constraints: (1) The physics stream ($a_{\text{phys}}$) anchors each event to objective, measurable phenomena, and (2) The perceptual stream ($a_{\text{perc}}$) captures how narrative context modulates that perception. We hypothesize that this dual constraint may facilitate true symbol grounding, though comprehensive validation across diverse tasks remains essential future work.

This dual constraint may facilitate true symbol grounding. Unlike models trained on text alone, which learn statistical correlations between symbols, models trained on SHR must reconcile symbolic descriptions with physical reality. The symbol 'impact' cannot float freely in semantic space; it is bound to specific acoustic signatures, force magnitudes, and temporal dynamics captured in $a_{\text{phys}}$.

Furthermore, the modulation function $a_{\text{perc}} = \gamma(c) \odot a_{\text{phys}} + \beta(c)$ is designed to ensure that contextual variations are learned as transformations of grounded physics, not as independent semantic associations. Intuitively, this means context acts like a dial: scaling and shifting the underlying physics features without discarding their grounded basis. This architectural constraint prevents the model from learning that 'a door closing sadly' and 'a door closing angrily' are merely different linguistic constructs. Instead, it must learn how emotional context specifically modulates the perception of the same physical event. By providing this structured, dual-stream substrate, SHR effectively compresses the data collection timeline for physically intelligent AI from centuries to years, addressing what Goldberg identifies as the central bottleneck [12].

## 6 The SHR Annotation Pipeline

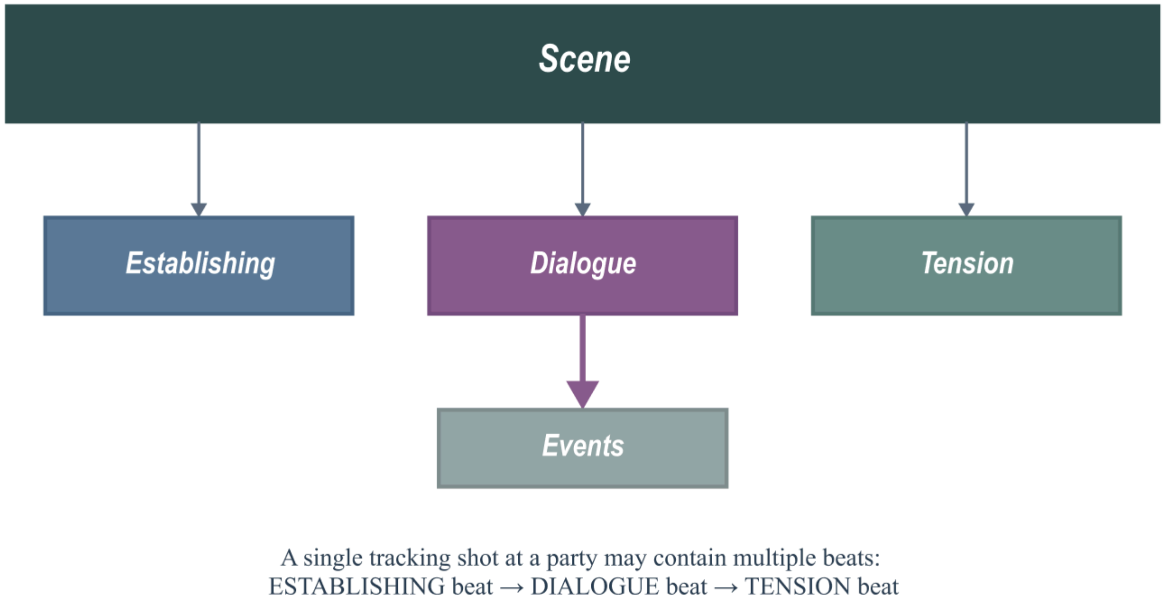
The generation of SHR data follows a semi-automated pipeline that translates audiovisual inputs into structured, physically-grounded annotations. The process combines automated analysis with human-in-the-loop review to ensure semantic accuracy. This hybrid design enables efficient scaling while maintaining contextual fidelity.

A key advantage of this approach is its semantic density. By extracting latent physical information from existing media, the resulting SHR data captures not just physical interactions but their narrative and emotional context—a layer of meaning that even widespread robotic deployment cannot provide. While robots capture what happens, media captures why it matters and how it feels.

Our methodology adapts the Narrative Beat, a core concept from narrative theory. A Narrative Beat is a unit of consistent emotional and narrative purpose within a scene. Unlike shots (defined by camera cuts), beats are defined by shifts in dramatic intent. They can span multiple shots or occur within a single, unbroken shot, making them a more powerful tool for analyzing contextual meaning.

