CHONKOLOGY: A MATHEMATICAL THEORY OF AUDIOVISUAL NARRATIVES

1. Introduction

A **chonk** is a short-form video artifact: a still image brought to life through orchestrated visual transformations synchronized with audio. This document develops the mathematical foundations of chonks, proceeding from primitive spaces through increasingly refined definitions.

We begin with the raw ingredients (images, transforms, audio), introduce the space of arbitrary audiovisual pairings, and then characterize chonks as those pairings satisfying a synchronization condition. This approach mirrors the construction of, say, measurable functions from arbitrary functions—the larger space provides context for understanding what makes the smaller space special.

2. Primitive Spaces

2.1. The Image Space.

Definition 2.1.1. Let $\mathcal{I}$ denote the space of images:

$$\mathcal{I} = \{f : [0,1]^2 \to [0,1]^3\}.$$

Remark. The codomain $[0,1]^3$ represents the RGB color space. In practice, images are discrete (pixel grids), but the continuous formulation generalizes cleanly and admits differential structure.

Remark. The domain $[0,1]^2$ is a normalization choice. Readers should mentally substitute the correct rectangle if their image is not square; none of what follows will depend delicately on this rectangle.

2.2. The Similarity Transform Group.

Definition 2.2.1. Let T denote the group of 2D similarity transformations (uniform scaling composed with translation):

$$T = \{\tau(s,t) : \mathbb{R}^2 \to \mathbb{R}^2 \mid s \in \mathbb{R}^+,\, t \in \mathbb{R}^2\}$$

where

$$\tau(s,t)(x) = sx + t.$$

Proposition 2.2.2. *T forms a 3-dimensional Lie group under composition:*

$$\tau(s_1,t_1) \circ \tau(s_2,t_2) = \tau(s_1 s_2,\, s_1 t_2 + t_1).$$

Proof. Direct computation. □

Remark. We restrict to similarity transforms (uniform scaling) rather than the full affine group to preserve aspect ratio—a chonk zooms and pans but does not shear or stretch.

2.3. The Time Domain.

Definition 2.3.1. For duration $D > 0$, let $T_D = [0, D] \subset \mathbb{R}$ denote the time domain.

2.4. **The Audio Space.**

Definition 2.4.1. Let A_D denote the space of audio signals of duration D:

$$A_D = \{a : [0, D] \to \mathbb{R}\}.$$

Remark. Stereo audio lives in $A_D \times A_D$. We treat audio as given data rather than constructed.

2.5. **The Mood Manifold.**

Definition 2.5.1. Let M be a low-dimensional smooth manifold parameterizing narrative/emotional states:

$$M \approx \mathbb{R}^k \quad \text{for small } k.$$

Possible coordinates on M include:

- **Tension** $\in [0, 1]$: suspense vs. resolution
- **Energy** $\in [0, 1]$: calm vs. intense
- **Valence** $\in [-1, 1]$: dark vs. bright

Remark. The exact dimensionality and coordinates are aesthetic choices. The essential property is that M is **shared** between audio and visual modalities.

2.6. **The Narrative Bundle.**

Definition 2.6.1. The **narrative bundle** is the product manifold

$$N = M \times T.$$

Definition 2.6.2. The canonical projections are

$$\pi_M : N \to M \qquad \pi_T : N \to T.$$

3. The Pairing Space

3.1. **Raw Pairings.**

Definition 3.1.1. A **raw pairing** is a 3-tuple

$$P = (I, \tau, a)$$

where $I \in \mathcal{I}$ is a source image, $\tau : T_D \to T$ is a transform trajectory, and $a \in A_D$ is an audio signal.

Definition 3.1.2. The **pairing space** is

$$\mathcal{P} = \mathcal{I} \times C^0(T_D, T) \times A_D$$

where $C^0(T_D, T)$ denotes the space of continuous maps from the time interval T_D into the transform group T.

Remark. A raw pairing is any image with any trajectory and any audio. There is no requirement that the visual motion "makes sense" with the audio. The space $\mathcal{P}$ is the ambient space in which chonks will live as a distinguished subspace.

3.2. **Rendering.**

Definition 3.2.1. The **rendered frame** of a pairing P at time t is

$$\text{Frame}(P, t) = I \circ \tau(t)^{-1}.$$

Remark. This is standard rasterization: to zoom in, we scale coordinates down before sampling.

