Full Mind Transfer:
Bandwidth vs Fidelity in Activation-Level Coordination

Jeremy McEntire[1]

March 2026

Abstract

Paper XVI showed that INLP projection injection produces marginal improvement (+1.9%) over text-only coordination when measured by domain classification — a metric that saturates at 95%. This paper replaces classification with Representational Similarity Analysis (RSA) and next-token KL divergence to probe fine-grained structure. Eight transmission conditions span from text-only (0 injected dimensions) through INLP (36d), PCA tiers (50d, 100d, 200d), and full activation (3584d), plus a scaffold-free condition (BOS token + full activation).

The central finding: **text and activation injection carry fundamentally different information.** All text-based conditions cluster at RSA ≈ 0.11, regardless of injection bandwidth (0–3584 dimensions). Injection does not improve representational fidelity when text is present. However, activation injection *without* text (BOS + full) achieves RSA $= 0.47$ — $4\times$ higher. The activation channel preserves geometric structure that text destroys. Within-domain RSA shows a domain reversal: text best preserves legal structure (0.23), while activation injection best preserves science (0.64).

Despite $4\times$ higher geometric fidelity, activation injection does not improve output alignment: KL divergence is similar across all conditions (6.97–7.35 nats). The forward pass from layer 10 to the output head dilutes the structural advantage. Geometric preservation and functional alignment are dissociated.

1 Introduction

Paper XVI established that domain classification accuracy is a poor dependent variable for coordination experiments — the text baseline reaches 95%, leaving only 5% headroom. INLP injection captured 38% of that headroom (+1.9%), but the metric is too coarse to probe what text does and does not preserve.

This paper uses two finer-grained metrics:

- **RSA (Representational Similarity Analysis)**: Spearman correlation between pairwise cosine distance matrices at the terminal layer. Captures the full geometry of

[1]Correspondence: jmc@cageandmirror.com

the representational space — not just domain labels but which probes are similar to which.

- **Next-token KL divergence**: Approximate $\mathrm{KL}(P_{\text{sender}} \| Q_{\text{receiver}})$ from top-100 token probabilities. Measures functional output alignment.

RSA decomposes into between-domain RSA (does the receiver preserve which domain each probe belongs to?) and within-domain RSA (within a domain, does the receiver preserve which probes are similar to which?). Within-domain RSA is the critical metric: it probes fine-grained expertise rather than coarse domain identity.

2 Methods

2.1 Transmission conditions

All conditions use Qwen 2.5-7B with injection at layer 10 ($\alpha = 1.0$, the optimal value from Paper XVI) and capture at the terminal layer (27).

1. **Text only**: 3-sentence summary, no injection.
2. **Text + INLP (36d)**: Summary + INLP projection of sender's centered layer-10 activation.
3. **Text + PCA-50**: Summary + projection onto top-50 PCA components.
4. **Text + PCA-100**: Summary + projection onto top-100 PCA components.
5. **Text + PCA-200**: Summary + projection onto top-200 PCA components.
6. **Text + Full (3584d)**: Summary + full centered layer-10 activation.
7. **BOS + Full (3584d)**: Single BOS token + full centered activation. No text scaffold.
8. **BOS only**: Single BOS token, no injection. Baseline control.

2.2 Projection bases

INLP directions (36) are computed at layer 10 via iterative ridge regression (9 per domain). PCA components are computed from the centered sender activations. Injection vectors are centered (mean-subtracted) before projection.

The projection norms provide context: INLP 36d captures mean norm 12.4 (47% of full norm 26.6), consistent with the 22.6% variance fraction at layer 10. PCA-50 captures 91% of the full norm; PCA-100 captures 97%; PCA-200 and full are identical (100%).

3 Results

3.1 RSA: text dominates injection

Table 1: RSA and KL divergence across transmission conditions. RSA decomposed into overall, between-domain, and within-domain (mean across four domains).

Condition	Dims	RSA	RSA_{bw}	RSA_{wi}	KL	Cos
Text only	0	0.107	0.063	0.112	7.35	0.677
Text + INLP	36	0.116	0.069	0.116	7.35	0.678
Text + PCA-50	50	0.116	0.067	0.126	7.30	0.684
Text + PCA-100	100	0.115	0.067	0.125	7.30	0.685
Text + PCA-200	200	0.114	0.065	0.124	7.30	0.685
Text + Full	3584	0.114	0.065	0.124	7.30	0.685
BOS + Full	3584	**0.466**	**0.403**	**0.469**	**6.97**	0.398
BOS only	0	0.000	0.000	0.000	6.98	0.398

All text-based conditions cluster at RSA ≈ 0.11, with injection providing $< 1\%$ improvement regardless of bandwidth. The text representation determines the terminal-layer geometry; injection is overwritten.

BOS + Full achieves RSA $= 0.47$ — $4\times$ higher than any text condition. Without text to overwrite the injected signal, the full activation propagates through the forward pass and preserves the sender's representational geometry at the terminal layer. BOS alone produces RSA $= 0$: a single token without injection generates a fixed representation with no probe-specific structure.

