WHEN TRANSFORMERS COPY INSTEAD OF COMPOSE: CIRCUIT COMPETITION UNDER SHARED SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models are often given the same semantic event realized in several modes at once: an image and its description, a sentence and its translation, a specification of a program and its implementation. The semantics are shared across views while the realization is not, so a model can form two competing circuits: one that computes the content in the target view and one that copies it from the source view. We study this phenomenon in detail with synthetic data generated from probabilistic context-free grammars (PCFGs). Specifically, we hold view difficulty fixed and dial the levels of shared structure: a controlled hierarchical setting where two views can share semantic structure down to a precisely controlled level L_S, with exact Bayes-optimal ceilings determining how much information is available from each view at each level. Probing the residual stream at every layer, we find the following: First, fusion between the views happens at the deepest shared latent level (the fusion boundary), the level L_S where the shared information first becomes sufficient for the target view. Second, the availability of a copying route weakens composition at semantic levels above L_S: once the model can copy the shared content from the source view, its further composition in the target view weakens. Third, this weakening is reversible: removing the target's access to the source, or otherwise reducing the copying route's reliability, restores the missing computation, which coexists with the copying route rather than being lost. Finally, we observe a consistent pattern in LLaVA-1.5-7B, where blocking attention to a conflicting caption restores image-based answers, and discuss implications for fusion in multimodal models.

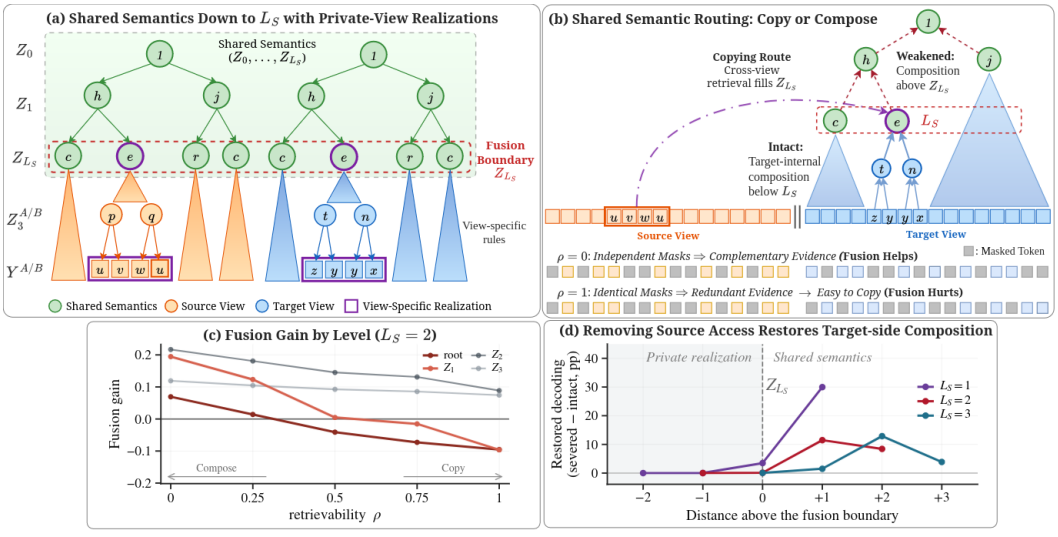

Figure 1: **Shared semantics create a competition between copying and composition.** (a) Two views share a hierarchy through the fusion boundary Z_{L_S}, below which their realizations are view-specific. (b) At the fusion boundary, the target can either compose internally or retrieve the corresponding representation from the source; above Z_{L_S}, these routes compete. (c) As source evidence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Figure 1: (Continued) becomes more retrievable and redundant ($\rho \uparrow$), fusion gain above the boundary decreases and becomes negative: despite access to more shared information, the target recovers less through its own composition. (d) Severing target-to-source attention restores this missing computation specifically above Z_{L_S}, showing that the target-side composition circuit remains available. **Thus, a second view can provide useful evidence while simultaneously suppressing target-side composition when shared semantics are easy to copy.**

1 INTRODUCTION

Multi-view and multimodal learning are motivated by a simple idea: multiple views of the same semantic event can provide additional information for learning richer representations. By a *view* we mean one realization of an underlying semantic event: the same content expressed in a different modality, language, or format, so that views share semantics but differ in surface form. An image may be accompanied by a caption, a video by its audio track, or a document by a summary. In each case, information from one view can be used to construct a representation of the other. This intuition underlies a broad range of multimodal and multi-view methods (Baltrušaitis et al., 2019; Liang et al., 2024; Li et al., 2019; Yu et al., 2024): access to additional views should provide additional evidence for learning.

Yet an additional view does not always help. Multimodal models can underperform their strongest unimodal counterpart, and prior work largely attributes this to optimization: one modality may learn faster (Wang et al., 2020), dominate the training signal (Wu et al., 2022), or leave another modality under-trained (Peng et al., 2022). These accounts explain *which view* the model relies on, but not *how access to another view changes the way information is obtained and represented*. The question is sharpest when views share semantics but realize them differently. Given an image and a caption of the same scene, a model can compose the answer within the image, or retrieve it from the caption. Here *composition* means constructing a higher-level feature from the representations of its lower-level parts, as when a scene is built from its objects or a sentence's meaning from its phrases, whether those parts come from one view alone or have already been combined across views. We call the use of information across views to construct a representation *fusion*. Shared semantics open several routes to the same content (Figure 1b): the target can compose it from its own parts, fuse its own evidence with the source's, compose it from parts already fused with the source, or *copy* it from the source. We use *copy* as an umbrella term for any mechanism by which the target obtains a shared latent representation from the source's representation of it rather than composing it. This is not token-level copying, since the two views realize it differently, and Section 5 localizes a mechanism that implements it. Our central question is therefore: *when does fusion enrich a representation, and when does access to another view reduce reliance on within-view composition?*

Real multimodal data let us label some shared semantics, such as whether both an image and its caption mention a dog, but not the full latent hierarchy behind them, where the sharing between views stops, or how much each view determines at each level. Answering this question requires all three. We therefore turn to synthetic data, where the generative process is fully specified. Probabilistic context-free grammars (PCFGs) are a natural choice: they generate sequences through a hierarchy of latent variables, the classical model of compositional structure in language (Chomsky, 1956; Manning & Schütze, 1999; Lari & Young, 1991), and exact inference over them is tractable. The Random Hierarchy Model (RHM) (Cagnetta et al., 2024; Cagnetta & Wyart, 2024) turns PCFGs into a controlled testbed for deep networks by sampling unambiguous grammars with fixed branching, depth, and number of synonymous rules, so that hierarchical composition can be studied without hand-designed rules. The correspondence to multimodal data is at the level of shared content and its realization: an image and its caption share high-level semantics (a dog, on the left) but realize them through different low-level rules, patches in one view and words in the other. In a two-view RHM, shared latent levels play the role of this content and view-specific rules below them play the role of modality-specific realization. We extend the RHM to two views that share a known hierarchy down to a controlled depth L_S and are realized by independent rules below it; the shared latent representation Z_{L_S} defines a *fusion boundary*. Because the generative process is known exactly, we compute Bayes-optimal ceilings that separate the information available from the inputs from the information recovered by the representation.

We train a causal transformer on source–target pairs, in which the target can attend to the source but not the reverse, and probe its residual stream at every layer against these ceilings. We vary the

shared depth L_S and the retrievability ρ, which controls how redundant the evidence in the two views is while keeping each view's difficulty fixed. Fusion helps when the evidence is complementary, but as it becomes redundant, target-side composition above Z_{L_S} weakens even though more shared information is available, while Z_{L_S} itself is recovered essentially perfectly. The target develops a distributed cross-view retrieval route to the source span of Z_{L_S}, which strengthens with ρ. Severing this route at inference restores the missing composition in an already-trained model, showing that the two circuits coexist; comparing against a model trained without the route shows that joint training scaffolds composition when little is shared and starves it when much is. Finally, a consistent pattern appears in LLaVA-1.5-7B (Liu et al., 2024): with a conflicting caption, answers agree with the caption on 98.3% of examples, and blocking attention to the caption's answer word restores agreement with the image to 71.0%, close to the 73.0% the model reaches on the same examples without a caption.

Together, our results show that shared semantics open specific new routes to a latent representation, and that which route the model uses changes what becomes recoverable within the target view, even when the view's own evidence is unchanged.