3.3. **Keyframe Representation.**

Definition 3.3.1. A **keyframe sequence** is a finite set

$$K = \{(t_0, \tau_0, \iota_0), (t_1, \tau_1, \iota_1), \ldots, (t_n, \tau_n, \iota_n)\}$$

where $t_i \in T_D$, $\tau_i \in T$, and $\iota_i \in \{\text{hold}, \text{linear}\}$ specifies interpolation type.

Definition 3.3.2. The **interpolation operator**

$$\text{Interp}(K, \cdot) : T_D \to T$$

reconstructs a continuous trajectory from keyframes via piecewise-constant (hold) or linear interpolation.

4. The Audio Mood Function

4.1. **Definition.**

Definition 4.1.1. An **audio mood function** for signal $a \in A_D$ is a map

$$m_a : T_D \to M$$

that extracts the emotional or narrative content of the audio at each moment.

Example. Possible constructions of m_a include:

- Loudness envelope $\to$ Energy coordinate
- Spectral centroid $\to$ Brightness or Valence
- Onset density $\to$ Energy derivative
- Harmonic tension (dissonance) $\to$ Tension coordinate

Remark. For a given audio file, m_a is considered fixed data, though its computation may be approximate or learned. The timestamps in `effectTimestamps` can be viewed as samples of m_a at moments of high saliency.

Remark. The mood function m_a is **model-dependent**, not intrinsic to the audio signal. Different mood extractors (manual annotation, learned models, rule-based heuristics) induce different coherence geometries. There is no canonical m_a—the "mood" is imposed by a choice of model, not discovered in the signal. This is not a flaw; it reflects that synchronization is an aesthetic judgment, not a physical measurement.

5. Chonks as Synchronized Pairings

5.1. **The Narrative Path.**

Definition 5.1.1. A **narrative path** is a continuous map

$$\gamma : T_D \to N = M \times T.$$

At each moment, $\gamma(t)$ specifies both a mood state and a visual transform.

Definition 5.1.2. The **visual trajectory** derived from γ is

$$\tau_\gamma = \pi_T \circ \gamma : T_D \to T.$$

5.2. **The Synchronization Condition.**

Definition 5.2.1. A narrative path γ is **synchronized** with audio a if

$$\pi_M(\gamma(t)) \approx m_a(t) \quad \text{for all } t \in T_D.$$

Remark. The symbol "$\approx$" allows for degrees of synchronization.

5.3. **The Chonk.**

Definition 5.3.1. A **chonk** is a 3-tuple

$$C = (I, \gamma, a)$$

where:

- $I \in \mathcal{I}$ is a source image,
- $\gamma : T_D \to N$ is a narrative path synchronized with a,
- $a \in A_D$ is an audio signal.

The visual trajectory is derived as $\tau = \pi_T \circ \gamma$.

Definition 5.3.2. The **chonk space** $\mathcal{C} \subset \mathcal{P}$ consists of all synchronized pairings (viewing $\mathcal{C}$ as $(I, \pi_T \circ \gamma, a)$).

Remark. A chonk is minimal: image, narrative path, audio. Everything else is derived:

- The trajectory $\tau = \pi_T \circ \gamma$
- The mood arc $m = \pi_M \circ \gamma$
- The focal point f, defined as the limit point of $\tau(t)$ as zoom increases
- The final scale ϕ, defined as the scale component of $\tau(D)$

The chonk contains exactly the information needed to render, nothing more.

5.4. **The Coherence Measure.**

Definition 5.4.1. The **coherence** of a pairing $P = (I, \tau, a)$ is

$$\rho(P) = \exp\left(-\frac{1}{D}\int_0^D \|m_\tau(t) - m_a(t)\|^2 \, dt\right),$$

where $m_\tau : T_D \to M$ is a mood function induced by the trajectory (e.g. from its velocity or acceleration profile).

Remark. The exponential form is a convenient normalization that converts average squared mood mismatch into a similarity score in $(0, 1]$. Other monotone transforms would yield equivalent coherence orderings but different sensitivity profiles.

Proposition 5.4.2. *Coherence satisfies:*

- $\rho(P) \in (0, 1]$ *for all* $P \in \mathcal{P}$
- $\rho(P) = 1$ *if and only if* $m_\tau = m_a$ *(perfect synchronization)*
- $\rho(P) \to 0$ *as synchronization degrades*

5.5. **The Coherence Filtration.**

Definition 5.5.1. For $\varepsilon \in [0, 1]$, the **ε-coherent pairings** are

$$\mathcal{P}_\varepsilon = \{P \in \mathcal{P} \mid \rho(P) \geq \varepsilon\}.$$

Proposition 5.5.2. *The coherence filtration satisfies:*

- $\mathcal{P}_0 = \mathcal{P}$ *(all pairings)*
- $\mathcal{P}_1 \subset \mathcal{P}_\varepsilon \subset \mathcal{P}_{\varepsilon'} \subset \mathcal{P}_0$ *for* $1 \geq \varepsilon \geq \varepsilon' \geq 0$
- $\mathcal{C} = \mathcal{P}_\varepsilon$ *for some aesthetic threshold* $\varepsilon > 0$

Remark. This filtration is useful for optimization, interpolation, and generation.

