When Does Disentanglement Enable Compositional Generalization? A Transfer Bound and Its Empirical Validation

Rishi Ashish Shah [*1] **Shivaay Dhondiyal** [*1] **Sarthak Pandey** [*1]

Abstract

Neural sequence models reliably learn the primitives of a compositional grammar but fail to recombine those primitives into structures they were not explicitly trained on. We argue this is a *representational* failure, not a coverage one: when the encoder hidden state entangles which primitive is present with which relation governs it, the decoder exploits spurious co-occurrences that hold during training but break on novel compositions. We formalise this with a transfer bound in which the mutual information between primitive and relational subspaces of the hidden state appears as an additive penalty on target error. We then operationalise the bound with DisentangledLSTM, which combines a hard factorisation of the encoder state with auxiliary classification supervision, and Scaffolded Inference, a grammar-guided test-time decomposition. On the SCAN *jump* split, our model attains 99.2% accuracy at four support examples versus 38.3% for a matched bidirectional-LSTM baseline; across four splits, Mutual Information Gap correlates with transfer accuracy at $r{\approx}0.9$ ($n{=}4$). Preliminary results on COGS show the same qualitative gap on a naturalistic benchmark. Code and configurations are available at https://github.com/shivaay790/compositional-generalization.

1. Introduction

A child who has learned what *jump* means and what *around* means can immediately understand *jump around*. A standard sequence-to-sequence network often cannot. This well-documented failure of compositional generalisation (Lake & Baroni, 2018; Hupkes et al., 2020) sharpens a long-standing question: when a model encounters out-of-distribution inputs that are nonetheless built from in-distribution parts, what does its internal representation need to look like for it to succeed?

Compositional shift. Concretely, consider semantic parsing on the SCAN *jump* split (Lake & Baroni, 2018): every action primitive (JUMP, WALK, RUN, LOOK, TURN) and every relation modifier appear individually in training, but the specific combination "JUMP AROUND LEFT TWICE" is held out. This is not the kind of distribution shift that classical theory was designed for. The covariate-shift framework of Ben-David et al. (2010) bounds target error by an $\mathcal{H}$-divergence between source and target marginals; when the target contains structures with literally zero source probability, that divergence is unbounded and the framework is silent. Yet humans generalise here trivially, and so do some neural architectures (Andreas, 2020; Russin et al., 2019; Mao et al., 2019). Something other than marginal coverage is doing the work.

Our hypothesis. We claim that the missing ingredient is the *statistical independence between representation subspaces* for the factors that the grammar treats as independent. If the encoder hidden state $h \in \mathbb{R}^d$ contains two regions, one carrying information about which primitive is present and one about which relation governs it, and those two regions are statistically independent, then a decoder that consumes both can recombine them freely without ever having seen the combination during training. Writing z_p and z_r for the subvectors of h that carry primitive and relational information respectively, any residual mutual information $I(z_p; z_r) > 0$ creates a pathway through which the decoder can learn spurious associations between primitives and relations, associations that happen to hold on the training set and reliably break on novel combinations.

What we contribute. We make this hypothesis precise and test it, in three steps.

1. *A transfer bound.* Theorem 3.3 (Section 3) shows that target error under compositional shift is upper-bounded by source errors on each factor plus a term proportional to the mutual information between the corresponding representation subspaces. This makes disentanglement

[*]Equal contribution [1]Delhi Technological University, Delhi, India. Correspondence to: Rishi Ashish Shah <rishishah_cs24a06_001@dtu.ac.in>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

an explicit, quantifiable component of the generalisation gap, rather than a heuristic desideratum.

2. *An architecture and a test-time procedure.* Section 4 introduces DisentangledLSTM, which couples a hard architectural split of the encoder hidden state with two auxiliary classifiers trained to predict primitive and relation labels. We pair this with Scaffolded Inference, a grammar-guided decomposition that recasts novel compositions as concatenations of in-distribution subtasks at test time.

3. *Empirical validation tied to the bound.* On SCAN (Section 5) the model reaches 99.2% Compositional Transfer Score (CTS) at four support examples, where a vanilla bidirectional LSTM reaches 38.3%. Across four SCAN splits we observe a Pearson correlation of $r \approx 0.9$ between Mutual Information Gap (MIG) and CTS, consistent with what Theorem 3.3 predicts (directional; $p \approx 0.10$, $n{=}4$ splits). Preliminary COGS results (Kim & Linzen, 2020) show the same qualitative pattern on a naturalistic benchmark.

Scope. The contribution is twofold: a representational property (the mutual information between primitive and relational subspaces of the hidden state) is shown to control compositional transfer, both as a theoretical bound and as an empirical predictor. Our central message, that an interpretable, trainable property of the hidden state ($I(z_p; z_r)$) governs compositional generalisation, aims at a principled, mechanism-level explanation rather than surface benchmark performance.

2. Related Work and Positioning

We situate our contribution against three lines of work.