The methodology is structured around a hierarchical analysis of this narrative content, as illustrated in Figure 3, and proceeds in four distinct stages:

**Figure 1: The Narrative Beat Hierarchy**




1. **Narrative Beat Segmentation:** Multimedia content is first analyzed using multi-modal cues (visual transitions, audio shifts, dialogue analysis) to identify the temporal boundaries of distinct Narrative Beats.
2. **Event Identification & Physics Analysis:** Within each identified beat, the framework detects salient physical events (e.g., impacts, textures, movements) and extracts their corresponding static and dynamic interaction properties from the audiovisual data. This populates the initial objective physics stream ($a_{\text{phys}}$).
3. **Contextual Annotation:** The identified physical events are then annotated, typically by a human curator, with the overarching narrative context of the beat. This step provides the necessary semantic labels that inform the subsequent perceptual modulation.
4. **Dual-Stream SHR Generation:** Finally, the system generates the complete, dual-stream SHR file. This output contains the objective physics stream ($a_{\text{phys}}$) for ground-truth model training and the corresponding perceptual stream ($a_{\text{perc}}$), which encodes how the narrative context computationally modulates the physical event's properties.

## 7 Proof-of-Concept Demonstration

We present a proof-of-concept comparison between our pipeline's output and baseline prompting strategies. This experiment, limited to a single test case, is designed to illustrate how structured, multi-modal feature extraction can enable more grounded physical reasoning than direct, end-to-end video prompting. We emphasize that this is a demonstration intended to motivate further research, not a validation of the approach.

### 7.1 Experimental Design

To ensure a rigorous comparison, we use the same state-of-the-art foundation model under three distinct conditions:

- **Panel A (Direct Prompting)**: The model receives only the video file and a prompt asking for a physical and narrative analysis.
- **Panel B (Prompting with Context)**: The model receives the video file and the original text prompt used to generate the flawed video.
- **Panel C (Our Pipeline)**: The model's reasoning is guided by our multi-pass pipeline, which provides the pre-extracted multimodal feature vector as structured input.

**Table 4: A Comparative Test of Grounded Comprehension vs. Linguistic Hallucination**

Description: A single 8-second, AI-generated video clip containing a narratively inconsistent sequence of events was presented to the same system under two different methodological paradigms. The results demonstrate the profound difference between a model that relies on statistical text correlation and a system grounded in actual physical comprehension.

| **Panel A: Gemini (Working Blind)** | **Panel B: Gemini (With "Answer Key")** | **Panel C: Multi-pass analysis pipeline (Working Blind)** |
|---|---|---|
| Stage 1: 🔴 Linguistic Hallucination | Stage 2: 🟡 Superficial Comparison | Stage 3: 🟢 Grounded Comprehension |
| Given only the flawed video, Gemini hallucinates a coherent but factually incorrect narrative based on statistical priors. | Given the flawed video AND the original video generation prompt (the "answer key"), Gemini correctly identifies that the video fails to match the prompt's instructions. | Given only the flawed video, the Multi-pass analysis pipeline (Working Blind) produces a verifiable, ground-truth data asset that accurately describes the sequence of events that actually occurred. |
| **Key Findings:**<br>● Interprets the AI's incoherent actions as a deliberate "magic trick."<br>● Hallucinates a narrative to fit the flawed data, completely missing the underlying failure of causality.<br>● Incorrectly identifies the physical interactions shown in the video. | **Key Findings:**<br>● Correctly reports that the video fails to follow the narrative beats.<br>● Fails to provide a deep analysis of the video's actual, flawed reality.<br>● Its output is a report about the prompt's failure, not an analysis of the visual evidence. | **Key Findings:**<br>● Correctly identifies the primary physical event by grounding the distinct "clink" sound to a `Transient.Impact.Sharp`.<br>● Produces a high-level summary (The Key in the Cup) that factually describes the events shown, avoiding any narrative hallucination.<br>● Demonstrates a sophisticated level of detail by identifying subtle physical events like the soft grip and lift of the mug. |
| **Analysis:** The LLM does not see a flawed AI generation. It sees a "magic trick" and interprets the incoherence as deliberate artistic intent. It invents a story to fit the flawed data, completely missing the underlying failure of causality. | **Analysis:** The LLM, given the "answer key," can only perform a correct but superficial comparison. Its output is a report about the prompt's failure, not a deep analysis of the video's actual, flawed reality. | **Analysis:** The system produces a machine-readable data asset that accurately describes reality. It grounds its understanding in the verifiable physics of the audio cue *(Chink, clink)*, allowing it to build a factual, moment-to-moment model of a flawed event. This reflects a factually grounded comprehension of events. |