Our contributions are:

- **A controlled framework for studying multi-view fusion.** An extension of the Random Hierarchy Model with controlled shared semantics and exact Bayes-optimal ceilings, separating the information available across views from the representations formed within a view.
- **Fusion can hurt, through circuit competition.** A second view improves the target's representation when its evidence is complementary, but weakens target-side composition above the fusion boundary as the evidence becomes redundant, through a distributed, span-level retrieval route that strengthens with retrievability ρ.
- **Causal evidence that the two circuits coexist.** Severing the route at inference restores composition above Z_{L_S} without changing weights or inputs, and a model trained without it shows that joint training both scaffolds and starves target-side composition.
- **Consistent evidence in a pretrained multimodal model.** In LLaVA-1.5-7B, a conflicting caption overrides the image, and blocking attention to the caption's answer restores agreement with the image; we read this as consistent with, not a demonstration of, the synthetic mechanism.

2 Related work

We summarize the closest lines of work here; Appendix G gives an extended discussion.

When fusion helps and when it hurts. Multi-view and multimodal learning assume that views carry complementary information, so fusing them should be no worse than the best single view (Baltrušaitis et al., 2019; Li et al., 2019; Yu et al., 2024; Huang et al., 2021; Federici et al., 2020). Failures are usually explained by optimization and concern *which* view the model relies on (Wang et al., 2020; Wu et al., 2022; Peng et al., 2022; Liang et al., 2022), while alignment objectives aim to bring the views into a common representation (Radford et al., 2021; Li et al., 2023a; Huh et al., 2024). We instead hold each view's difficulty fixed, vary the shared depth L_S and the redundancy of the views' evidence ρ, and ask which route supplies the target's representation.

Controlled hierarchical data and shortcuts. The Random Hierarchy Model and related synthetic hierarchies built on PCFGs (Cagnetta et al., 2024; Cagnetta & Wyart, 2024; Allen-Zhu & Li, 2023; 2024; Chomsky, 1956; Manning & Schütze, 1999; Lari & Young, 1991) make every latent level explicit, which lets us score probes against exact ceilings and address the known limits of probing (Alain & Bengio, 2017; Belinkov, 2022; Hewitt & Liang, 2019; Voita & Titov, 2020; Pimentel et al., 2020). Compositional-generalization benchmarks (Lake & Baroni, 2018; Kim & Linzen, 2020; Hupkes et al., 2020; Dziri et al., 2023) vary recombination within one stream, not structure shared across views. Networks prefer predictive structure that is cheap to compute (McCoy et al., 2019; Geirhos et al., 2020; Pezeshki et al., 2021; Shah et al., 2020); here the shortcut is created by shared semantics rather than by a spurious correlate.

Retrieval circuits and conflicting sources. Mechanistic work identifies copying and retrieval heads (Olsson et al., 2022; McDougall et al., 2023; Wu et al., 2025) and circuits via activation patching (Wang et al., 2023; Meng et al., 2022; Zhang & Nanda, 2024). The route we find is cross-view and reads a span realized differently from the target's own realization. Work on conflicting sources, including image against caption, studies which source a model trusts (Longpre et al., 2021; Xie et al., 2024; Frank et al., 2021; Hua et al., 2025; Ortu et al., 2025; Li et al., 2023c); our LLaVA experiment adds a causal intervention.

3 SYNTHETIC FRAMEWORK AND EXPERIMENT SETUP

We describe the two-view generative model, the transformer and probes we train on it, their Bayes-optimal ceilings, and the masking and attention metrics we use.

3.1 HIERARCHICAL GENERATIVE MODEL WITH CONTROLLABLE SHARED SEMANTICS

We construct our hierarchical generative model from the Random Hierarchy Model (RHM) (Cagnetta et al., 2024), an L-level compositional framework based on an ensemble of PCFGs (Chomsky, 1956; Manning & Schütze, 1999; Lari & Young, 1991). Each node in the hierarchy expands into b lower-level nodes, giving a tree with fixed branching factor and depth L. Thus, each view contains b^L leaves, which form the observed input sequence, while the internal nodes represent progressively higher-level features obtained by composing these leaves. We index the hierarchy by depth d, from the root ($d = 0$) to the leaves ($d = L$), so there are b^d nodes at depth d. At each level, a higher-level feature has m alternative production rules, yielding m equivalent lower-level representations of the same feature, which the RHM calls *synonyms*. Rules are unambiguous: no lower-level tuple is produced by more than one higher-level feature, so a fully observed span determines its parent uniquely (Cagnetta et al., 2025). The composition rules are sampled uniformly at random, giving a fully specified but non-hand-designed hierarchical structure (Cagnetta et al., 2024; Cagnetta & Wyart, 2024; Allen-Zhu & Li, 2023).

We extend the RHM to generate two views that realize the same *shared semantics*: they share the hierarchy from the root down to depth L_S (Figure 1), and at $L_S = 0$ only the root class. We call Z_{L_S}, the deepest shared level, the *fusion boundary*, and say that a level lies *above* it when it is closer to the root ($d < L_S$) and *below* it when it is closer to the leaves ($d > L_S$). Beyond the fusion boundary, the two views develop independently using view-specific composition rules, producing different realizations of the same shared semantics. Thus, L_S directly controls how deeply the semantic structure is shared between the views, while b, m, and L control the branching, number of alternative realizations, and overall hierarchical depth of each view.

3.2 MODEL AND EXPERIMENT SETTING

We present the two views to a causal transformer (Vaswani et al., 2017; Radford et al., 2019) as a source–target pair. The source view is placed before the target view in the sequence, so target positions may attend to the source, while source positions cannot attend to the target. We train an 8-layer transformer from scratch with hidden dimension 128, 4 attention heads, an inner dimension of 512, and RoPE positional embeddings (Su et al., 2024). We probe the residual stream at every layer to measure the information about the hierarchical target variables that has been computed at each stage (Alain & Bengio, 2017; Elhage et al., 2021; Li et al., 2023b; Allen-Zhu & Li, 2024; Meng et al., 2022; Wang et al., 2023). Probes are logistic classifiers fit on an 80/20 split; a fresh probe is fit for every model and every intervention, including the severed runs of Section 6. To isolate the role of cross-view information, we can sever the retrieval route by masking the target-to-source attention submatrix, leaving the target to solve the task using only information available within its own view.

3.3 READOUT: DECODING AND PREDICTION

We use two complementary readouts of the residual stream, which probe different latent representations at the same position: the canonical span boundary where a hierarchical latent representation can first have been formed. For a node at depth d and position q, this is the first position at which the node's entire leaf span has been observed under the causal mask.

Decoding. The probe predicts the latent RHM node that has just closed at the current position. Its entire leaf span is therefore visible. Decoding asks whether the model has *computed and retained* the corresponding hierarchical representation $Z^B_{d,q}$ from the residual stream at its span boundary. This readout measures what is recoverable about the latent representation where it can first have been composed, not whether the model's output depends on it (Section 9); throughout, weakened and restored composition refer to this recoverability. It is therefore the primary readout for circuit-competition and severance experiments (Wang et al., 2023; Zhang & Nanda, 2024; Tenney et al., 2019).

Prediction. The probe instead predicts the *next* latent RHM node, whose leaf span has not yet been observed. Because of the causal mask, the model must already have computed the information needed to predict that latent node from the preceding context. Prediction therefore asks whether the relevant computation is already present *before the target span is observed*. Prediction never means next-token prediction here, which remains the training objective throughout. We average probe accuracy over the b^d nodes at each depth and, unless otherwise stated, report the probe at the output layer for each level (Belinkov, 2022; Alain & Bengio, 2017).

3.4 Bayes-optimal reference ceilings

Raw probe accuracy does not by itself reveal whether a representation contains information that the model had to compute or information that was already available from the target prefix (Hewitt & Liang, 2019; Voita & Titov, 2020). Because the complete generative process is known, we compare probe performance against exact Bayes-optimal ceilings. We define a *prefix ceiling* ($\mathrm{BayesPrefix}$) as the optimal accuracy achievable from the target prefix alone, a *shared ceiling* ($\mathrm{BayesInput}$) as the optimal accuracy when the source is available under the known grammar, and an *oracle ceiling* ($\mathrm{BayesSharedOracle}$) as the optimal accuracy when the true shared latent representations are known. The gap between the prefix ceiling and the other two quantifies the target information available through the shared semantics beyond what can already be inferred from the target prefix. Because RHM rules are unambiguous, the shared and oracle ceilings coincide when the source is fully observed; they differ only under the masking of Section 3.5. Appendix C provides their definitions and the belief-propagation procedure used to compute them.