5.6. **Synchronization Anchors.**

Definition 5.6.1. A set of **synchronization anchors** is

$$\Phi = \{(t_1, \tau_1), (t_2, \tau_2), \ldots, (t_n, \tau_n)\} \subset T_D \times T,$$

specifying that at audio time t_i, the visual transform should be τ_i.

Proposition 5.6.2. *Given anchors Φ and an interpolation scheme, the trajectory $\tau = \mathrm{Interp}(\Phi)$ satisfies the synchronization condition at the anchor points.*

Remark. This is exactly how the mood strategies work: timestamps provide t_i, and strategies compute the corresponding τ_i.

5.7. **The Synchronization Spectrum.**

Definition 5.7.1. The **coupling type** of a chonk characterizes how synchronization is achieved.

Coupling	Description	Example
Tight	Visual events at exact audio times	Jump cut on drum hit
Loose	Visual intensity follows audio intensity	Faster zoom during crescendo
Emergent	Overall mood alignment	Slow zoom with ambient audio

Remark. All three are valid chonks; they differ in how densely the anchors Φ sample the synchronization condition.

6. Specification vs. Artifact

It is useful to distinguish the **specification** (what the user requests) from the **artifact** (what gets rendered).

6.1. **The Specification.**

Definition 6.1.1. A **chonk specification** is a tuple

$$S = (I, W, \text{strategy}, \text{audio\_file}),$$

where:

- I is the source image,
- $W = [(f_1, \phi_1, t_1), \ldots, (f_n, \phi_n, t_n)]$ is a list of **waypoints**,
- strategy is a mood or interpolation strategy identifier,
- audio_file is an audio asset reference.

Remark. The waypoints W are constraints on the narrative path γ: at time t_i, the trajectory should be near focal point f_i at scale ϕ_i. The strategy determines how to interpolate between waypoints and synchronize with audio.

Example. The current Chonke implementation uses a single waypoint

$$W = [(f, \phi, D)],$$

specifying only the endpoint.

More complex specifications could include:

- Multiple focal points (pan from face A to face B)
- Scale keyframes (zoom in, hold, zoom out)
- Intermediate timing constraints

6.2. **The Generation Map.**

Definition 6.2.1. The **generation map** is

$$\text{Generate} : \text{Spec} \to \mathcal{C}, \qquad \text{Generate}(S) = ([I], \gamma_S, a),$$

where γ_S is the narrative path produced by applying the strategy to satisfy the waypoint constraints while synchronizing with audio.

Remark. The specification provides **boundary conditions** on γ. The strategy provides the **dynamics**. The audio provides **timing cues**. Together they determine the narrative path.

7. The Image Orbit Structure

7.1. **The Group Action.**

Definition 7.1.1. The similarity group T acts on images by

$$(\tau \cdot I)(x) = I(\tau^{-1}(x)).$$

7.2. **Orbit Equivalence.**

Definition 7.2.1. Two images are **orbit-equivalent** if

$$I_1 \sim I_2 \iff \exists \tau \in T : I_2 = \tau \cdot I_1.$$

Definition 7.2.2. The **orbit** of an image is its equivalence class

$$[I] = \{\tau \cdot I \mid \tau \in T\}.$$

Remark. An orbit represents visual content independent of framing—the scene itself, not how it is cropped.

7.3. **The Principal Bundle.**

Proposition 7.3.1. *The orbit projection $\pi : I \to I/T$ defines a principal T-bundle*

$$T \hookrightarrow I \to I/T.$$

Remark. A chonk trajectory $\tau : T_D \to T$ is a path in the fiber over a fixed point $[I]$ in the base. This is the same structure as gauge theories, frame bundles, and camera rigs.

7.4. **Refined Chonk Definition.**

Definition 7.4.1. A chonk depends on an orbit, not an image:

$$C = ([I], \gamma, a),$$

where $[I] \in I/T$ is the image orbit.

