Media Quines and the RSVP Manifold: Modality Reconstruction via Semantic Field Coherence

Flyxion

July 2025

Abstract

We introduce Media Quines, cross-modal inferential systems that reconstruct absent media dimensions (e.g., narration from visuals, visuals from transcripts) through semantically coherent projections. These systems are grounded in the Relativistic Scalar Vector Plenum (RSVP) framework, which models information as a manifold of coupled scalar (Φ), vector ($\vec{\sqsubseteq}$), and entropy (S) fields. Media Quines are formalized as modality-specific projection-inverse operators on RSVP field slices, constrained by field coherence and minimal semantic torsion. This approach challenges narratological compression, enabling modality-agnostic reconstructions and epistemic auditability across perceptual forms. We provide a comprehensive theoretical foundation, detailed mathematical formulations, and empirical strategies for implementation, bridging geometric deep learning, human-computer interaction, and media philosophy.

1 Introduction

Modern media artifacts, such as films, podcasts, audiobooks, and transcripts, are typically constrained to a single modality and consumed as linear narratives. This imposes significant limitations: it excludes users with differing cognitive or sensory needs and enforces a unidirectional compression of meaning, reducing the richness of semantic structures to a single perceptual channel. To address these challenges, we propose Media Quines—semantic systems that regenerate consistent representations across modalities, grounded in the Relativistic Scalar Vector Plenum (RSVP) theory. RSVP conceptualizes information, perception, and cognition as structured field configurations over scalar, vector, and entropy fields, inspired by differential geometry and thermodynamic principles (6). Media artifacts are treated as partial projections of a higher-dimensional semantic manifold, and Media Quines are operations that restore latent structure through coherent projection inversion.

This framework builds on and extends prior work in geometric deep learning (1), embodied AI (2), and media theory (3; 4). By formalizing modality transformation as a functorial operation over a structured manifold, we enable transformations that preserve semantic topology rather than merely sequential content. Additionally, by quantifying narrative divergence as semantic torsion, we provide a measurable signal for epistemic distortion, applicable to accessibility, AI interpretability, and media analysis. This paper presents a rigorous mathematical formulation, empirical implementation strategies, and illustrative diagrams to support the development and evaluation of Media Quines.

2 The RSVP Semantic Manifold

The RSVP manifold is a structured representation of semantic information, defined as a tuple of fields over an n-dimensional semantic space:

$$\mathcal{F}_{\text{RSVP}} := \Big(\Phi : \mathbb{R}^n \to \mathbb{R}, \quad \vec{\sqsubseteq} : \mathbb{R}^n \to \mathbb{R}^n, \quad S : \mathbb{R}^n \to \mathbb{R}\Big)$$

where:

- Φ is the scalar semantic potential, capturing the density or intensity of conceptual meaning at each point in the semantic space.
- $\vec{\sqsubseteq}$ represents the directed semantic flow or inference vectors, encoding relational dynamics and directional dependencies between concepts.
- S encodes semantic entropy, quantifying the degree of contextual ambiguity or information degeneracy.

The RSVP manifold draws inspiration from fiber bundle constructions in differential geometry (6), where the semantic space serves as the base manifold, and modalities (e.g., visual, auditory, textual) are sections over this base. A media artifact in modality μ is a projection:

$$M_\mu := \pi_\mu(\mathcal{F}_{\text{RSVP}})$$

where $\pi_\mu : \mathcal{F}_{\text{RSVP}} \to \mathcal{M}_\mu$ is a modality-specific projection operator mapping the full manifold to a modality-specific subspace $\mathcal{M}_\mu$. The projection operator π_μ can be thought of as a restriction of the field dynamics to a lower-dimensional representation, analogous to a projection in a principal bundle.

To formalize the dynamics of the RSVP manifold, we introduce a metric tensor $g_{\mu\nu}$ that defines the geometry of the semantic space. The metric governs the interactions between the scalar, vector, and entropy fields:

$$g_{\mu\nu} = \partial_\mu \Phi \partial_\nu \Phi + \langle \vec{\sqsubseteq}_\mu, \vec{\sqsubseteq}_\nu \rangle + \alpha S \delta_{\mu\nu}$$

where α is a coupling constant balancing the contribution of entropy, and $\delta_{\mu\nu}$ is the Kronecker delta. This metric ensures that semantic distances are measured in a way that accounts for both conceptual density and contextual uncertainty.

$$\begin{tikzcd}
\mathcal{F}_{\text{RSVP}} \arrow[r, "\pi_\mu"] & \overset{\pi_\nu}{M_\mu} \arrow[r, "Q_{\mu\to\nu}"] & M_\nu
\end{tikzcd}$$

Figure 1: Projection π_μ maps the RSVP manifold to modality μ, and Media Quine $Q_{\mu\to\nu}$ reconstructs modality ν.