3.5 Masking and retrievability

Without masking, every leaf of both views is observed, and because the grammar is unambiguous each fully observed span determines its ancestors. The source then repeats what the target already contains, so it can offer an alternative route to the target's latent levels but no additional evidence. To study when a second view contributes information rather than only a route, we mask each leaf independently with probability 0.5 in both views, in the ρ experiments of Section 4 and Section 5. Masking makes each view informative but incomplete, so the source can in principle fill in what the target is missing: this is the setting in which *fusion*, the target using information from the source to construct its own representation, can help.

Masking alone does not control whether the source's evidence complements the target's or duplicates it. We therefore introduce *retrievability* $\rho \in [0, 1]$: for each leaf, with probability ρ the target reuses the source's masking decision at the same position, and otherwise draws its own. The marginal masking probability is 0.5 in each view for every ρ, so the difficulty of each view on its own is unchanged; only the joint pattern varies. At $\rho = 0$, the patterns are independent, and a leaf masked in the target is often visible in the source, so the two views provide complementary evidence. At $\rho = 1$, the patterns are identical, so the source's visible leaves sit exactly where the target's do: its evidence is aligned with the target's and most directly retrievable, but adds the least new information. Retrievability ρ leaves the prefix and oracle ceilings unchanged, since the target's own masking distribution does not depend on it, but lowers the shared ceiling, which conditions on the observed source: it is highest at $\rho = 0$, where the source can reveal leaves masked in the target, and lowest at $\rho = 1$. Increasing ρ therefore reduces the information the source adds while making it easier to retrieve, which lets us separate the value of the source's evidence from the availability of a route to it. In Figure 2a and Section 6, we deliberately remove masking, which takes this redundancy to its limit and, in Section 6, makes the decoding ceiling exactly 100%.

3.6 Measuring cross-view retrieval

To locate where the target reads the source, we measure how each attention head distributes its cross-view attention. The shared latent representation Z_{L_S} spans $g = b^{L-L_S}$ leaves. Below Z_{L_S}, the two views are realized by different rules, so the g source leaves under a shared node form a span whose content corresponds to that node but is not a token-level copy of the target's span. This span is fixed by the hierarchy and sequence layout, independently of the model, and gives a ground-truth location against which attention can be scored.

Let $\alpha_{\ell h}(i, j)$ denote the attention of head h in layer ℓ from target position i to source position j, and let Δ be the separation between the two blocks. The cross-view attention of a head at relative offset δ is

$$A_{\ell h}(\delta) = \frac{1}{|B|} \sum_{\substack{i \in B \\ i-\delta \in A}} \alpha_{\ell h}(i,\, i-\delta), \qquad A_{\ell h} = \sum_{\delta} A_{\ell h}(\delta), \tag{1}$$

where B and A are the target and source positions. Its *span concentration* is the fraction of this mass that falls on the source counterparts of the next g target positions, which form the shared node's source span at block boundaries:

$$\sigma_{\ell h} = \frac{\sum_{\delta \in S} A_{\ell h}(\delta)}{A_{\ell h}}, \qquad S = \{\Delta - g, \ldots, \Delta - 1\}. \tag{2}$$

$A_{\ell h}$ measures how much a head reads the source, and $\sigma_{\ell h}$ measures how specifically it reads the source realization of the shared latent level. Descriptively, we track how $A_{\ell h}$ and $\sigma_{\ell h}$ change with ρ and L_S; causally, we rank heads by $\sigma_{\ell h}$ and block the cross-view attention of the top k. Since blocking any source-reading head removes some source information, we compare against two controls with the same k: *volume-matched* heads, which remove a comparable total cross-view attention $A_{\ell h}$ without selecting for the span, and *random* heads. A larger effect than volume-matched blocking shows that it matters *where* heads read the source, not *how much*.

4 Fusion can help or hurt

We now measure how access to the source changes what the target represents at each level. We report the prediction readout against the exact ceilings of Section 3.4: because it probes a latent representation before its span is observed, it isolates what the target infers from its prefix and from the source.

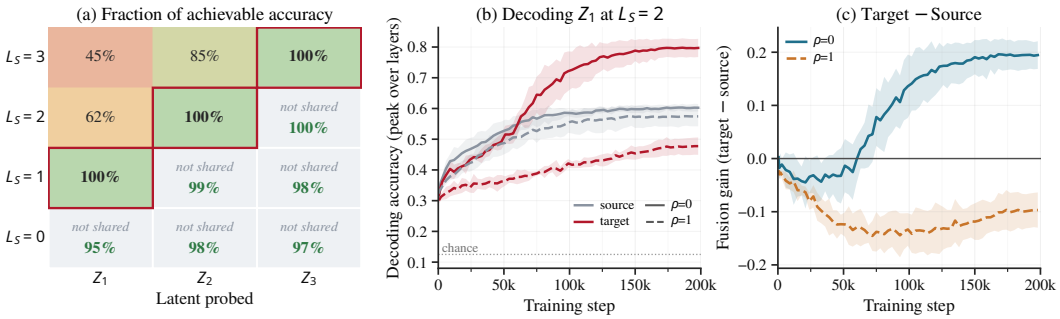

Figure 2: **A second view can help or hurt, and where it hurts is set by the fusion boundary.** (a) Target prediction accuracy without leaf masking, as a fraction of the oracle ceiling; rows are L_S, columns the latent representation probed, the red outline marks Z_{L_S}, and grey cells are private levels (Table 1). (b) Decoding accuracy of Z_1 at $L_S = 2$ (peak over layers) with 50% leaf masking, for source and target at $\rho = 0$ (solid) and $\rho = 1$ (dashed). (c) Target minus source from (b), paired within seed. Five seeds, ± 1 s.d.

Private levels are unaffected, and the boundary is recovered completely. Below the boundary ($d > L_S$), the two views share nothing, and the target reaches 97–100% of the Bayes ceiling at every L_S (Figure 2a, grey cells). Sharing structure higher in the hierarchy does not change how the target

composes its private levels. At Z_{L_S}, the target predicts the shared node at 100% of the ceiling for every $L_S \geq 1$. Its span has not yet been observed, and the target prefix alone supports only 25–51% (Table 1), so this accuracy comes from the source: fusion at the boundary is complete.

Above the boundary, the target falls short. The shared levels above Z_{L_S} ($d < L_S$) also have a ceiling of 100%, yet the target recovers only part of it: 62% at Z_1 for $L_S = 2$, and 85% at Z_2 and 45% at Z_1 for $L_S = 3$. The shortfall grows with distance above the boundary. It is not explained by the target's own evidence: the prefix ceiling at Z_1 is about 25% for every L_S, and with no shared structure below the root ($L_S = 0$), the target reaches 95% of the ceiling at the same latent representation. In terms of the normalized recovery η (Appendix C), the target recovers 92% of the available information at Z_1 with no sharing, but only 49% at $L_S = 2$ and 26% at $L_S = 3$. Sharing more structure makes more information available at Z_1, and the target recovers less of it.

Redundant evidence turns a gain into a loss. The heatmap is the limit of maximal redundancy; we now introduce masking and vary retrievability ρ (Section 3.5), using the decoding readout at Z_1 for $L_S = 2$, one level above the boundary (Figure 2b,c). The source cannot attend to the target, so its accuracy reflects composition from its own masked view, and it is nearly the same at both values of ρ (60.2% and 57.4%). At $\rho = 0$, where the views' evidence is complementary, the target ends 19.5 points above the source. At $\rho = 1$, where it is aligned, the target ends 9.7 points *below* the source, despite having strictly more input. Since each view's masking rate is fixed, ρ changes only how the two views' evidence overlaps: when that evidence is aligned, the target's representation above the boundary is worse than what the source builds from its own view alone. Section 5 identifies the cross-view route through which the target reads the source, and how ρ shapes it.

5 The target develops a span-level retrieval route

Section 4 showed that the target recovers the fusion boundary Z_{L_S} completely but falls short above it, most when the two views' evidence is aligned. We now locate where the target reads the source. Using the metrics of Section 3.6, we measure for every head how much cross-view attention it sends ($A_{\ell h}$) and what fraction of it falls on the source span of the shared node ($\sigma_{\ell h}$), at $L_S = 1, 2, 3$, where this span covers $g = 8, 4, 2$ source leaves.

The route reads the shared node's span. The head with the highest span concentration places 97% of its cross-view attention on the four-leaf source span at $L_S = 2$, and 96% on the two-leaf span at $L_S = 3$. At $L_S = 1$ the head with the highest span concentration places 73% on the copy offset $\Delta - 1$, so at this depth the route acts more like an induction-style pointer than a span read (Figure 3a).