Corollary 7.4.2. *Two chonks are **fiber-compatible** if $[I_1] = [I_2]$. Fiber-compatible chonks can be interpolated without image blending.*

8. Algebraic Structure

The chonk space admits natural algebraic operations, with some important constraints.

8.1. **Temporal Concatenation.**

Definition 8.1.1. For fiber-compatible chonks

$$C_1 = ([I], \gamma_1, a_1), \quad C_2 = ([I], \gamma_2, a_2),$$

define **concatenation**

$$C_1 \oplus C_2 = ([I], \gamma_{12}, a_1 \oplus a_2),$$

where γ_{12} plays γ_1 then γ_2, and $a_1 \oplus a_2$ concatenates audio.

Proposition 8.1.2. *For each orbit $[I]$, the fiber*

$$\mathcal{C}_{[I]} = \{C \in \mathcal{C} \mid C \text{ has orbit } [I]\}$$

forms a monoid under $\oplus$ as raw pairings, with identity $\varepsilon_{[I]}$.

Remark. This is a fibered monoid: concatenation only makes sense within a fiber.

Remark. Concatenation does not preserve coherence in general:

$$\rho(C_1) \geq \varepsilon \quad \text{and} \quad \rho(C_2) \geq \varepsilon \quad \not\Rightarrow \quad \rho(C_1 \oplus C_2) \geq \varepsilon.$$

Coherence can degrade due to trajectory discontinuity, mood mismatch, or non-local effects.

Definition 8.1.3. A **chonk sequence** is an ordered list $(C_1, C_2, \ldots, C_n)$ where consecutive chonks may have different orbits.

Remark. When $[I_1] \neq [I_2]$, transitions require cuts or blends.

8.2. **Time Scaling.**

Definition 8.2.1. For $\lambda > 0$, define **time scaling**

$$\lambda \cdot C = ([I], \gamma \circ (t \mapsto t/\lambda), a \circ (t \mapsto t/\lambda)).$$

Proposition 8.2.2. *Time scaling defines an action of $(\mathbb{R}^+, \cdot)$ on $\mathcal{C}$.*

8.3. **Temporal Reversal.**

Definition 8.3.1. Define **reversal**

$$\widetilde{C} = ([I], \gamma \circ (t \mapsto D - t), \tilde{a}).$$

Proposition 8.3.2. *Reversal is an involution satisfying*

$$\widetilde{\widetilde{C}} = C, \qquad (C_1 \oplus C_2)^{\sim} = \widetilde{C}_2 \oplus \widetilde{C}_1.$$

9. Differential Structure

Since trajectories are paths in a Lie group, we have calculus.

9.1. **Trajectory Derivatives.**

Definition 9.1.1. For $\tau : T_D \to T$, the **velocity** is

$$\tau'(t) = \frac{d\tau}{dt} \in \mathfrak{t},$$

where $\mathfrak{t}$ is the Lie algebra of T.

For $\tau(t) = (s(t), x(t))$, the velocity in log-coordinates is

$$\tau'(t) = \left(\frac{s'(t)}{s(t)},\ x'(t) - \frac{s'(t)}{s(t)} x(t) \right).$$

9.2. **Kinetic Energy.**

Definition 9.2.1. The **kinetic energy** of a trajectory is

$$E(\tau) = \int_0^D \|\tau'(t)\|^2 \, dt.$$

Remark. High energy corresponds to fast, dramatic motion. Different mood strategies implicitly optimize different energy profiles.

9.3. **Jerk.**

Definition 9.3.1. The **total jerk** is

$$J(\tau) = \int_0^D \|\tau'''(t)\|^2 \, dt.$$

Remark. Jerk measures abruptness. Jump cuts have infinite jerk; smooth zooms have low jerk.

10. Mood Strategies as Trajectory Generators

10.1. **The Strategy Interface.**

Definition 10.1.1. A **mood strategy** is a function

$$\text{Strategy} : \text{RuntimeInfo} \to K,$$

where

$$\text{RuntimeInfo} = (I, f, \phi, D, \Phi_{\text{audio}}).$$

10.2. **Examples.**

Example (Dramatic). Jump cut at first timestamp to 50% of final zoom, followed by exponential zoom progression.

Example (Ominous). Approach–retreat oscillations toward the focal point with damped amplitude.

Example (MicDrop). Smooth buildup to a slam point with damped harmonic oscillation:

$$s(t) = s_{\text{target}} \left(1 + Ae^{-\gamma t} \cos(\omega t)\right).$$

11. Information Budget (Informal)

Note. This section presents an intuition about information content, not a rigorous entropy decomposition.