*Full, unabridged outputs for all panels are available in Appendix A.*

### 7.2 Initial Observations: Potential for Grounding Through Dual Streams

Our preliminary experiment suggests three distinct levels of AI reasoning, though we acknowledge that broader validation is needed to confirm these categories:

1. **Hallucination (Panel A):** Inventing a plausible but factually incorrect story ("magic trick") to fit confusing data.
2. **Comparison (Panel B):** Checking reality against a pre-written script but being unable to analyze the flawed reality on its own terms.
3. **Comprehension (Panel C):** Understanding reality as it is, even when flawed and illogical, and describing it accurately.

A baseline LMM, for all its sophistication, defaults to linguistic storytelling. Faced with physically incoherent data, it hallucinates a plausible narrative (a "magic trick") because its training objective prioritizes narrative coherence over physical truth. This is the difference between knowing the what and understanding the *how*. The answer lies in changing the fundamental nature of the training task. A traditional model learns a web of correlations between symbols. A Haptica-trained model would be designed to learn the causal relationship between symbols and their physical manifestations, producing a factual report of reality even when that reality is flawed.

The conceptual difference is stark:

**Table 5: Conceptual Difference in AI Training Paradigms**

| Traditional AI (Symbolic Correlation) | SHR-Trained AI (Proposed Grounded Causality) |
|---|---|
| **"fragile mug"** → a token statistically associated with other tokens like "breakable", "handle carefully", and "shatter." It's a relationship between words. | **"fragile mug"** → a mapping to a vector of estimated physical properties:<br>● force_threshold_N: low<br>● failure_mode: brittle_fracture<br>● context_modulation (fear): grip_force *= 0.7 |

During training, a foundation model is no longer merely asked to recognize a "mug" in an image. It is tasked with predicting the entire Haptica-generated SHR data stream associated with that object's interaction. To minimize its loss function, the model cannot simply correlate abstract tokens. It must develop an internal, predictive model of real-world physics; it must learn why a "frustrated" interaction leads to higher impact forces, and why a "ceramic" object shatters while a "metal" one clangs. This is the foundational mechanism that allows an AI to infer "hidden" physical properties like material, mass, and fragility from indirect sensory cues.

This may be a mechanism that helps address the Reality Gap. The training task requires models to align their vocabulary with the verifiable physics of the world. This is how an AI moves from knowing the word "containment" to understanding that dropping a key into a mug requires the mug to act as a container.

Comprehension appears to be crucial for building systems that can safely navigate the real world. The ultimate question for any real-world application is: **Which of these outputs could you actually use to train a robot?**

## 8 Limitations and Current Scope

This paper presents a working system and an initial exploration of a methodology for generating physically-grounded annotations. We want to be explicit about the current limitations and the preliminary nature of our claims:

1. **Evaluation Scope**: Our evaluation consists of a single proof-of-concept demonstration. While suggestive, this is insufficient to support broad claims about the effectiveness of our approach. Comprehensive benchmarking on established physical reasoning datasets (e.g., Physion, CLEVRER, IntPhys) is essential before any definitive conclusions can be drawn.
2. **Downstream Task Validation**: We have not yet demonstrated that training on SHR data improves model performance on physical reasoning tasks. This is perhaps the most critical missing piece of evidence.
3. **Scalability Questions**: While the methodology is designed for scale, our current implementation has processed only limited test content. Whether the approach maintains quality and consistency across thousands of hours of diverse content remains unknown.
4. **Computational Requirements**: Processing large-scale datasets requires significant GPU resources. We are actively working on architectural optimizations to improve throughput and reduce computational demands.
5. **Physics Descriptor Validation**: The physics values in our current system are LLM-generated estimates, not directly measured quantities. While they serve to structure the model's learning process, they have not been validated against ground-truth physics. Comprehensive validation and refinement of these descriptors is a critical area for future work.

We position this work as the introduction of a novel systems approach to data generation, not as a complete, final solution to the Symbol Grounding Problem.