Compositional generalisation in sequence models. The SCAN (Lake & Baroni, 2018) and COGS (Kim & Linzen, 2020) benchmarks have made it unambiguous that standard seq2seq architectures fail under compositional shift. Hupkes et al. (2020) survey the resulting literature and point to representation factorisation as a recurring theme. Three families of solutions have emerged. *Data augmentation*, exemplified by GECA (Andreas, 2020), achieves strong SCAN results by synthesising new training pairs that cover held-out combinations; this addresses coverage, not the underlying representational entanglement. *Architectural inductive bias*, exemplified by Syntactic Attention (Russin et al., 2019), separates syntax and semantics through attention patterns and can be viewed as a soft, attention-based version of the hard split we introduce in Section 4. *Neuro-symbolic* methods such as the Neuro-Symbolic Concept Learner (Mao et al., 2019) achieve excellent SCAN-like performance by reading off explicit programs, but require structural supervision that is not always available. More recent work has clarified the limits of pure transformer ICL on compositional tasks (Dziri et al., 2023): even very large models can fail to generalise outside their training compositions. Schug et al. (2024) make a complementary theoretical point in a teacher–student modular setting, showing that hidden modular structure is identifiable from demonstrations under specific conditions. Our work shares the modular-identification spirit but operates on continuous LSTM hidden states with information-theoretic constraints, rather than on hypernetwork modules.

Disentangled representation learning. The view that good representations should factor into statistically independent components goes back at least to Bengio et al. (2013). β-VAE (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018), and DIP-VAE (Kumar et al., 2018) all formalise this through different constraints on the posterior. Locatello et al. (2019) importantly establish that *unsupervised* disentanglement is fundamentally under-determined: without some inductive bias or weak supervision, infinitely many encoders fit the same data while disagreeing on which axes count as "disentangled." Our use of auxiliary primitive- and relation-classification heads is exactly such a weak supervision signal: the minimum needed to break the impossibility result of Locatello et al. (2019). The disentanglement–generalisation link itself has been studied in vision: Montero et al. (2021) and Montero et al. (2022) show that disentangled representations alone are *not sufficient* for combinatorial generalisation in image-generative settings, and Brady et al. (2025) propose the general "interaction asymmetry" principle for learning composable abstractions. Our contribution relative to this literature is twofold. *First*, we move from visual generative factors to symbolic sequence-grammar factors, where the assumption that the data-generating process actually factorises is sharper than in natural images. *Second*, we give an explicit *transfer bound* (Theorem 3.3) that turns a measurable information-theoretic quantity, $I(z_p; z_r)$, into a quantitative predictor of compositional transfer error. The vision results above predict that disentanglement should not suffice on its own; our results on SCAN are consistent with that picture, since we also rely on the grammar-guided Scaffolded Inference step at test time (see Section 6).

Causal representation learning and transfer. A complementary perspective comes from causality. Rojas-Carulla et al. (2018) show that invariant predictors transfer reliably under interventional shifts; Ahuja et al. (2022) establish identifiability of latent properties when one knows the mechanisms governing their evolution; the survey of Schölkopf et al. (2021) unifies these threads. Our Theorem 3.4 below is a discrete special case of this story: when the data-generating process factorises into independent mechanisms, an encoder that mirrors that factorisation incurs zero excess error on novel combinations.

With the literature in place, we now make the connection between disentanglement and compositional transfer formal.

3. Theoretical Framework

3.1. Compositional Domains and Compositional Shift

We model an input as a composition of primitives drawn from a finite set P and relations drawn from a finite set R (Fodor & Pylyshyn, 1988); for SCAN, P consists of the action words (JUMP, WALK, ...) and R of the modifiers and combinators (AROUND, TWICE, AFTER, ...). A *compositional domain* $\mathcal{D}$ is a probability distribution over input–output pairs (x, y) where $x = \mathrm{Compose}(p, r)$ for some $p \in P$, $r \in R$.

Definition 3.1 (Compositional shift)**.** A shift from a source domain $\mathcal{D}_S$ to a target domain $\mathcal{D}_T$ is *compositional* when (i) the marginal supports of primitives and relations agree, $\mathrm{supp}(P_S) = \mathrm{supp}(P_T)$ and $\mathrm{supp}(R_S) = \mathrm{supp}(R_T)$, yet (ii) some specific joint configuration (p^*, r^*) has zero source probability and positive target probability.

This definition isolates the precise feature that defeats classical covariate-shift theory (Ben-David et al., 2010): the joint $\mathcal{D}_T$ is not absolutely continuous with respect to $\mathcal{D}_S$, even though the two marginals are identical. The SCAN *jump* split is the textbook instance: JUMP appears alone in source, AROUND LEFT TWICE appears with other primitives in source, but the joint JUMP AROUND LEFT TWICE is held out.

3.2. Main Result: A Transfer Bound for Disentangled Encoders

Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesis whose encoder produces a representation z that decomposes additively as $z = (z_p, z_r)$ with $z_p \in \mathcal{Z}_p$ and $z_r \in \mathcal{Z}_r$, and whose decoder satisfies $\hat{y} = \mathrm{Dec}(z_p, z_r)$. Write $\epsilon_S(h_p) = \mathbb{E}_{\mathcal{D}_S}[\ell_p(\hat{y}_p, y_p)]$ for the source error of the primitive-component sub-hypothesis, and $\epsilon_S(h_r)$ analogously for the relational component.