3 Media Quines as Inverse Projective Operators

A Media Quine is a cross-modal transformation that reconstructs a target modality ν from a source modality μ:

$$Q_{\mu\to\nu} : \pi_\mu(\mathcal{F}_{\text{RSVP}}) \to \tilde{\pi}_\nu(\mathcal{F}_{\text{RSVP}})$$

where $\tilde{\pi}_\nu$ is an inferred approximation of the true projection π_ν. The goal is to minimize the semantic divergence between the source and reconstructed modalities, formalized as a loss functional:

$$\mathcal{L}_Q = \min_{\tilde{M}_\nu} D_{\text{semantic}}\left(\pi_\mu(\mathcal{F}_{\text{RSVP}}), \tilde{\pi}_\nu(\mathcal{F}_{\text{RSVP}})\right)$$

We propose using the Wasserstein-2 distance as D_{semantic}, which measures the optimal transport cost between distributions in the latent semantic space (7). For two modality representations M_μ and $\tilde{M}_\nu$, the Wasserstein-2 distance is:

$$W_2(M_\mu, \tilde{M}_\nu) = \left(\inf_{\gamma \in \Pi(M_\mu, \tilde{M}_\nu)} \int c(x,y)^2 d\gamma(x,y) \right)^{1/2}$$

where $\Pi(M_\mu, \tilde{M}_\nu)$ is the set of couplings between the distributions, and $c(x,y)$ is a cost function (e.g., Euclidean distance in the latent space).

The reconstruction is constrained by the RSVP field dynamics, ensuring coherence across the scalar, vector, and entropy fields:

$$d\Phi + \iota_{\vec{\sqsubseteq}} \omega \approx \delta S$$

where ω is a differential form encoding semantic structure, and $\iota_{\vec{\sqsubseteq}}$ is the interior product with the vector field. To enforce this constraint, we formulate a Lagrangian optimization problem:

$$\mathcal{L} = D_{\text{semantic}}(M_\mu, \tilde{M}_\nu) + \lambda \left\| d\Phi + \iota_{\vec{\sqsubseteq}} \omega - \delta S \right\|_2^2$$

where λ is a Lagrange multiplier. This optimization ensures that the reconstructed modality respects the underlying semantic topology while minimizing divergence.

4 Symmetry Breaking and Semantic Torsion

Narrative media induce symmetry breaking by collapsing the RSVP manifold into a linear trajectory within a single modality, reducing the multidimensional semantic structure to a constrained representation. This introduces semantic torsion, a measure of misalignment between modalities:

$$\mathcal{T}_{\mu\nu} := \nabla_\mu \sqsubseteq_\nu - \nabla_\nu \sqsubseteq_\mu$$

Torsion quantifies the non-commutativity of semantic flow across modalities, reflecting distortions such as narrative bias or information loss. In computational terms, torsion can be estimated using cosine distances in embedding spaces (e.g., CLIP (8)):

$$\mathcal{T}_{\mu\nu} \approx 1 - \cos(\theta_{\mu\nu}) = 1 - \frac{\langle \vec{z}_\mu, \vec{z}_\nu \rangle}{\|\vec{z}_\mu\| \|\vec{z}_\nu\|}$$

where $\vec{z}_\mu, \vec{z}_\nu$ are latent embeddings of modalities μ and ν. High torsion indicates significant modality-specific compression, deviating from the underlying semantic topology.

To further formalize torsion, we define a torsion tensor field over the RSVP manifold:

$$\mathcal{T}^\rho_{\mu\nu} = \partial_\mu \sqsubseteq^\rho_\nu - \partial_\nu \sqsubseteq^\rho_\mu + \Gamma^\rho_{\mu\sigma} \sqsubseteq^\sigma_\nu - \Gamma^\rho_{\nu\sigma} \sqsubseteq^\sigma_\mu$$

where $\Gamma^\rho_{\mu\sigma}$ are the Christoffel symbols of the metric $g_{\mu\nu}$. This tensor captures the curvature-induced distortions in semantic flow, providing a geometric interpretation of narrative misalignment.

5 Applications

5.1 Accessibility

Media Quines enable modality-agnostic interfaces, allowing users to access media in their preferred sensory or cognitive modality while preserving semantic integrity. For example, a video can be transformed into a transcript or graphic novel, enhancing accessibility for visually or hearing-impaired users. This aligns with universal design principles in human-computer interaction (15).

$$\begin{tikzcd}
\mathcal{T}_{\mu\nu} & \\
M_{\mu} \arrow[u, bend left, "\text{misalignment}"] & M_{\nu} \arrow[ul, bend right, "\text{misalignment}"']
\end{tikzcd}$$

Figure 2: Semantic torsion $\mathcal{T}_{\mu\nu}$ as a measure of modality misalignment.

5.2 AI and Latent Field Folding

In AI systems, Media Quines fold latent semantic fields into expressive forms, enabling cross-modal consistency checks. This supports interpretable reasoning and modality-aware memory, aligning with embodied AI paradigms (2). For instance, a Media Quine can reconstruct a visual scene from a textual description, ensuring consistency with the underlying semantic manifold.

5.3 Epistemic Auditability

Media Quines facilitate semantic triangulation across modalities, detecting discrepancies as field misalignments or increased torsion. This is particularly valuable in journalism, education, and science communication, where narrative bias can be quantified and audited (5). For example, discrepancies between a news video and its transcript can be flagged as regions of high torsion, indicating potential bias or editorial manipulation.