Retrievability strengthens and focuses the route. As ρ increases, span concentration above the null rises from 18% to 36% at $L_S = 1$, from 28% to 44% at $L_S = 2$, and from 26% to 45%

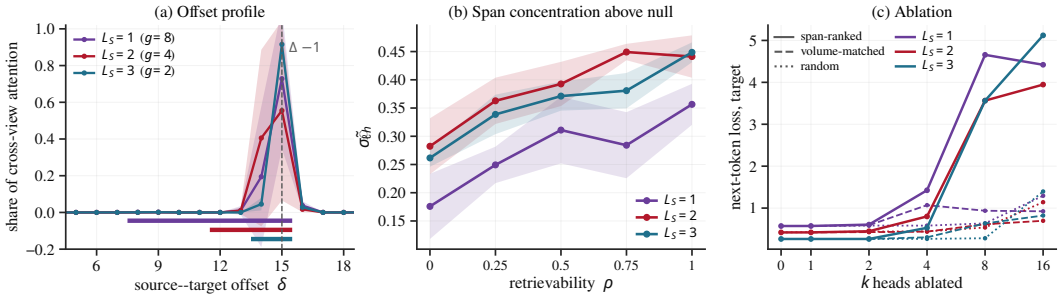

Figure 3: **The target develops a span-level retrieval route.** (a) Cross-view attention by source–target offset for the head with the highest span concentration at $\rho = 1$; bars mark the source span of the shared node. (b) Span concentration above the uniform-attention null, $\tilde{\sigma} = (\sigma - \sigma_0)/(1 - \sigma_0)$, against ρ, mass-weighted mean over heads. (c) Target next-token loss after blocking the cross-view attention of the top-k span-ranked heads at $\rho = 1$, against volume-matched and random heads.

at $L_S = 3$ (Figure 3b), and total cross-view attention grows as well. The more the two views' evidence is aligned, the more the target reads the source, and the more of that reading falls on the shared node's span.

The route is distributed and causally relevant. Blocking the cross-view attention of the single head with the highest span concentration has little effect at any shared depth (Figure 3c). The effect grows as more heads are blocked, and by $k = 8$ target next-token loss is five to six times higher than with the same number of volume-matched heads. The loss increase therefore depends on which source positions the blocked heads attend to, the span of the shared node, rather than on how much cross-view attention is removed. In the limit $L_S = L$, the span is a single token, the two views are token-identical, and the route reduces to token copying.

6 Source access suppresses composition, and removing it restores it

We next ask whether the weaker composition above Z_{L_S} is reversible. As in Figure 2a, we run these experiments *without leaf masking*, the limit of maximal redundancy that $\rho = 1$ approaches in Section 4, where copying is most reliable. Because the grammar is unambiguous, the decoding ceiling is then exactly 100% at every latent level and every L_S, so any shortfall is missing computation rather than missing information.

Severing source access at inference restores composition. Blocking the target-to-source attention submatrix at every layer, with weights and inputs unchanged, raises decoding at every latent level above Z_{L_S} and leaves it essentially unchanged at and below (Figure 4): the root rises from 68% to 98% at $L_S = 1$, and Z_1 from 85% to 96% at $L_S = 2$ and from 72% to 85% at $L_S = 3$. The circuit that composes these latent representations from the target alone is therefore present in the weights; in the intact forward pass, the retrieval route determines what is recoverable. The blocked pass is a counterfactual the model never ran in training, so we read it as evidence that the two circuits coexist, not that the representation was unused in the residual stream.

Severed versus retrained. A model trained from scratch without the route shows the same localization, but the two recoveries at the root differ, and their order flips with shared depth. At $L_S = 1$, the severed model far exceeds the retrained one at the root (98% vs. 69%); at $L_S = 2$ the two are close; at $L_S = 3$ the retrained model is ahead (37% vs. 46%). Joint training thus both scaffolds and starves target-side composition, with shared depth setting the balance. Joint training thus both scaffolds and starves target-side composition, with shared depth setting the balance. We hypothesize that this reflects reuse of the circuit that composes the target's $L - L_S$ private levels, which retrieval cannot replace: at $L_S = 1$ three private levels exercise it, and composition above Z_{L_S} comes nearly free once the route is removed; at $L_S = 3$ only one does, and the circuit above the boundary ends up weaker than in a model that built it alone. At fixed L, this cannot yet be separated from distance above the boundary, which varies with L_S in the same way.

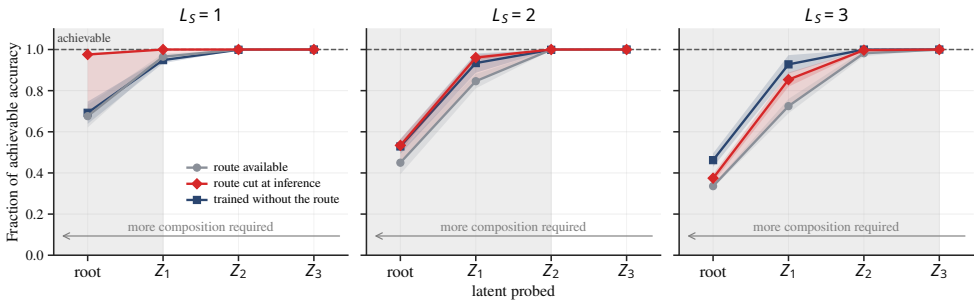

Figure 4: **Removing source access restores target-side composition.** Target-view decoding without leaf masking, where the exact ceiling is 100%. Grey: intact model; red: target-to-source attention blocked at inference, same weights and inputs; blue: trained with attention blocked throughout. Shading marks latent levels above Z_{L_S}. Five seeds, ± 1 s.d.

7 SIMILAR ROUTING COMPETITION IN A PRETRAINED MULTIMODAL MODEL

We next ask whether a similar pattern appears in a pretrained multimodal model, without retraining. We study LLaVA-1.5-7B (Liu et al., 2024), in which the image tokens precede the text: the answer position can attend to both the image and a caption, while the text cannot influence the image tokens. The model can therefore reach an answer about the image in two ways, from the image tokens or by reading a caption that states it. In the terms of Section 3, the object's side is a shared latent representation that the image realizes through many patch tokens, which must be composed, and the caption realizes as a single token, which can be read directly; this is the regime in which copying is cheapest. Unlike the synthetic setting, the model is pretrained, the variable takes only two values (left or right), and we make image and caption disagree, since only then does the answer reveal which route the model used. We ask which side a salient object is on in 400 balanced COCO (Lin et al., 2014) examples, with no caption, an agreeing caption, or a conflicting caption, and score the logits of the two side tokens (Appendix E, Table 2).

A conflicting caption overrides the image, and blocking it reverses this. Answers agree with the image on 72.0% of all 400 examples without a caption and on 100% with an agreeing one, but with a conflicting caption they follow the caption on 98.3%: the model can answer from the image, yet a disagreeing caption almost always wins. We rank heads by the attention the answer position sends to the caption's answer word on one half of the examples, and block that attention in the top-k heads on the other half. At $k = 512$, agreement with the image rises from 2.0% to 71.0%, close to the 73.0% the model reaches on the same examples without a caption; the same number of volume-matched heads, which attend to other caption tokens, gives 1.0%, and random heads give 12.5%. The recovery requires blocking half of the model's 1024 heads, so, as in the synthetic setting, the route is distributed. At this scale the intervention comes close to hiding the answer word, so it shows that the image-based answer returns once the caption's answer is unavailable, rather than which heads implement the route.

Downstream representations follow the caption. A probe trained on caption-free examples and applied, without retraining, to the conflicting-caption examples reads out the caption's side rather than the image's (Appendix E).

These results are qualitatively consistent with the synthetic pattern: when the prompt offers a direct textual route to the answer, both the answer and the downstream representation follow it, and removing access to it restores agreement with the image. Because the routes conflict and the intervention is coarse, we read this as consistent with, not a demonstration of, the synthetic mechanism.

8 CONCLUSION

A second view adds information, but shared semantics also open new routes to a latent representation, and which route the model uses changes what becomes recoverable within the target view. In our two-view extension of the RHM, with exact Bayes-optimal ceilings, a second view helps when its evidence is complementary. However, when its evidence is redundant, the target recovers the fusion boundary, while the levels above it become less recoverable. A distributed cross-view retrieval route reads the source span of the shared node, and severing it restores the missing composition. A pretrained multimodal model shows a similar competition: in LLaVA-1.5-7B, blocking attention to a conflicting caption restores image-based answers. These findings suggest treating fusion not only as combining information, but as a competition among routes to shared semantics: which route supplies a latent representation, and therefore what becomes recoverable within the target view, depends on the information each view provides and on how reliably the other view can be read.