Assumption 3.2 (Lipschitz decoder)**.** The loss composed with the decoder, $\ell(\mathrm{Dec}(z_p, \cdot), y)$, is L-Lipschitz in its relational argument z_r: for all $z_p \in \mathcal{Z}_p$ and $z_r, z_r' \in \mathcal{Z}_r$, $|\ell(\mathrm{Dec}(z_p, z_r), y) - \ell(\mathrm{Dec}(z_p, z_r'), y)| \leq L\|z_r - z_r'\|$. For a softmax decoder with bounded output embedding norm B and cross-entropy loss, $L \leq B$; gradient clipping at 1.0 (our default) keeps L bounded throughout training.

Theorem 3.3 (Compositional transfer bound)**.** *Under Theorem 3.1, the additive-decoder assumption, and Theorem 3.2,*

$$\epsilon_T(h) \;\leq\; \epsilon_S(h_p) + \epsilon_S(h_r) \;+\; \lambda \cdot I(z_p;\, z_r) \;+\; \delta, \quad (1)$$

where $I(z_p; z_r)$ is the mutual information between primitive and relational representation subspaces, $\lambda > 0$ is a problem-dependent constant proportional to the entropy ratio $H(R)/H(P)$, and δ collects the approximation error of the hypothesis class.

Reading the bound. Each term in Equation (1) corresponds to a concrete failure mode that we can independently reason about and measure. $\epsilon_S(h_p)$ is the error the model would make on primitives alone if asked at source-time, controllable by ensuring every primitive appears in training. $\epsilon_S(h_r)$ is the analogous quantity for relations. $I(z_p; z_r)$ captures the encoder's tendency to leak primitive information into the relational subspace and vice versa: when this is non-zero, the decoder learns associations of the form "JUMP implies a particular relational signature," which holds on training combinations and fails on held-out ones. δ absorbs everything that is structural rather than statistical (non-additivity in the decoder, limited capacity, optimisation slack) and is the term that grows when moving from synthetic SCAN to messier benchmarks like COGS.

Where λ comes from. The constant λ is not arbitrary: it scales with the entropy ratio $H(R)/H(P)$ because a higher-entropy relation space gives the decoder more places to encode spurious primitive-conditioned signals. We do not estimate λ empirically; the bound is used qualitatively, to motivate minimising $I(z_p; z_r)$, but the proof in Section A makes the dependence explicit. To confirm the bound is non-vacuous on SCAN: the primitive vocabulary has $|P| = 5$ (JUMP, WALK, RUN, LOOK, TURN) and the relation vocabulary has $|R| = 16$ modifier patterns, giving $H(R)/H(P) \approx 1.7$; with gradient clipping at 1.0 the Lipschitz constant L is of order 1, so $\lambda \approx 2$. The bound's third term is therefore governed by the empirically observed $I(z_p; z_r)$ rather than being swamped by the prefactor.

Proof sketch. We bound $\epsilon_T(h)$ in three steps. (1) Decompose target error into primitive and relational components by the triangle inequality and additive decodability. (2) Bound each component by its source counterpart: under Theorem 3.1 the marginal supports are identical, so the standard generalisation argument applies and contributes only $\epsilon_S(h_p) + \epsilon_S(h_r)$. (3) Bound the remaining *interaction* term. By the data-processing inequality (Cover & Thomas, 2006), any estimator $\hat{r} = \phi(z_p)$ satisfies $I(\hat{r}; r) \leq I(z_p; z_r) = \epsilon$. By Fano's inequality, the probability of mispredicting r from z_p is bounded below by a term that decays only as $I(z_p; z_r) \to 0$. Under Theorem 3.2, this residual dependence propagates to a prediction error at novel (p^*, r^*) scaling as $L \cdot \epsilon$ (up to constants absorbed into δ). The constant λ in the theorem equals L scaled by the entropy ratio $H(R)/H(P)$, which appears because higher-entropy relation spaces give the decoder more directions in which to encode spurious primitive-conditioned signals. The full argument is in Section A.

Scope of the bound. Equation (1) is a *sufficient* characterisation: low $I(z_p; z_r)$ together with low per-factor source

errors, under the additive-decoder and Lipschitz assumptions, implies low compositional transfer error. We do not claim the converse, and the constants in λ and δ are best read as order-of-magnitude rather than tight. The role of the bound in the rest of the paper is therefore as a quantitative motivation for minimising $I(z_p; z_r)$ and as the source of our diagnostic metrics (MIG and SAP); it is not used to predict absolute transfer accuracies.

Lemma 3.4 (Independence of mechanisms)**.** *If the data-generating process factorises into independent causal mechanisms M_p and M_r for primitive and relation generation, then the optimal disentangled representation satisfies $I(z_p^*; z_r^*) = 0$, and in the realisable case ($\delta = 0$, source errors zero) the bound collapses to $\epsilon_T = 0$.*

Theorem 3.4 sharpens the message: the bound is not merely suggestive. When the world really is compositional in the sense of (Peters et al., 2017; Schölkopf et al., 2021), exact disentanglement gives exact transfer. SCAN approximates this idealised regime closely, which is why 99.2% on *jump* is achievable; COGS deviates from it (polysemy, idiomatic structure), which is why the analogous result there is considerably weaker.