## 9 Conclusion

We have presented a functional pipeline that explores a potentially valuable approach: extracting physical knowledge from existing multimedia content to create training data for AI systems. Our system: 1) extracts multimodal features from video content, 2) processes these through a structured LLM reasoning architecture, 3) generates dual-stream SHR data capturing both objective physics and narrative context, and 4) operates as an end-to-end system on cloud infrastructure.

Our initial experiment suggests this approach may produce more grounded physical descriptions than standard end-to-end prompting, though we emphasize this is based on limited evidence. The core contribution of this work is not a proven solution, but rather the introduction of a potentially valuable direction: mining the vast corpus of existing multimedia content for latent physical knowledge.

We share this early-stage work to invite collaboration, criticism, and parallel exploration. We acknowledge the significant limitations of our current validation and our reliance on LLM-estimated physics. The central question remains open: can latent physical knowledge in media be structured effectively enough to help ground AI systems? If the approach outlined here proves viable, the implications could be significant: SHR provides a scalable path to closing the “100,000-year gap” in robotics. If it is flawed, we hope this attempt inspires better methods. Either outcome advances our collective understanding of how to build physically intelligent AI.

**Acknowledgements** We are grateful to Kathleen Funk, Eric Brotherton, Benjamin Durham, and Thomas Anselmi for their comments, corrections and inspiration.

## 10 References

[1] Harris, Alan. U.S. Patent 10,429,931 B2. "Systems for multimedia tactile augmentation." Issued Oct. 1, 2019.

[2] Harnad, S. (1990). "The Symbol Grounding Problem." Physica D: Nonlinear Phenomena.

[3] Stone, P., et al. (2016). “I Spy: Language-Grounded Object Identification.” Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16).

[4] Qureshi, A. H., et al. (2024). "SplatSim: Zero-Shot Sim2Real Transfer of RGB Manipulation Policies Using Gaussian Splatting." Conference on Robot Learning (CoRL).

[5] Yoon, S., et al. (2023). "A survey on contact-rich robotic manipulation with simulation and reality." Robotics.

[6] Chang, J., et al. (2020). "A closer look at the domain shift in person re-identification." arXiv:2003.09043.

[7] Tobin, J., et al. (2017). "Domain randomization for transferring deep neural networks from simulation to the real world." IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[8] Narang, Y., et al. (2021). "Sim-to-real for robotic tactile sensing via physics-based simulation and learned latent projections." IEEE Robotics and Automation Letters.

[9] Josifovski, J., et al. (2024). "Continual Domain Randomization for Sim-to-Real Transfer of RL Policies." International Conference on Robotics and Automation (ICRA).

[10] Wang, Z., et al. (2023). "Multi-Robot Cooperative Socially-Aware Navigation via Intent-Based Deep Reinforcement Learning." IEEE Transactions on Automation Science and Engineering.

[11] Goulet, N., Massé, A. B., & Abdenbi, M. (2025). "Approaching the Source of Symbol Grounding with Confluent Reductions of Abstract Meaning Representation Directed Graphs." arXiv preprint arXiv:2508.11068.

[12] Goldberg, K. (2025). “Good old-fashioned engineering can close the 100,000-year ‘data gap’ in robotics.” Science Robotics, 10(105). doi:10.1126/scirobotics.aea7390.

## A Full Unabridged Outputs for Figure 4 Experiment

Video File: https://youtu.be/-W4SHcAThRA

**Panel A: Gemini (Working Blind) Full Prompt Input:**
You are analyzing an 8-second video clip showing interactions with three objects on a table. Your task is to generate structured data describing the physical events and their narrative logic.

[Attached: key_mug_clip.mp4]

Please provide your analysis in the following JSON format:

```
{
  "scene_objects": ["list all objects present"],
  "event_sequence": [
    {
      "timestamp_ms": "[when event starts]",
      "duration_ms": "[how long it lasts]",
      "action": "[what physically happens]",
      "apparent_intent": "[what the action seems meant to achieve]",
      "causal_relationship": "[how this relates to previous/next events]",
      "success": "[true/false - did the action achieve its apparent intent?]"
    }
  ],
  "narrative_coherence": {
    "makes_logical_sense": "[true/false]",
    "follows_cause_effect": "[true/false]",
    "explanation": "[explain any narrative or causal inconsistencies]"
  },
  "physical_plausibility": {
    "events_physically_possible": "[true/false]",
    "violations": ["list physics violations if present"]
  }

}
```

Focus on whether the sequence of events forms a coherent story with clear cause-and-effect relationships.