3.2 Bandwidth saturation at 100 dimensions

PCA-100, PCA-200, and Full produce identical results (RSA 0.114–0.115, KL 7.30, Cos 0.685). The effective bandwidth ceiling is $\sim$100 dimensions. Above this, additional dimensions carry no variance and add no information. This is consistent with the effective dimensionality $d_{\text{eff}} \approx 20$ from Paper XI: the activation energy is concentrated in a low-dimensional subspace, and 100 PCA components capture it all.

INLP (36d) captures less energy (47% of norm) but provides comparable RSA (0.116 vs 0.116 for PCA-50). The domain-discriminative subspace overlaps substantially with the top PCA directions.

Table 2: Within-domain RSA by domain and condition. Text and activation injection preserve different domain structures.

Condition	Medical	Legal	Code	Science
Text only	0.121	0.234	0.108	−0.015
Text + INLP	0.129	0.238	0.107	−0.013
Text + PCA-50	0.142	0.258	0.109	−0.006
Text + Full	0.140	0.258	0.106	−0.008
BOS + Full	0.398	0.355	**0.485**	**0.638**

3.3 Within-domain RSA: a domain reversal

Text-based conditions show a clear domain hierarchy: Legal (0.23) $\gg$ Medical (0.12) $>$ Code (0.11) $\gg$ Science (-0.01). Legal domain structure survives text compression best, likely because legal language has distinctive vocabulary and phrasing that summaries preserve. Science structure is effectively destroyed by text (RSA ≈ 0).

BOS + Full reverses this hierarchy: Science (0.64) $\gg$ Code (0.48) $>$ Medical (0.40) $>$ Legal (0.35). Activation injection best preserves science domain structure — the same structure that text completely destroys. This reversal indicates that the two modalities carry complementary domain-specific information: text carries lexical/semantic structure (strong for legal), while activations carry computational/geometric structure (strong for science).

3.4 KL divergence: geometric preservation $\neq$ functional alignment

Despite 4$\times$ higher RSA, BOS + Full has KL divergence (6.97) only marginally better than text conditions (7.30–7.35). The 5% geometric improvement translates to $<$ 5% output distribution improvement. The forward pass from layer 10 to the output head is a many-to-one mapping that dilutes the geometric advantage: many different layer-10 geometries produce similar output distributions.

4 Discussion

4.1 Two modalities, two types of information

The central result is a dissociation: text preserves *functional* content (what the model says) while activations preserve *geometric* structure (how the model represents). These are not the same information. A text summary that says "the medical analysis focused on cardiac arrhythmias" carries the semantic content without the computational geometry — which specific activation patterns were active during the analysis.

The domain reversal in within-domain RSA (Table 2) makes this concrete. Legal language has distinctive surface features ("whereas," "pursuant to," "the Court finds") that text summaries preserve, maintaining legal's within-domain structure. Science has less distinctive vocabulary but more distinctive computational patterns (mathematical reasoning, experimental logic, causal inference chains) that text strips but activations retain.

4.2 Text as the binding constraint

When text is present, it determines the receiver's representational geometry at the terminal layer. Injection at layer 10 is overwritten by the time the activation reaches layer 27: the 17 intervening transformer layers, each performing attention over the text tokens, reshape the representation to match the text content. The injected signal is not destroyed — it is *dominated*.

This explains why all text conditions have identical RSA regardless of injection bandwidth: the text contribution to the terminal representation is $\gg$ the injection contribution at $\alpha = 1.0$. Higher α might change this balance, but at the cost of distorting the representation away from meaningful processing.

4.3 Implications for coordination protocols

1. **Text is the primary channel.** For functional coordination (getting agents to produce compatible outputs), text summaries preserve 63.7% of the relevant information (Paper XVIII). Activation injection adds $< 2\%$.

2. **Activation injection preserves complementary structure.** The $4\times$ RSA advantage of BOS + Full indicates that activation-level communication carries information text cannot. Whether this information is *useful* depends on the task.

3. **The effective bandwidth is $\sim$100 dimensions.** Beyond 100 PCA components, additional bandwidth is wasted. The domain-discriminative subspace (36 INLP directions) is a near-optimal compression: it achieves comparable RSA to PCA-50 with fewer dimensions.

4. **Multi-layer injection may be necessary.** Single-layer injection at layer 10 is overwritten by text processing. Injection at multiple layers, or at the terminal layer, might maintain the injected signal through the forward pass. This is a design question for activation-sharing protocols.

5 Conclusion

Text and activation injection are complementary coordination modalities that carry different types of information. Text preserves functional content (what was done) and determines the receiver's terminal representation when present. Activation injection preserves geometric structure (how it was represented) and achieves 4× higher RSA when text is absent. But geometric preservation does not translate to functional alignment: KL divergence is similar across all conditions.

The effective bandwidth of the activation channel is ∼100 dimensions. The INLP projection (36 dimensions) provides near-optimal domain-discriminative compression. The bandwidth vs fidelity curve saturates rapidly — there is no gradual improvement from INLP to full activation when text is present.

Data Availability

All results are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper17/`.

Series: Activation Geometry of Domain-Selective Noise Injection, Paper XVII.