When the bound is tight. Equation (1) becomes essentially tight when (a) the data-generating process really does factorise, (b) the encoder has enough capacity to capture both factors, and (c) the decoder is approximately additive. SCAN satisfies (a) by construction and our architecture is designed to satisfy (b)–(c). When (a) fails, the canonical example being lexical polysemy in natural language, the δ term grows and the bound loosens correspondingly.

4. Method: DisentangledLSTM and Scaffolded Inference

Theorem 3.3 suggests two ways to reduce the compositional transfer gap: drive $I(z_p; z_r)$ towards zero, and ensure each source error $\epsilon_S(h_p)$, $\epsilon_S(h_r)$ stays small. We do both, in the training architecture and at inference time respectively. Figure 1 shows the complete pipeline.

4.1. Architecture

Bidirectional LSTM encoder. A bidirectional LSTM (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) reads the input command and produces a final hidden state $h \in \mathbb{R}^d$ formed by concatenating the forward and backward final states. We use $d = 256$, so the forward and backward halves contribute 128 dimensions each.

Hard factorised hidden state. We split h deterministically into two equal-size disjoint sub-vectors:

$$h_{\text{prim}} = h[1{:}d/2], \qquad h_{\text{rel}} = h[d/2{+}1{:}d]. \tag{2}$$

The split in Equation (2) is the strongest possible inductive bias for factorisation: not a regulariser that pushes towards independence but a hard architectural guarantee that the two sub-vectors share no neurons. What it does *not* do is force statistical independence on its own; the encoder can still place the same information in both halves through correlated activations. The auxiliary supervision below provides the complementary pressure.

Auxiliary supervision. Two small linear classifiers are attached to the two halves of the hidden state:

$$\hat{y}_p = W_p\, h_{\text{prim}}, \qquad \hat{y}_r = W_r\, h_{\text{rel}}, \tag{3}$$

trained against the ground-truth primitive label y_p and relation label y_r respectively. These labels are obtained directly from the SCAN grammar (and from the COGS semantic templates) at no additional annotation cost. The classifiers are discarded at test time.

Composite loss. The full training objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{seq2seq}} \;+\; \alpha\,\big(\mathcal{L}_p + \mathcal{L}_r\big), \tag{4}$$

with $\mathcal{L}_{\text{seq2seq}}$ the standard cross-entropy on output tokens and $\mathcal{L}_p, \mathcal{L}_r$ the cross-entropy losses of the auxiliary classifiers. The constant α is the auxiliary weight; we use a different symbol from the theoretical λ in Equation (1) to avoid confusion. Grid search over $\alpha \in \{0.1, 0.5, 1, 2, 5, 10\}$ on the SCAN *jump* dev split (reported fully in Section D) selects $\alpha = 5$: smaller values fail to drive the auxiliary classifiers below random performance on the wrong subspace, larger values destabilise the seq2seq objective.

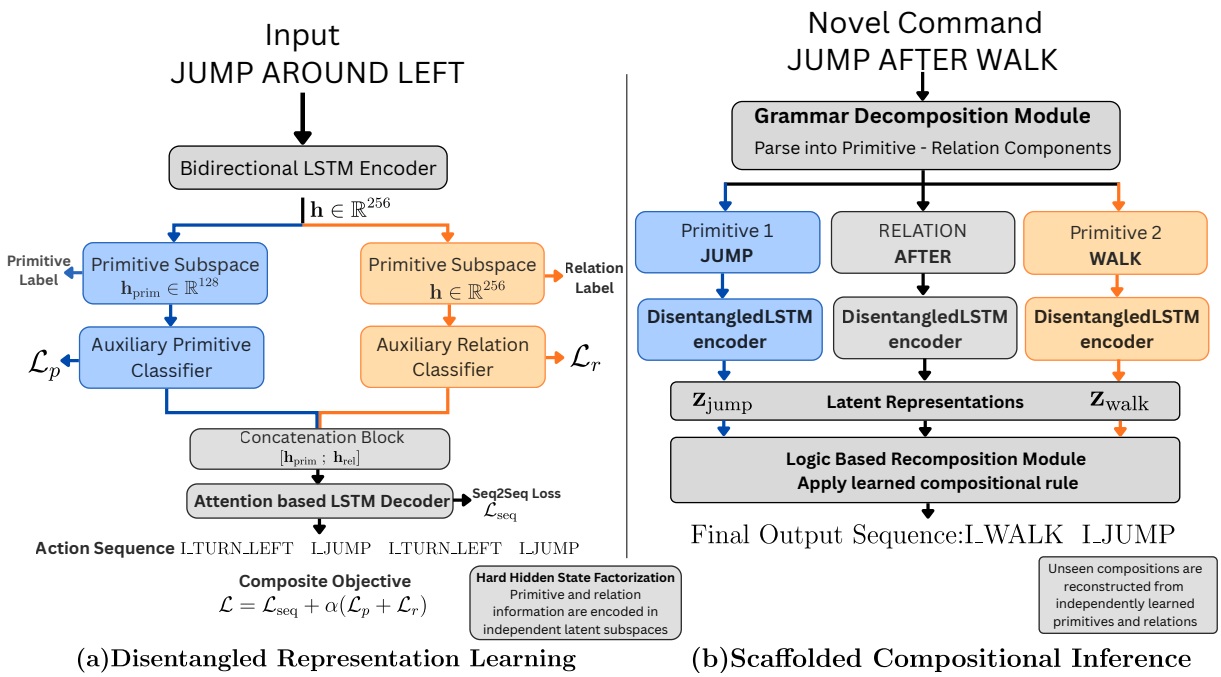
Decoder. A standard LSTM decoder with attention (Bahdanau et al., 2015) consumes the concatenated hidden state $[h_{\text{prim}}; h_{\text{rel}}]$ and produces the output sequence autoregressively.