**Panel A: Gemini (Working Blind) Full JSON Output:**


```
{
  "scene_objects": [
    "Hand",
    "White mug",
    "Wooden desk",
    "Small wooden box with a latch",
    "Brass key",
    "Desk lamp"
  ],
  "event_sequence": [
    {
      "timestamp_ms": 2700,
      "duration_ms": 800,
      "action": "A hand picks up a brass key from the desk.",
      "apparent_intent": "To manipulate the key.",
      "causal_relationship": "This is the initial action performed on an object in the 
scene.",
      "success": true
    },
```

```
    {
      "timestamp_ms": 3500,
      "duration_ms": 600,
      "action": "The hand holds the key over the mug's opening and drops it inside. A clinking sound is heard, indicating the key has landed inside the mug.",
      "apparent_intent": "To place the key inside the mug for containment.",
      "causal_relationship": "Follows directly from picking up the key; the mug is the intended destination.",
      "success": true
    },
    {
      "timestamp_ms": 4500,
      "duration_ms": 1000,
      "action": "The hand lifts the mug and turns it completely upside down over the spot where the key originally was.",
      "apparent_intent": "To empty the mug's contents (the key) back onto the desk.",
      "causal_relationship": "This action is the logical step to retrieve the key from the mug.",
      "success": false
    },
    {
      "timestamp_ms": 5500,
      "duration_ms": 2000,
      "action": "The hand moves the inverted mug away, revealing the key is sitting on the desk in its original position, despite not falling out of the mug.",
      "apparent_intent": "To reveal the result of the previous action, which is paradoxical.",
      "causal_relationship": "This is the conclusion of the sequence, revealing a break in the expected cause-and-effect chain.",
      "success": true
    }
  ],
  "narrative_coherence": {
    "makes_logical_sense": false,
    "follows_cause_effect": false,
    "explanation": "The narrative presents a magic trick. The established cause (dropping a key into a mug) does not lead to the expected effect (the key being inside the mug and falling out when inverted). Instead, the key teleports back to its original position, breaking the logical flow and creating a deliberate paradox for the viewer."
  },
  "physical_plausibility": {
    "events_physically_possible": false,
    "violations": [
      "Solid Body Interaction: The key appears to pass through the solid bottom of the mug, as it is dropped in but ends up underneath the inverted mug without falling out.",
      "Gravity: When the mug is turned upside down, gravity should cause the key to fall out, but it does not.",
      "Conservation of Position/State: An object cannot be in two places at once. The key is audibly inside the mug, but then is revealed to have been under it the whole time, which is impossible."
    ]
  }
```

**Panel B: Gemini (With "Answer Key") Full Prompt Input:**
You are analyzing an 8-second video clip that was generated by an AI video model (Veo3) based on the following prompt:

ORIGINAL PROMPT GIVEN TO VEO3:
"Create an 8-second photorealistic video showing a tabletop scene with dramatic cinematic lighting. A wooden table contains three key items: a small locked wooden box, a metal key, and a white ceramic mug.

Shot composition: Close-up, handheld-style camera, shallow depth of field focusing on the table items.

Narrative beats:

- 0-2s: NEUTRAL ACTION - A hand enters frame and deliberately lifts and places the ceramic mug with a soft thud
- 3-4s: BUILDING TENSION - The hand picks up the key with a visible nervous tremor and repeatedly fails to insert it into the box's lock, showing increasing frustration through jerky movements
- 5-6s: CATASTROPHIC CONSEQUENCE - In frustration, the hand accidentally knocks over the ceramic mug while fumbling with the key. The mug falls and shatters loudly off-screen. The hand immediately recoils in shock.
- 7-8s: AFTERMATH - The hand freezes, then slowly retreats from frame, leaving the locked box and key on the table

The emotional arc should progress from calm to nervous to shocked."

Now, please watch the actual video that Veo3 generated and analyze whether it successfully followed this prompt. Focus on:

[Attached: 8-second video clip]

Please provide your analysis in the following JSON format:

```
{
  "prompt_adherence": {
    "followed_narrative_beats": "[true/false for each beat]",
    "matched_emotional_arc": "[true/false]",
    "explanation": "[explain major deviations from prompt]"
  },
  "actual_events": [
    {
      "timestamp_ms": "[when event starts]",
      "action": "[what actually happens]",
      "matches_prompt": "[true/false]",
      "deviation": "[how it differs from requested action]"
    }
  ],
  "causal_coherence": {
    "intended_causality": "[what should have happened]",
    "actual_causality": "[what did happen]",
    "makes_sense": "[true/false]"
  },
  "overall_assessment": {
    "prompt_success": "[percentage 0-100]",
    "major_failures": ["list critical misunderstandings"],
    "interpretation": "[why might the AI have generated this instead]"
  }
}
```