9 LIMITATIONS

Our views are contiguous, fixed-length, and free of distractor tokens, with a known source span; real multimodal inputs have none of these properties, and the strength of copying may differ there. Our evidence that composition weakens is probe decodability: the interventions show that cross-view routes causally affect predictions, but we do not show that reduced decodability degrades output, and linear probes may miss information encoded non-linearly or in the source's format.

AI USE STATEMENT

Generative AI tools were used to assist with editing and polishing the language and presentation of this manuscript. All scientific content, experimental design, results, and conclusions are entirely the work of the authors. All co-authors have reviewed the final manuscript, take full responsibility for its content, and consent to its submission.

ETHICS STATEMENT

This work does not involve human-subject recruitment or the collection of new personal data. Our experiments primarily use synthetically generated data, with a supplementary evaluation based on existing COCO images and annotations. The main ethical consideration is the potential dual-use of our mechanistic findings: understanding cross-modal shortcuts and conflicting information could help improve model robustness, but could also inform attempts to manipulate multimodal model behavior.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of our experiments in Appendix F. This includes the complete synthetic-data construction and fixed random seeds, model architecture and parameter count, masking and retrievability procedures, optimization and training schedules, probing protocols, multi-seed evaluation, and compute requirements. The appendix also specifies the distinct constructions used for the main experiments, as well as the evaluation procedure for the multimodal experiment. All reported synthetic results use fixed data-generation procedures and five training seeds per configuration, with uncertainty reported across seeds. **To preserve double-blind anonymity, all code, data generation scripts, and configuration files will be publicly released upon publication.**

ACKNOWLEDGMENTS

We are grateful to a colleague for insightful discussions and constructive feedback that helped shape this work. We acknowledge access to high-performance computing resources, which were used to carry out the computational experiments reported in this work. Details are omitted for anonymous review.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. In *International Conference on Learning Representations*, 2017. URL https://mlanthology.org/iclr/2017/alain2017iclr-understanding/.

David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, 2004. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2004.01.029. URL https://www.sciencedirect.com/science/article/pii/S0960982204000430.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures. *arXiv preprint arXiv:2305.13673*, 2023.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1067–1077. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/allen-zhu24a.html.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *Advances in Neural Information Processing Systems*, 2024.

Francesco Cagnetta, Leonardo Petrini, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: the random hierarchy model. *Physical Review X*, 14(3):031001, 2024.

Francesco Cagnetta, Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Scaling laws and representation learning in simple hierarchical languages: Transformers versus convolutional architectures. *Phys. Rev. E*, 112:065312, Dec 2025. doi: 10.1103/qtd6-nl8p. URL https://link.aps.org/doi/10.1103/qtd6-nl8p.

N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956. doi: 10.1109/TIT.1956.1056813.

Francis B. Colavita. Human sensory dominance. *Perception & Psychophysics*, 16(2):409–412, 1974. doi: 10.3758/BF03203962.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002. doi: 10.1038/415429a.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9847–9857. Association for Computational Linguistics, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Asif A. Ghazanfar and Charles E. Schroeder. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6):278–285, 2006. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2006.04.008. URL https://www.sciencedirect.com/science/article/pii/S1364661306001045.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*, 2025.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

Konrad P. Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B. Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLOS ONE*, 2(9):1–10, 09 2007. doi: 10.1371/journal.pone.0000943. URL https://doi.org/10.1371/journal.pone.0000943.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

K. Lari and S.J. Young. Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech & Language*, 5(3):237–257, 1991. ISSN 0885-2308. doi: https://doi.org/10.1016/0885-2308(91)90009-F. URL https://www.sciencedirect.com/science/article/pii/088523089190009F.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023a.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *International Conference on Learning Representations*, 2023b. URL https://mlanthology.org/iclr/2023/li2023iclr-emergent/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023c.

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2019. doi: 10.1109/TKDE.2018.2872063.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10), June 2024. ISSN 0360-0300. doi: 10.1145/3656580. URL https://doi.org/10.1145/3656580.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024. doi: 10.1109/CVPR52733.2024.02484.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262133601.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. doi: 10.1038/264746a0.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. *arXiv preprint arXiv:2507.13868*, 2025.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In *Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022. doi: 10.1109/CVPR52688.2022.00806. URL https://mlanthology.org/cvpr/2022/peng2022cvpr-balanced/.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Tim Rohe and Uta Noppeney. Cortical hierarchies perform bayesian causal inference in multisensory perception. *PLOS Biology*, 13(2):1–18, 02 2015. doi: 10.1371/journal.pbio.1002073. URL https://doi.org/10.1371/journal.pbio.1002073.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12692–12702, 2020. doi: 10.1109/CVPR42600.2020.01271.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wu22d.html.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, pp. 33762–33775, 2025.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and C. L. Philip Chen. A review on multi-view learning. *Frontiers of Computer Science*, 19(7):197334, 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40004-w. URL https://doi.org/10.1007/s11704-024-40004-w.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: metrics and methods. In *International Conference on Learning Representations*, 2024.

A Discussion

Our results suggest that shared semantics change not only what information is available, but how that information is computed. At the fusion boundary, the source provides information beyond the target prefix, and the target recovers this shared representation essentially completely. Above the boundary, however, the source also provides an alternative route: rather than continuing to compose from target-side components, the model can retrieve the corresponding source representation. As the source evidence becomes more redundant and retrievable, this route increasingly suppresses target-side composition even though the target's own evidence is unchanged. Fusion is therefore not purely additive: access to another view can increase available information while changing which computation is used to obtain it.

Importantly, the resulting reduction in target-side composition does not imply that the computation has been lost. Severing source access at inference restores composition with the same weights and inputs, indicating that the model retains both routes and that changing the available pathway changes which route is exercised. In this sense, the shortcut we identify is better viewed as *representation reuse enabled by shared semantics* than as simple token copying: the source and target can realize the same latent representation differently, yet the target can retrieve the source realization instead of reconstructing it. The LLaVA experiment provides qualitative evidence for a similar competition outside the synthetic setting, although we regard it as consistent with, rather than a demonstration of, the synthetic mechanism.

Biological systems and the case for parallel routes. The tension we identify, between a shortcut that retrieves from another view and a circuit that computes from the target, has a familiar counterpart in multisensory perception. Humans combine cues in close agreement with reliability-weighted Bayesian integration (Ernst & Banks, 2002; Alais & Burr, 2004), but they do not always fuse. Körding et al. (2007) model perception as inferring whether two cues share a common cause and blending fusion and segregation accordingly, while Rohe & Noppeney (2015) report that different stages of the cortical hierarchy carry segregated, fused, and causal-inference-weighted estimates, suggesting that parallel computations can remain available. Unisensory processing also persists alongside multisensory integration, including in areas once thought to be purely unisensory (Ghazanfar & Schroeder, 2006), while visual dominance (Colavita, 1974) and the McGurk effect (McGurk & MacDonald, 1976) show that one route can nevertheless override another, much like the caption dominance we observe in LLaVA (Liu et al., 2024). This provides a useful perspective on our results: maintaining an alternative target-only route may be important when the cross-view route is unreliable, misleading, or unavailable. Our experiments suggest a possible computational reason for this: when a cross-view route provides a reliable substitute, it can reduce the pressure to exercise target-side composition even when that computation remains encoded in the weights. A natural architectural direction is therefore to preserve both fused and target-only pathways and arbitrate between them according to source reliability, for example through source dropout or a source-masked auxiliary loss with a learned gate. We regard this as a direction for future work, not a claim supported directly by our experiments.

B Future work: toward a multimodal grammar

In our framework both views are strings generated by the same kind of grammar, so the source realization of a shared node is a contiguous span of known length. Real image–caption pairs are not like this: the two views share semantics but realize them through different processes, one spatial and one sequential. A natural next step is a *vision PCFG* that keeps the RHM's latent class hierarchy but realizes it as an image rather than a sequence. A shared node would then no longer correspond to a clean, contiguous span in the source, which removes one of the main simplifications noted in our limitations. Studying such a grammar on its own first would show whether the composition we probe in strings carries over to spatial structure.

Pairing a vision grammar with a text grammar that shares its hierarchy down to depth L_S would give a fully multimodal version of our setting, with exact knowledge of what the two modalities share. This would let us ask whether the competition between copying and composition depends on the modality of the source, for instance whether copying is easier from text than from an image, as our

LLaVA experiment suggests. It would also let us study content that only one modality carries, and test whether a model still composes that content when the rest can be copied from the other view.

C EXACT BAYES-OPTIMAL CEILINGS

Because the complete generative process is known, we can compute the Bayes-optimal accuracy for predicting each target latent representation from different information sets. These ceilings provide reference points for distinguishing information that is already inferable from the target prefix from information that becomes available through the shared semantics.