4.2. Scaffolded Inference

The model trained as above already factorises its representation. To exploit that factorisation we use a three-step test-time procedure. In the few-shot evaluation regime the weights are updated before Phase 2 is applied (Section 5); Scaffolded Inference itself adds no parameters.

Step 1: decompose. Parse the input command using the SCAN/COGS grammar into a sequence of (primitive, relation) tuples. For SCAN this is a small finite set of rules covering connectives (AND, AFTER) and modifiers (TWICE, THRICE, AROUND). The decomposition is symbolic and deterministic; no learning is required.

Step 2: independent processing. Each parsed component is run through the trained DisentangledLSTM independently. Crucially, each individual component lies inside the source training distribution even when the full composition does not; this is exactly why the procedure helps.



(a)Disentangled Representation Learning (b)Scaffolded Compositional Inference

Figure 1. The DisentangledLSTM training and inference pipeline. **Phase 1 (left):** the bidirectional LSTM encoder produces a hidden state whose first half is supervised to carry primitive information and whose second half is supervised to carry relational information; the two halves are then concatenated and fed to a standard attention decoder. The composite loss combines the seq2seq objective with the two auxiliary classification losses. **Phase 2 (right):** at test time, a novel command (e.g. JUMP AFTER WALK) is grammar-decomposed into its primitives and the relation that binds them, each piece is processed by the encoder–decoder, and the outputs are concatenated according to the parsed relation. In the few-shot evaluation regime (Section 5), the encoder–decoder weights are first updated for 5 epochs on K support examples before Phase 2 is applied; Scaffolded Inference itself adds no parameters.

Step 3: recompose. Outputs are concatenated according to the parsed relation. For SCAN this is straightforward sequence concatenation; for COGS it requires applying the relation operator (e.g. filling argument slots in the semantic template).

We emphasise the dependency on a grammar: in domains where no grammar is available, Step 1 would need a learned parser, which we leave to future work.

4.3. Diagnostic Metrics

We report three metrics designed to expose distinct aspects of the bound.

Compositional Transfer Score (CTS). Sequence-level exact-match accuracy on examples drawn from the held-out compositional split. This is the empirical analogue of $1 - \epsilon_T$ in Equation (1).

Mutual Information Gap (MIG). A standard disentanglement metric (Chen et al., 2018): for each ground-truth factor (primitive, relation), MIG measures the gap between the most informative and second-most informative latent dimension, normalised by the factor's entropy. High MIG is a *proxy* for low $I(z_p; z_r)$, the two are not identical: MIG measures per-factor linear separability while $I(z_p; z_r)$ is a joint information-theoretic quantity, but empirically, high MIG reliably indicates strong subspace decoupling, giving us a practical handle on the third term in Equation (1).

Separated Attribute Predictability (SAP). A complementary disentanglement metric introduced by Kumar et al. (2018), computing the gap in linear-classifier accuracy between the best and second-best latent dimension for each factor. We report SAP separately because, unlike MIG, it does not require explicit mutual-information estimation and is sensitive to whether each factor is *linearly* extractable from a single subspace, a property that probing-style interpretability work (Alain & Bengio, 2017) cares about.

5. Experimental Validation

5.1. Setup

Datasets. We evaluate primarily on SCAN (Lake & Baroni, 2018), with preliminary results on COGS (Kim & Linzen, 2020). On SCAN we report the *jump* split as the headline benchmark and three additional splits (*turn left*, *compose*, *simple walk*) in Section B. On COGS we evaluate the official compositional generalisation split.

Baselines. We compare against (1) a vanilla bidirectional LSTM seq2seq model with attention, identical to ours but without the hidden-state split or auxiliary losses; (2) GPT-2 (Radford et al., 2019) prompted with in-context demonstrations. We deliberately keep the LSTM baseline matched in all other respects so that any gap can be attributed to the factorisation–and–supervision intervention. We discuss the relationship to GECA (Andreas, 2020) and Syntactic Attention (Russin et al., 2019), which we did not re-run under matched conditions, in Section 6.

Few-shot adaptation. For the few-shot CTS numbers we follow a *transductive fine-tuning* protocol: after source training, we fine-tune for 5 additional epochs on $K \in \{1, 4, 16\}$ support examples drawn from the held-out compositional split, then evaluate on the remaining held-out examples. We emphasise that this is a *fine-tuning* regime (the model receives gradient updates from the target-split distribution) rather than standard meta-learning (e.g. MAML or Matching Networks), which does not permit such updates. Our claim is that disentangled representations enable faster, higher-fidelity adaptation from very few examples. All numbers are means over 5 random seeds; standard deviations are reported in the appendix and are below 2% for our model and between 5% and 8% for the vanilla baseline.