11.1. **The Information Budget Intuition.** Informally, the information content of a chonk can be expressed as

$$\text{“Info”}(C) \approx \text{Info}(I) + \text{Info}(\gamma \mid I) + \text{Info}(a \mid I, \gamma).$$

11.2. **Compression Implication.** The representation $(I, \text{mood\_name}, f, \phi, \text{audio\_ref})$ is parsimonious and compresses well.

12. Summary

Theorem 12.0.1 (Chonk Characterization)**.** *A chonk is a synchronized path through the narrative bundle $N = M \times T$, where synchronization means the mood component tracks the audio's emotional content:*

$$C = ([I], \gamma, a) \quad \textit{with} \quad \pi_M \circ \gamma \approx m_a.$$

13. Future Directions

13.1. **Optimization-Based Generation.**

Definition 13.1.1. Given image I, audio a, and waypoints W, the coherence-optimal trajectory problem is

$$\max_{\tau} \ \rho(I, \tau, a) \quad \text{subject to waypoint constraints.}$$

13.2. **Energy-Constrained Generation.**

Definition 13.2.1. The energy-constrained objective is

$$\max_{\tau} \ \rho(I, \tau, a) - \lambda J(\tau).$$

13.3. **Search Over Strategy Space.**

Definition 13.3.1. The **coherence distribution** of a strategy S is the distribution of

$$\rho(\text{Generate}(S, \cdot))$$

over random inputs.

13.4. **Learning.** Problems include learning m_a, learning m_τ, and learning direct audio-to-trajectory mappings.

13.5. **Composable Narratives.** Longer-form content can be structured via chonk concatenation with appropriate boundary conditions.

13.6. **Multi-Image Extensions.**

Definition 13.6.1. A **transition chonk** between images I_1 and I_2 is

$$C_{\text{trans}} = (I_1, I_2, \gamma, a, \text{blend}).$$

13.7. **Tooling Implications.** The specification identifies exactly what the user must provide: image, waypoints, strategy, and audio reference.

Appendix A. Notation

Symbol	Meaning
$\mathcal{I}$	Image space
T	Similarity transform group
M	Mood manifold
$N = M \times T$	Narrative bundle
$\mathcal{P}$	Pairing space: $\mathcal{I} \times C^0(T_D, T) \times A_D$
$\mathcal{C}$	Chonk space (synchronized pairings)
γ	Narrative path: $T_D \to N$
τ	Visual trajectory ($\pi_T \circ \gamma$)
f	Focal point (derived from τ)
ϕ	Final scale (scale component of $\tau(D)$)
m_a	Audio mood function
ρ	Coherence measure
$\mathcal{P}_\varepsilon$	ε-coherent pairings
Φ	Synchronization anchors
$\oplus$	Temporal concatenation
$\lambda \cdot C$	Time scaling
$\widetilde{C}$	Temporal reversal
$[I]$	Image orbit

Appendix B. Implementation Mapping

Mathematical Object	Code
Image orbit $[I]$	Source `UIImage` identity (provenance)
Narrative path γ	`MoodEffectStrategy.computeKeyframeTransforms()` output
Audio a	Audio file asset
Focal point f	`cropRectangleCenter: CGPoint`
Target scale ϕ	`cropRectangleScale: CGFloat`
Mood strategy	`MoodEffectStrategy` protocol
Sync anchors Φ	`effectTimestamps` array
Trajectory τ	`[KeyframeTransform]` + interpolation
Audio mood m_a	Implicit in timestamp placement
Coherence ρ	(not yet implemented)

Appendix C. Category-Theoretic Perspective (Speculative)

Note. This appendix presents a categorical framing that is mathematically valid but currently underpowered.

C.1. **Chonks as Morphisms.**

Definition C.1.1. The category **Chonk** has:

- Objects: pairs $([I], f)$ — an orbit with focal point
- Morphisms: chonks $C : ([I_1], f_1) \to ([I_2], f_2)$
- Composition: temporal concatenation $\oplus$
- Identity: zero-duration chonks

Remark. This is equivalent to the monoid structure described earlier.

C.2. **Functors.**

Definition C.2.1. A **mood functor** is a map

$$M : \text{Spec} \to \textbf{Chonk}$$

assigning chonks to specifications.

Remark. Natural transformations between mood functors represent systematic mood transformations.

C.3. **Future Directions.** The categorical perspective becomes more powerful when extended to functors between mood manifolds, fibered categories, and operads for multi-input composition.