For a target node $Z^B_{d,q}$, let $P_{d,q}$ be the target-view token prefix before its span-closing position, Y^A_{obs} the source sequence presented to the model (including any mask symbols), and $S = Z_{0:L_S}$ the shared latent classes at all positions and depths $0, \ldots, L_S$. The ceilings are Bayes accuracies under three information sets:

$$\text{BayesPrefix}(d,q) = \mathbb{E}\Big[\max_c P(Z^B_{d,q} = c \mid P_{d,q})\Big], \tag{3}$$

$$\text{BayesInput}(d,q) = \mathbb{E}\Big[\max_c P(Z^B_{d,q} = c \mid P_{d,q}, Y^A_{\text{obs}})\Big], \tag{4}$$

$$\text{BayesSharedOracle}(d,q) = \mathbb{E}\Big[\max_c P(Z^B_{d,q} = c \mid P_{d,q}, S)\Big]. \tag{5}$$

They are, respectively, the ceilings from the target prefix, the model's observed input, and oracle access to the shared latent representations (the prefix, shared, and oracle ceilings of Section 3.4).

For $d > L_S$, define the position of the target node's depth-L_S ancestor by

$$a(q) = \left\lfloor \frac{q}{b^{d-L_S}} \right\rfloor, \tag{6}$$

where b is the hierarchy's branching factor and positions are indexed from zero. Equivalently, the depth-d positions are partitioned into consecutive blocks of b^{d-L_S} descendants, and $a(q)$ is the index of the block containing position q. Write $F = Z_{L_S,a(q)}$. Given $(P_{d,q}, F)$, the target node is conditionally independent of the other shared latent representations: the private target expansion below F depends on no other shared node. Hence its posterior given $(P_{d,q}, S)$ equals its posterior given $(P_{d,q}, F)$. Thus, for nodes below the fusion boundary, the relevant frontier node suffices to compute the oracle ceiling.

Theorem 1 (Ceiling Ordering)**.** *For any two-view PCFG whose private expansions are conditionally independent given the shared latent representations S, and every target node (d, q),*

$$\text{BayesPrefix}(d,q) \le \text{BayesInput}(d,q) \le \text{BayesSharedOracle}(d,q). \tag{7}$$

Proof. Since $\sigma(P_{d,q}) \subseteq \sigma(P_{d,q}, Y^A_{\text{obs}})$, conditional Jensen for the convex function $p \mapsto \max_c p_c$ gives

$$\mathbb{E}_{Y^A_{\text{obs}} \mid P_{d,q}}\Big[\max_c P(Z^B_{d,q} = c \mid P_{d,q}, Y^A_{\text{obs}})\Big] \ge \max_c P(Z^B_{d,q} = c \mid P_{d,q}).$$

Taking expectation over $P_{d,q}$ proves the first inequality.

For the second inequality, the source and target private expansions are conditionally independent given the shared latent representations S. Therefore

$$P(Z^B_{d,q} = c \mid P_{d,q}, Y^A_{\text{obs}}) = \mathbb{E}_{S \mid P_{d,q}, Y^A_{\text{obs}}}\big[P(Z^B_{d,q} = c \mid P_{d,q}, S)\big].$$

Jensen applied to this mixture yields

$$\max_c P(Z^B_{d,q} = c \mid P_{d,q}, Y^A_{\text{obs}}) \le \mathbb{E}_{S \mid P_{d,q}, Y^A_{\text{obs}}}\Big[\max_c P(Z^B_{d,q} = c \mid P_{d,q}, S)\Big].$$

Averaging over $(P_{d,q}, Y^A_{\text{obs}})$ proves the second inequality. □

Corollary 1 (Fully Observed RHM)**.** *If the rules are unambiguous (Section 3), then for a fully observed (unmasked) source yield and every target node,*

$$\text{BayesInput}(d,q) = \text{BayesSharedOracle}(d,q). \tag{8}$$

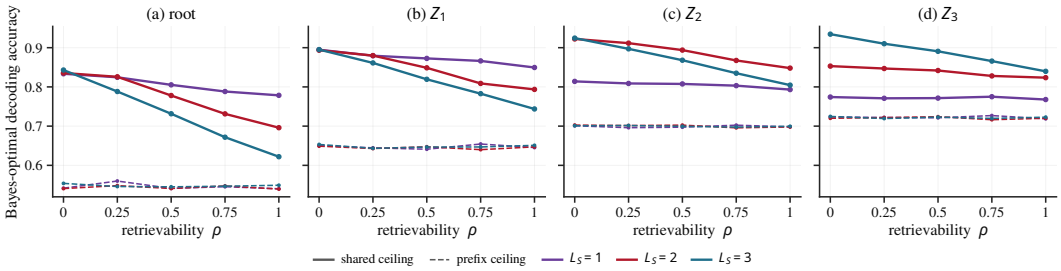

Figure 5: **Retrievability lowers the shared ceiling but not the prefix ceiling.** Bayes-optimal decoding accuracy under 50% leaf masking, against ρ, for each latent level and shared depth. Solid: shared ceiling; dashed: prefix ceiling. The prefix ceiling is invariant to ρ; the shared ceiling falls with ρ but remains above the prefix ceiling at $\rho = 1$.

Proof. Parse the source tree from leaves to root. The leaves are observed. If the child classes of a node are known, unambiguity identifies its parent class uniquely. Induction therefore recovers every source-side class, including all shared latent representations S, as a deterministic function of Y^A_{obs}. Hence $\sigma(P_{d,q}, S) \subseteq \sigma(P_{d,q}, Y^A_{\text{obs}})$, so refinement of information gives $\text{BayesInput} \geq \text{BayesSharedOracle}$. The reverse inequality is equation 7; therefore equality holds. With masking, bottom-up recovery can fail and the ceilings need not coincide. □

The extraction gap is the oracle ceiling minus the shared ceiling. The two gains over the prefix ceiling are

$$\Delta_{\text{observed}}(d, q) = \text{BayesInput}(d, q) - \text{BayesPrefix}(d, q), \tag{9}$$

$$\Delta_{\text{oracle}}(d, q) = \text{BayesSharedOracle}(d, q) - \text{BayesPrefix}(d, q), \tag{10}$$

with $\Delta_{\text{observed}} \leq \Delta_{\text{oracle}}$: the oracle gap is total shared-latent headroom, while the observed gap is the part available from the model's input. They coincide in the unmasked RHM by equation 8.

We compute each posterior exactly by belief propagation and estimate the expectations over $4{,}000$ sampled sequences, then average over the span-closing positions used by the probes to obtain $\text{BayesPrefix}(d)$, $\text{BayesInput}(d)$, and $\text{BayesSharedOracle}(d)$.

Separating information availability from representation recovery. The Bayes ceilings characterize available information; they do not measure how much the transformer represents. Let $\text{Acc}_B(d)$ be target probe accuracy at depth d. The gain over the prefix ceiling is

$$\text{gain}(d) = \text{Acc}_B(d) - \text{BayesPrefix}(d), \tag{11}$$

and normalized recovery of shared-latent headroom is

$$\eta(d) = \frac{\text{Acc}_B(d) - \text{BayesPrefix}(d)}{\text{BayesSharedOracle}(d) - \text{BayesPrefix}(d)} \quad \left(= \frac{\text{gain}(d)}{\Delta_{\text{oracle}}(d)} \right). \tag{12}$$

The probe state is a deterministic function of the tokens available at its position, hence of the model input $(P_{d,q}, Y^A_{\text{obs}})$. Therefore $\text{gain}(d) \leq \Delta_{\text{observed}}(d) \leq \Delta_{\text{oracle}}(d)$. Whenever $\text{Acc}_B(d) \geq \text{BayesPrefix}(d)$, $\eta(d)$ lower-bounds the fraction of model-accessible headroom recovered. When the source is unmasked, the two headrooms coincide and this bound is tight.

How retrievability acts on the ceilings. Figure 5 computes the decoding ceilings under the 50% leaf masking of Section 3.5. The prefix ceiling (BayesPrefix) conditions only on the target's own observed view, whose masking distribution does not depend on ρ, and is flat in ρ. The shared ceiling (BayesInput) also conditions on the observed source and falls with ρ, because aligned masks hide the same leaves in both views. It nevertheless stays above the prefix ceiling at $\rho = 1$ for every latent level and shared depth, so aligned evidence reduces what the source adds without removing it.