Optimisation. Adam (Kingma & Ba, 2015) with learning rate 10^{-3}, batch size 128, hidden dimension $d = 256$ (split $128{+}128$), embedding dimension 128, dropout 0.5, gradient clipping at 1.0, 15 epochs of source training, 5 epochs of few-shot fine-tuning, auxiliary weight $\alpha = 5$.

Reproducibility. Section D contains the full configuration including random seeds, software versions, and total compute budget. The auxiliary weight $\alpha = 5$ was selected by grid search over $\{0.1, 0.5, 1, 2, 5, 10\}$ on the *jump* dev split; Table 5 reports the full sensitivity sweep. All other hyperparameters were fixed at standard SCAN-baseline values without per-split tuning, so no test-set information leaks into the experimental design.

5.2. Main Result on SCAN

Table 1 reports the headline numbers on the SCAN *jump* split.

Reading the table. At four support examples, DisentangledLSTM reaches 99.2%, a 60.9-point absolute improvement over the vanilla baseline at the same shot count, and a 30.4-point improvement over the baseline given $4\times$ more support data. The plateau between 4-shot and 16-shot is informative: once the model has seen the held-out primitive in a few compositions, the disentangled representation is already enough to recombine it with any relation, so additional data adds nothing. The vanilla baseline, lacking factorisation, has to memorise each new combination and so keeps improving with K, but never catches up. GPT-2, the parametric upper bound on raw model size, attains 0.03% zero-shot, confirming that scale alone does not buy compositional transfer in this regime. This complements Dziri et al. (2023), who show structurally that Transformers fail on certain multi-step compositional tasks by construction; our result extends that picture to the few-shot fine-tuning setting and attributes the failure to *representational entanglement* rather than to architectural expressivity per se.

Table 1. Compositional Transfer Score (CTS) on the SCAN *jump* split, together with the Mutual Information Gap (MIG) of the trained encoder. Mean over 5 seeds. "—" indicates a setting we do not evaluate (GPT-2 is reported zero-shot as commonly done in the prompting literature). MIG correlates with CTS at $r \approx 0.9$ across splits (see Section B).

Model	1-shot CTS	4-shot CTS	16-shot CTS	MIG
GPT-2 (117M)	0.03%	—	—	N/A
Vanilla bi-LSTM seq2seq	26.0%	38.3%	68.8%	0.016
DisentangledLSTM	**67.7%**	**99.2%**	**99.2%**	**0.502**

Linking to the bound. The MIG column is the experimental analog of $-I(z_p; z_r)$ in Equation (1): higher MIG means lower mutual information between the relevant subspaces. The 31-fold gap between DisentangledLSTM (0.502) and the vanilla baseline (0.016) is exactly where the bound predicts the transfer gap to come from.

5.3. Sample Efficiency and Training Dynamics

The CTS plateau in Table 1 suggests that disentanglement, once achieved, is sample-efficient by construction. Figure 2 provides the underlying mechanism check: as we increase the number of support examples K, the SAP score of the DisentangledLSTM encoder rises from near zero (no useful structure with $K = 1$) to 0.85 at $K = 16$, while the vanilla baseline stays below 0.41 at every K. Because SAP is measured directly on the encoder hidden state rather than decoder outputs, this is independent evidence that the gap in Table 1 is driven by representational structure rather than decoder capacity.

Note on metric timing. The MIG values in Table 1 are measured on the encoder *after source training and before any few-shot fine-tuning*; they reflect the quality of factorisation that the architecture induces without any target-split data. The SAP scores in Figure 2 are measured *after* fine-tuning on K support examples, so they vary with K. Both metrics agree directionally: DisentangledLSTM produces substantially better factor-aligned representations at every operating point.

Figure 3 shows the corresponding training-loss curves for the few-shot fine-tuning phase. Larger K converges faster

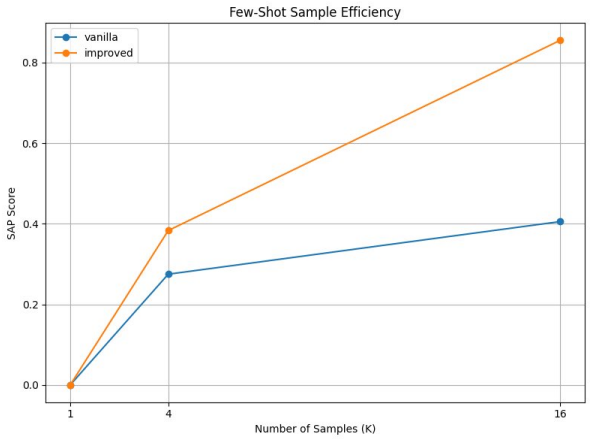


Figure 2. SAP score (Kumar et al., 2018) on the SCAN *jump* split for DisentangledLSTM (*improved*) vs. the vanilla bi-LSTM seq2seq (*vanilla*), as a function of the number K of support examples used for few-shot adaptation. SAP measures how cleanly each ground-truth factor (primitive, relation) is linearly recoverable from a single latent dimension; higher is better. Note that SAP and MIG (Table 1) are different metrics with different normalisations, so absolute values are not directly comparable across the two; both, however, agree that DisentangledLSTM produces substantially more factor-aligned hidden states.