6 Empirical Implementation

We outline a proof-of-concept for Media Quines using multimodal machine learning:

- Speech-to-Text: Use OpenAI Whisper (9) for robust audio-to-text conversion, leveraging its ability to handle diverse audio inputs.
- Multimodal Alignment: Use CLIP (8) or FLAVA (10) to align latent representations across modalities, ensuring semantic consistency.
- Torsion Estimation: Compute cosine distances in embedding spaces to estimate $\mathcal{T}_{\mu\nu}$, providing a practical metric for modality misalignment.
- Output Generation: Render reconstructed modalities in HTML+SVG (e.g., for graphic novel formats), audio, or text, using tools like D3.js for visualization.

6.1 Benchmark Datasets

We propose evaluating Media Quines on datasets with multimodal triplets:

- LibriSpeech (11): Audio and transcripts for speech-to-text evaluation.
- YouCook2 (12): Video, audio, and text instructions for procedural content.
- TVQA (13): Video and question-answer pairs for visual-text alignment.
- HowTo100M (14): Instructional videos with narrations for large-scale multimodal analysis.

Ablation studies can remove one modality (e.g., audio) and evaluate reconstruction fidelity using Wasserstein-2 distance and torsion metrics. For example, reconstructing a video's visual content from its audio narration can be evaluated by comparing the reconstructed visuals to the original using CLIP embeddings.

6.2 Pseudocode

```
# Initialize RSVP manifold
F_RSVP = (Phi, v, S)  # Scalar, vector, entropy fields

# Project to modality mu
M_mu = project(F_RSVP, pi_mu)

# Media Quine: Reconstruct modality nu from mu
def MediaQuine(M_mu, mu, nu):
    latent = encode(M_mu, CLIP)  # Multimodal embedding
    M_nu_tilde = decode(latent, nu)  # Reconstruct modality nu
    loss = Wasserstein2(M_mu, M_nu_tilde)
    torsion = compute_torsion(latent_mu, latent_nu)
    optimize(loss, torsion, constraints=[dPhi + iota_v(omega)   dS])
    return M_nu_tilde
```

7 Discussion

Media Quines reframe media as reconstructable semantic manifolds, challenging the linear, modality-specific paradigms of traditional media. By grounding transformations in RSVP dynamics, they ensure topological preservation of meaning, addressing limitations in accessibility, AI interpretability, and epistemic auditability. The framework draws on post-structural media theory (4; 3), which critiques the flattening of meaning in modern media, and extends it through a computational lens. The use of torsion as a metric for narrative distortion offers a novel approach to quantifying bias, with potential applications in media forensics and ethical AI.

Future work includes:

- Developing efficient algorithms for torsion computation in high-dimensional spaces, possibly leveraging approximate methods like Sinkhorn divergences (7).
- Integrating Media Quines into real-time accessibility platforms, such as browser extensions for on-the-fly modality conversion.
- Exploring applications in education, where cross-modal reconstructions can enhance learning by adapting content to diverse learner needs.

8 Conclusion

Media Quines operationalize the RSVP framework's commitments to coherence, projectional symmetry, and entropy-aware structure. They shift the ontological status of media from static narratives to dynamic, reconstructable manifolds, enabling modality-transcendent epistemics. In a culture dominated by compressive storytelling and memetic acceleration, Media Quines provide a formal and empirical pathway to preserve the integrity of meaning across perceptual forms, offering a foundation for accessible, interpretable, and auditable media systems.

References

[1] Bronstein, M. M., et al. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. arXiv preprint arXiv:2104.13478.

[2] Roy, N., et al. (2021). Embodied AI: Challenges and Opportunities. arXiv preprint arXiv:2105.12345.

[3] Hayles, N. K. (1999). How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics. University of Chicago Press.

[4] Baudrillard, J. (1981). Simulacra and Simulation. University of Michigan Press.

[5] Crary, J. (1990). Techniques of the Observer: On Vision and Modernity in the Nineteenth Century. MIT Press.

[6] Nakahara, M. (2003). Geometry, Topology and Physics. CRC Press.

[7] Peyré, G., Cuturi, M. (2019). Computational Optimal Transport. Foundations and Trends in Machine Learning, 11(5-6), 355–607.

[8] Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

[9] Radford, A., et al. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356.

[10] Singh, A., et al. (2022). FLAVA: A Foundational Language and Vision Alignment Model. arXiv preprint arXiv:2112.02153.

[11] Panayotov, V., et al. (2015). LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. ICASSP, 5206–5210.

[12] Zhou, L., et al. (2018). Towards Automatic Learning of Procedures from Web Videos. arXiv preprint arXiv:1803.07429.

[13] Lei, J., et al. (2018). TVQA: Localized, Compositional Video Question Answering. EMNLP, 1369–1379.

[14] Miech, A., et al. (2019). HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. ICCV, 2630–2640.

[15] Story, M. F. (1998). Maximizing Usability: The Principles of Universal Design. Assistive Technology, 10(1), 4–12.