Table 1: **Absolute numbers behind Figure 2a.** Target prediction accuracy at the output layer, with two exact Bayes ceilings: Prefix ($\mathrm{BayesPrefix}$), what the target's own causal prefix determines, and Oracle ($\mathrm{BayesSharedOracle}$), what becomes available when the shared latent representations are known. η is the normalized recovery of shared-latent headroom (Appendix C). At $d \leq L_S$ the level is shared and its ceiling is exactly 1.0. * deepest shared latent representation; shaded rows lie above it. Mean $\pm$ s.d. over $n = 5$ seeds. The root is omitted, since the prediction readout has no target there.

L_S	Latent level	Accuracy	Prefix	Oracle	Acc/Oracle	η
0	Z_1	0.749 ± 0.017	0.256	0.791	95%	$+0.92$
	Z_2	0.587 ± 0.003	0.333	0.598	98%	$+0.96$
	Z_3	0.600 ± 0.008	0.530	0.621	97%	$+0.77$
1	$Z_1^\star$	1.000 ± 0.000	0.253	1.000	100%	$+1.00$
	Z_2	0.718 ± 0.005	0.344	0.726	99%	$+0.98$
	Z_3	0.662 ± 0.004	0.525	0.673	98%	$+0.93$
2	Z_1	0.615 ± 0.056	0.253	1.000	62%	$+0.49$
	$Z_2^\star$	1.000 ± 0.000	0.347	1.000	100%	$+1.00$
	Z_3	0.776 ± 0.003	0.505	0.776	100%	$+1.00$
3	Z_1	0.447 ± 0.030	0.253	1.000	45%	$+0.26$
	Z_2	0.845 ± 0.031	0.347	1.000	85%	$+0.76$
	$Z_3^\star$	1.000 ± 0.000	0.511	1.000	100%	$+1.00$

D Absolute results for shared-semantic prediction

Table 1 reports the absolute prediction accuracies and Bayes-optimal reference ceilings underlying Figure 2a. The table makes two points explicit. First, when no semantic structure except the root is shared ($L_S = 0$), the target already recovers most of the information available from the oracle ceiling. Second, as the fusion boundary moves toward the leaves, prediction remains near-perfect at and below Z_{L_S}, while accuracy at levels above Z_{L_S} declines even though the oracle ceiling increases to 1.0. The normalized recovery η captures the same effect: the fraction of the additional information supplied by the shared semantics that the model recovers decreases above the fusion boundary.

E Multimodal experiment details

We evaluate LLaVA-1.5-7B (Liu et al., 2024) on balanced image-question examples constructed around the horizontal position of a salient object.

Setup. We pair COCO (Lin et al., 2014) object annotations with images from LLaVA-Instruct-150K. For each image we select the largest annotated object whose centroid, after the model's resize and center-crop, lies clearly to one side ($|c_x - 0.5| > 0.15$ in normalized width) and which covers at least 1% of the image; images with no such object are skipped. We retain 400 examples, balanced between left and right positions.

Each example is presented in three conditions, which differ only in whether a `Context` line is present and what it states:

```
no caption
USER: <image>
Which side of the image is the dog on, left or right?
ASSISTANT:
```

```
agreeing caption (object is on the left)
USER: <image>
Context:  the dog is on the left.
Which side of the image is the dog on, left or right?
ASSISTANT:
```

```
conflicting caption
USER: <image>
Context:  the dog is on the right.
Which side of the image is the dog on, left or right?
ASSISTANT:
```

At the answer position we compare the logits of the two side tokens, so the evaluation involves no generation or string matching. We say the model *follows the image* when it answers the annotated side, and *follows the caption* when it answers the stated one.

Table 2: **LLaVA evaluation data.** Of the $63{,}664$ COCO-annotated images in LLaVA-Instruct-150K, 93.7% contain an object clearly on the left or right ($|c_x - 0.5| > 0.15$) covering at least 1% of the image. We take 400 of these in catalog order, balanced across sides, and present each under three conditions: no caption, agreeing caption, and conflicting caption. Heads are ranked on one half of the examples and the intervention is evaluated on the other.

COCO-annotated images in LLaVA-Instruct-150K	63,664
with a qualifying object	93.7%
Examples taken	400
left / right	200 / 200
distinct images / object categories	400 / 64
most frequent category (*person*)	145
Prompts (3 conditions per example)	1,200
conflicting-caption	400
Head-ranking / evaluation examples	200 / 200

Intervention. For each head we measure the attention weight from the answer position to the caption's answer word, and rank heads by that weight. Ranking uses one half of the examples, and the intervention is evaluated on the disjoint other half, so heads are never selected on the examples they are scored on. For the top k heads we block attention to that word's key positions from every later query, leaving the prompt, the weights, and all other attention unchanged. Two controls use the same k: *volume-matched* heads carry a comparable amount of caption attention directed elsewhere, which separates the location of the attention from its magnitude, and *random* heads are sampled uniformly. Table 3 reports the resulting agreement with the image.

Table 3: **Blocking attention to the caption's answer word.** Agreement with the image on the evaluation half under a conflicting caption, after blocking attention to the caption's answer word in k heads: the top-k heads ranked on the other half, volume-matched heads that attend to other caption tokens, or random heads.

	Agreement with image (%)		
Heads blocked k	Top-k	Volume-matched	Random
0	2.0	2.0	2.0
16	2.0	3.0	2.0
64	16.0	2.0	1.5
512	**71.0**	1.0	12.5
No caption (image only): 73.0			

The image-only accuracy on the same evaluation set is 73.0%, so the largest intervention recovers 97.3% of the image-only accuracy. The volume-matched control removes a comparable amount of caption attention without targeting the selected heads, and does not produce the same recovery. The

model has 1024 heads, and substantial recovery requires blocking roughly half of them, consistent with a distributed route.

Probe. We fit a logistic probe on caption-free examples to predict the annotated side from the residual stream. We then apply it, frozen, to the agreeing and conflicting conditions at the same layer and position. Accuracy is always measured against the image, so a probe that has come to carry the caption's answer scores below chance under conflict. At the answer position, agreement with the image is 76.7% on caption-free examples and 8.0% under conflict; at the object token, the corresponding accuracies are 70.7% and 19.7%. Across all layers and read positions, the highest conflict accuracy is 58.0%. We report the full layer profile rather than a single layer, since the effect varies with depth. The image-side representations are unchanged across the no-caption, agreeing-caption, and conflicting-caption conditions, consistent with the causal ordering of the image and text streams.

F REPRODUCIBILITY

We provide the following details to make the synthetic experiments and analyses reproducible. The setup of the pretrained multimodal experiment, including the data construction, prompting, head-ranking procedure, intervention, controls, and probing protocol, is described separately in Appendix E.

Grammar. We use one Random Hierarchy Model throughout the synthetic experiments: depth $L = 4$, branching factor $b = 2$, $v = 8$ features, 8 root classes, and per-level synonym schedule $m = (4, 4, 2, 2)$ from the root down. Each view therefore contains $b^L = 16$ leaves, and the token vocabulary contains 9 symbols (eight content symbols and one mask symbol). The rule table is sampled once with seed_rules $= 0$ and the sample pool once with `seed_samples` $= 42$; these are then held fixed across runs. We generate $100{,}000$ training sequences and $10{,}000$ evaluation sequences.

Two views. The source and target views are concatenated without a separator, with the source first, giving a sequence of 32 tokens and a target block beginning at position 16. The views share levels $0, \ldots, L_S$. Below L_S, the target re-expands the shared frontier using the same rule table with an independent synonym draw (the shared-rule construction), generated with `seed_samples2` $=$ 1042. The heatmap in Section 4 uses a distinct split-rule construction: the target re-expands with a second independently sampled rule table, using `seed_rules2` $= 1$, while levels at and above L_S are copied from the first view. This construction is also the one that permits $L_S = 0$, in which case the views share only the root class. The headline pattern replicates under the shared-rule construction with masking: from $\rho = 0$ to $\rho = 1$, the fraction of the shared ceiling recovered falls above the boundary, changes by at most 3.1 points at private levels, and is unchanged at $L_S = 1$, which has no shared level above the boundary.

Masking and retrievability. In the retrievability experiments, each leaf is kept with probability 0.5 and masked otherwise (code parameter `fusion_rate`, the keep probability). For each leaf, with probability ρ, the target reuses the source's masking decision; otherwise it draws its masking decision independently. We use $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$. The severance experiments in Section 6 use unmasked inputs (`fusion_rate` $= 1.0$).