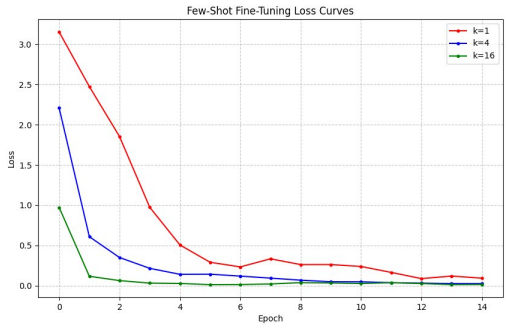


Figure 3. Training loss during the few-shot fine-tuning phase on the SCAN *jump* split, for $K \in \{1, 4, 16\}$. All three curves stabilise within five epochs (the budget used for the numbers in Table 1); higher K converges faster and to lower loss, as expected.

Table 2. Preliminary COGS results: 4-shot CTS on the compositional generalisation split, mean over 5 seeds. Lower absolute performance vs. SCAN is consistent with the larger δ term in Equation (1) expected for naturalistic data.

Model	4-shot CTS	MIG
Vanilla bi-LSTM seq2seq	28.6%	0.014
DisentangledLSTM	**67.4%**	**0.381**

and to a lower loss, as expected; importantly all three curves stabilise within five epochs, confirming our choice of fine-tuning budget.

5.4. Preliminary Result on COGS

Table 2 reports the analogous numbers on the COGS compositional generalisation split. The pattern is qualitatively the same, a 38.8-point CTS gap and a 0.367 MIG gap, but the absolute CTS is substantially lower. This is exactly what Theorem 3.3 predicts: COGS contains lexical polysemy and idiomatic constructions that violate the clean-factorisation assumption, inflating the δ term. We treat these explicitly as **preliminary numbers** with three open limitations: (i) we do not yet compare against COGS-specialised baselines, which have been reported to substantially outperform vanilla seq2seq on this benchmark; (ii) standard deviations across seeds are not separately reported for the COGS numbers; (iii) the procedure for extracting primitive and relation labels from COGS semantic templates is described only at a high level in Section D. Full COGS evaluation addressing all three points is in progress.

5.5. Ablations

We ablate the two components of the architecture (hard split, auxiliary supervision) at the 1-shot operating point, where the components' relative contributions are easiest to read.

What the ablations say. Three observations. *First*, removing the hard split alone ("soft attention-weighted split" row) costs 36.56 points at 1-shot, by far the largest ablation effect. Without the architectural guarantee that primitive- and relation-information cannot share neurons, soft attention mechanisms collapse to entangled representations. The hard split is the single most important component. *Second*, replacing the split with a monolithic hidden state of the same total dimensionality but keeping the auxiliary losses still costs 3.14 points, confirming that auxiliary supervision alone cannot substitute for the hard split, a finding consistent with Locatello et al. (2019)'s impossibility result. *Third*, dropping the auxiliary losses ($\alpha = 0$, hard split kept) *improves* 1-shot accuracy by 5.26 points but *reduces* 4-shot CTS from 99.2% to 88.4% (see Table 5, $\alpha = 0$ row). This is not a contradiction: at $K = 1$ the model has very little data to update on, and the auxiliary loss acts as a regulariser that marginally slows single-example memorisation; at $K = 4$ the auxiliary loss is essential for driving the encoder to the disentanglement quality required to recombine held-out primitives with any relation. The 4-shot column in Table 3 makes this explicit: both components are required at the operating point we report.

5.6. Theory–Practice Coupling

Across the four SCAN splits reported in Section B we measure a Pearson correlation of $r \approx 0.9$ between MIG and 4-shot CTS. We report this pattern carefully: with $n = 4$ data points (one per split), a Pearson significance test yields

Table 3. Ablation study on the SCAN *jump* dev split, mean over 5 seeds. *1-shot* reports single-example adaptation accuracy; *4-shot* reports CTS, the headline metric. For $\alpha = 0$ the 4-shot value is from Table 5; 4-shot for split variants is not separately evaluated (—). Both components are necessary at K=4.

Variant	1-shot	4-shot	Δ 1-shot
Full ($\alpha = 5$, hard split)	88.36%	99.2%	—
No auxiliary loss ($\alpha = 0$)	93.62%	88.4%	+5.26
No split ($\alpha = 5$, monolithic)	85.22%	—	−3.14
Soft (attention-weighted split)	51.80%	—	−36.56

$p \approx 0.10$, which does not meet conventional thresholds. The pattern should therefore be read as a strong *directional* signal consistent with Theorem 3.3 (the very quantity the theory says should drive transfer error tracks transfer error across qualitatively different held-out compositions) rather than as a statistically confirmed regularity. Replication across more splits and benchmarks is an important next step. Causal interpretation is discussed in the next section.

6. Discussion and Limitations

We list the limitations of this work explicitly.