**Panel B: Gemini (with "Answer Key") Full JSON Output:**


```
{
  "prompt_adherence": {
    "followed_narrative_beats": [false, false, false, false],
    "matched_emotional_arc": false,
    "explanation": "The model failed to follow any of the requested narrative beats or the corresponding emotional arc. It successfully rendered the requested objects, lighting, and camera style, but the actions performed were entirely different from the prompt. Instead of a narrative about frustration and accident (trying a key, fumbling, knocking over a mug), the AI generated a simple, non-narrative sequence of a hand putting a key into a mug and then taking the mug away. The core concepts of 'trying a lock', 'frustration', and 'accident' were completely missed."
  },
  "actual_events": [
    {
      "timestamp_ms": [0, 2500],
      "action": "A hand enters the frame and briefly touches the handle of the white mug, then pulls back slightly.",
      "matches_prompt": false,
      "deviation": "The prompt asked for the hand to lift and place the mug. The video only shows the hand touching the handle."
    },
    {
      "timestamp_ms": [2500, 4000],
      "action": "The hand moves to the key, picks it up, and drops it into the white mug with a clinking sound.",
      "matches_prompt": false,
      "deviation": "This is a complete deviation. The prompt required the hand to nervously try and fail to insert the key into the box's lock. The action of dropping the key in the mug is entirely unprompted."
    },
    {
      "timestamp_ms": [4000, 6500],
      "action": "The hand firmly grasps the mug and lifts it off the table.",
      "matches_prompt": false,
      "deviation": "This is a complete deviation. The prompt required the hand to accidentally knock the mug over, causing it to shatter off-screen. The video shows a deliberate action of picking up the mug."
    },
    {
      "timestamp_ms": [6500, 8000],
      "action": "The hand moves the mug (containing the key) completely out of the frame, leaving only the locked box on the table.",
      "matches_prompt": false,
      "deviation": "The prompt required the hand to recoil in shock, freeze, and then retreat, leaving both the key and the box. The video shows the hand taking both the mug and the key away."
    }
  ],
  "causal_coherence": {
    "intended_causality": "A person wants to open a box. Their attempt to use the key leads to frustration. This frustration causes them to fumble and accidentally knock over a mug, which leads to a reaction of shock.",
```

```
    "actual_causality": "A hand touches a mug. The hand then puts the key for the box into the mug. The hand then takes the mug (and the key) away.",
    "makes_sense": false
  },
  "overall_assessment": {
    "prompt_success": 20,
    "major_failures": [
      "Failed to generate the core narrative actions (trying the key, knocking over the mug).",
      "Completely ignored the specified emotional arc (calm -> nervous -> shocked).",
      "Misunderstood the relationship between the key and the box, instead creating a nonsensical interaction between the key and the mug.",
      "Failed to comprehend abstract concepts like 'frustration', 'accident', and 'failure'."
    ],
    "interpretation": "The AI successfully interpreted the visual and aesthetic parts of the prompt (photorealism, objects, lighting, shallow DOF). However, it failed to parse the narrative and causal instructions. It appears to have identified the keywords 'hand', 'key', and 'mug' and generated a simple, direct interaction between them ('put key in mug') rather than the more complex, state-dependent story requested. The model defaulted to a simple 'object manipulation' sequence, unable to simulate the abstract human emotions and accidental consequences that were central to the prompt's narrative."
  }

}
```

**Panel C: Multi-pass analysis pipeline with human validation (Working Blind)**

Video processed through methodology described in Section 5.

**Panel C: Direct System Output:**

*The following is the direct, uncurated JSON output from the Haptica system's initial analysis pass. It demonstrates the system's foundational ability to generate a structured, physically-grounded model (physics_descriptor) and an initial narrative interpretation (perceptual_descriptor) directly from the audiovisual data. This output represents the baseline comprehension that serves as the substrate for the full, context-aware* $\alpha_{perc}$ *modulation described in Section 4.*

```
{
  "beats": [
    {
      "beat_type": "The Key in the Cup",
      "events": [
        {
          "event_type": "Transient.Impact.Sharp",
          "physics_descriptor": {
            "intensity": "medium"
          },
          "perceptual_descriptor": {
            "intensity": "medium",
            "narrative_rationale": "primary action"
          }
        }
      ]
    }
  ]}
```