Model. We train an 8-layer pre-norm GPT-2-style transformer from scratch, with hidden dimension 128, 4 attention heads, feed-forward dimension 512, GELU activations, and RoPE positional embeddings. Dropout is 0.1 on the residual stream, embeddings, and attention; layer-normalization uses $\epsilon = 10^{-5}$, and parameters are initialized with standard deviation 0.02. The output head is untied from the input embedding. The model has $1{,}588{,}744$ parameters.

Optimization. We use AdamW with learning rate 10^{-3}, weight decay 0.01, cosine learning-rate decay with 500 warmup steps, gradient clipping at 1.0, batch size 64, no gradient accumulation, and mixed precision. The objective is next-token cross-entropy. The retrievability and severance experiments are trained for $200{,}000$ steps; the split-rule experiments used for the heatmap in Section 4 are trained for $100{,}000$ steps.

Probes. Probing is performed as a separate offline procedure over saved checkpoints. For each checkpoint, we fit one multinomial logistic probe per layer using $4{,}000$ sequences and 300 Adam steps at learning rate 0.05, with no weight decay. Features are standardized using statistics from the training split. The probe data are split sample-disjointly $80/20$, with the split fixed using seed 1234 so that probe curves are comparable across checkpoints. Every experimental arm fits its own probe, including the severed-at-inference arm, whose probe is fit on the blocked forward pass. The multimodal experiment is the exception: as described in Section 7, its probe is trained on caption-free examples and then frozen when applied to the other conditions.

Random seeds and uncertainty. We use five training seeds per configuration. A training seed determines model initialization and batch order; the grammar and sample pool remain fixed. Thus, reported variability reflects training stochasticity rather than variation in the generated grammar or dataset. Error bars denote ± 1 standard deviation across the five seeds. For severed-versus-intact comparisons, differences are computed within seed before averaging because the two conditions use the same weights and inputs.

Compute. Each training run uses a single NVIDIA A100 GPU.

G Extended related work

This appendix expands the summary of related work in Section 2.

Multiple views: when fusion helps and when it hurts. Multi-view and multimodal learning rest on the premise that views of one event carry complementary information, so fusing them should be no worse than the best view alone (Baltrušaitis et al., 2019; Li et al., 2019; Yu et al., 2024). Theory supports the premise under conditions: fusion lowers population risk when the shared latent representation is estimated more accurately from two views than from either (Huang et al., 2021), and information-bottleneck theory provides a general language for minimal sufficient representations (Tishby et al., 1999), which multi-view information-bottleneck work applies to multiple views (Federici et al., 2020). In practice the premise routinely fails, and the existing explanations are largely optimization-side: one modality learns faster (Wang et al., 2020), dominates the training signal (Wu et al., 2022), or leaves the other under-trained (Peng et al., 2022), with the two streams' representations geometrically separated as a related symptom (Liang et al., 2022). Those accounts explain *which* view the model uses. We hold each view's marginal difficulty fixed and vary two things separately: how deep the semantic structure the views share runs, L_S, and how retrievable one view's evidence is in the other, ρ, which is a dose–response knob on the redundancy of the two views' evidence rather than on its quantity. We then ask what the target view still computes once the source has made that computation unnecessary.

Alignment, versus the route that uses what is aligned. Multimodal practice pairs the two streams with an alignment objective, such as contrastive image–text pretraining (Radford et al., 2021) or a learned interface between a frozen encoder and the language model (Li et al., 2023a), on the assumption that the views should be brought into a common representation. The platonic-representation hypothesis states this as convergence on a shared abstraction (Huh et al., 2024), and an information-theoretic reading of the same premise asks that representation to be minimal and sufficient for the task (Federici et al., 2020). Our interventions separate having a shared representation from using one. Blocking target-to-source attention removes the only route through which any source-side representation could reach the target, and it raises target decoding above the fusion boundary both at inference, with the weights and the inputs unchanged, and during training, where the composition is then built from the start (Section 6). What those interventions change is which route supplies the information, not how similar the two views' representations are.

Hierarchical structure, with exactly computable reference points. The Random Hierarchy Model and its descendants generate data from an ensemble of PCFGs whose latent levels are explicit, and use that explicitness to predict sample complexity and learning time for deep networks (Cagnetta et al., 2024; Cagnetta & Wyart, 2024); the "Physics of Language Models" line trains on controlled synthetic hierarchies in the same ceiling-anchored spirit (Allen-Zhu & Li, 2023; 2024),

with PCFGs and their inference procedures as the classical substrate (Chomsky, 1956; Manning & Schütze, 1999; Lari & Young, 1991). A separate literature measures compositional generalization with benchmarks that hold the parts fixed and vary their recombination (Lake & Baroni, 2018; Kim & Linzen, 2020; Hupkes et al., 2020), and finds that transformers often reach such tasks by matching subgraphs rather than by composing systematically (Dziri et al., 2023). We use the RHM's explicit latent levels in a two-view construction, but vary the depth L_S to which the views share a hierarchy rather than hierarchy depth or rule complexity within a single stream. Because the generative process is known, every readout is scored against exact Bayes-optimal ceilings, which matters because probe accuracy alone cannot separate information a representation carries from information the probe itself supplies (Alain & Bengio, 2017; Belinkov, 2022; Hewitt & Liang, 2019; Voita & Titov, 2020; Pimentel et al., 2020); the ceiling is what turns "the probe did not find it" into a statement about the model.

Shortcuts, and the price of a cheaper route. That learned models prefer predictive structure that is cheap to compute is well established: reliance on surface heuristics (McCoy et al., 2019), features that are predictive in-distribution but not in the way the task requires (Geirhos et al., 2020), gradient starvation, where one easy feature saturates the loss and harder ones stop receiving signal (Pezeshki et al., 2021), and a simplicity bias strong enough to discard a more predictive but harder feature that is present in the data (Shah et al., 2020). Our finding sits in this family and we do not claim the phenomenon as new; what differs is the object. The computation the target stops performing above Z_{L_S} is not a spurious correlate but the composition of levels that *both* views carry, so the source holds the same content the target would have derived, and the objective is indifferent between the two routes: given the shared latent representation, the levels above it add exactly nothing to next-token prediction. The shortcut is created by the shared semantics of the two views rather than by an artifact of collection, and exact ceilings let us say which computation the model declined, not merely that it failed.

Circuits, copying and retrieval heads. Mechanistic work identifies copying and retrieval as concrete implementations: induction heads complete repeated patterns (Olsson et al., 2022), a single head in GPT-2 Small suppresses naive copying and thereby repairs other ablations (McDougall et al., 2023), a sparse set of retrieval heads carries information from arbitrary context positions and pruning them causes hallucination (Wu et al., 2025), and named circuits implement indirect object identification (Wang et al., 2023) and factual recall (Meng et al., 2022), with activation patching as the standard causal tool (Zhang & Nanda, 2024). Two properties separate the route we find. It is cross-view rather than within-stream, and it is not token copying: below Z_{L_S} the two views realize the same latent representation differently, so what a head can read is a span whose content must be decoded rather than copied (Figure 3), and only in the degenerate token-identical case does the source token at the copy offset $\Delta - 1$, which always lies in S, equal the target's next token. Because $\Delta - 1$ lies in S, and at $\rho = 1$ the two views' masks are aligned there, part of the span concentration may reflect an induction-style offset rather than a read of the whole span; separating the two is left to future work. As in this literature, the route is distributed: lesioning the single head with the highest span concentration has no effect (Section 5). One known reason a single-head lesion can be null is self-repair by other components (McDougall et al., 2023). Because ablating a component mid-pass takes the model off the distribution it was trained on, we read these lesions as localization rather than as a decomposition of the computation (Meng et al., 2022; Wang et al., 2023; Zhang & Nanda, 2024), and the claim they carry is cross-checked against two manipulations that are not inference-time lesions of a trained model: retrievability, which changes the data alone (Section 5), and severing the route from the first training step (Section 6).

Conflicting sources. Which of two conflicting sources a model trusts has been studied for context against parametric memory (Longpre et al., 2021; Xie et al., 2024) and, in vision-language models, between image and caption, where the two directions of cross-modal influence are markedly asymmetric (Frank et al., 2021), the favored modality is visible in representational structure and can be shifted by intervening on a small set of heads (Hua et al., 2025; Ortu et al., 2025), and language priors contribute to object hallucination (Li et al., 2023c). Our LLaVA results reproduce that asymmetry and add the causal step this literature usually omits: blocking attention to the one caption token that carries the answer restores agreement with the image to 97.3% of the model's own image-only accuracy, while volume-matched and random-head controls do not. The normative

counterpart is cue integration, where each cue is weighted by its reliability (Ernst & Banks, 2002); what we document is not misweighting but the replacement of one computation by another, and its reversal when the alternative route is cut.