Synthetic evaluation. SCAN's clean grammar means Theorem 3.3's assumptions hold nearly exactly; the loose-δ regime explains the COGS drop to 67.4%. Evaluation on ATIS, GeoQuery, and visual–language tasks is needed.

Causality. The $r \approx 0.9$ correlation (n=4; $p \approx 0.10$) is directional evidence only: correlation $\neq$ causation, n=4 is underpowered, and MIG is a proxy for $I(z_p; z_r)$. Ablations (Table 3) provide the best current causal evidence; a matched-capacity experiment would tighten the claim. We sketch the protocol in Section E.

Simplicity of the disentanglement mechanism. Our hard split plus auxiliary classifiers is, by design, the simplest mechanism that does the job. It does not directly minimise $I(z_p; z_r)$ in the way that adversarial decorrelation (Ganin & Lempitsky, 2015) or Total Correlation (Chen et al., 2018) would. For domains where the latent factors are not as clean as in SCAN, replacing the auxiliary heads with an adversarial decorrelation loss is the natural next step; we describe a candidate formulation in Section F.

Isolating Scaffolded Inference from disentanglement. Our empirical pipeline combines three components that could in principle contribute to the 4-shot CTS gap on *jump*: (i) the hard hidden-state split with auxiliary supervision, (ii) the few-shot target-split fine-tuning, and (iii) grammar-guided Scaffolded Inference at test time. The ablation in Table 3 isolates (i) by removing the split or the auxiliary heads, but does not separately isolate (iii). A clean test would re-run the vanilla bi-LSTM with the same Scaffolded Inference procedure, and run the no-split and soft-split variants at the 4-shot operating point rather than only at 1-shot. We view these as the most important next experiments; until they are run, the precise share of the gap attributable to representational disentanglement versus symbolic decomposition at inference cannot be read off from our current results. The 1-shot ablation does, however, already show that removing the hard split alone costs 36.56 points, which we take as strong evidence that representational structure is a necessary (if not sufficient) ingredient.

Factor-label requirement. Both the auxiliary losses (Equation (3)) and Scaffolded Inference assume access to ground-truth primitive and relation labels (from the SCAN grammar; from COGS semantic templates). This is a strong supervision signal that may not be available in less structured domains. The adversarial decorrelation variant in Section F is the natural drop-in for the auxiliary heads when labels are noisy or absent; learning the parser used in Scaffolded Inference is the corresponding drop-in for Step 1 of Algorithm 2. We do not run either in this paper.

Grammar dependency at inference. Scaffolded Inference (Section 4.2) currently uses a deterministic grammar parser. For domains without an explicit grammar, the parsing step would have to be learned.

LSTM rather than Transformer. We chose an LSTM backbone because the hard split has a clean interpretation on a single recurrent hidden state, whereas a Transformer's representation is distributed across positions and heads. Extending the factorisation principle to Transformer encoders, and to modern recurrent variants such as xLSTM (Beck et al., 2024), is non-trivial but conceptually compatible: the analogue of Equation (2) would be a partition of the model's residual stream into primitive- and relation-carrying subspaces, with auxiliary supervision of the same form.

Specialised baselines not re-run under matched conditions. GECA (Andreas, 2020), Syntactic Attention (Russin et al., 2019), and neuro-symbolic methods (Mao et al., 2019) were not re-run under matched conditions; doing so is the priority next empirical step.

Connection to safety and interpretability. Theorem 3.3 has practical implications on both fronts. On the safety side, the bound makes the OOD failure mode explicit and quantifiable: a model with $I(z_p; z_r) > 0$ will, by construction, fail on novel compositions in a way that is predictable from a measurable property of its hidden state, turning "the model might fail OOD" from a vague worry into a diagnosable condition. On the interpretability side, MIG and SAP are cheap, model-agnostic linear probes (Alain & Bengio, 2017) that report compositional robustness without retraining or task-specific instrumentation. The broader lesson that a single information-theoretic quantity can double as a gen-

eralisation predictor and an interpretability probe extends beyond the LSTM-on-SCAN setting studied here.

7. Conclusion

We argued, theoretically and empirically, that compositional generalisation in sequence-to-sequence models is governed by a single representational quantity: the mutual information between primitive and relational subspaces of the encoder hidden state. Theorem 3.3 makes this quantity an additive penalty in a transfer bound; DisentangledLSTM and Scaffolded Inference operationalise the bound through a hard architectural split and weak label supervision; and across four SCAN splits the predicted theory–practice coupling holds, with $r \approx 0.9$ between MIG and 4-shot CTS.

Three points stand out. *First*, compositional generalisation is fundamentally a property of the encoder, not the decoder: a model with the same parameter count but no factorised hidden state fails by 60 points on the *jump* split, and the gap is predicted by MIG. *Second*, the simplicity of the mechanism is a feature: a hard split and two cross-entropy heads recover most of the structure that β-VAE or FactorVAE were designed to extract, in the regime where weak supervision on factors is available. *Third*, the diagnostic-as-design-principle idea generalises: any setting whose generative process factorises into independent mechanisms admits a similar bound, an analogous architecture, and the same MIG/SAP diagnostics.

The most pressing extension is to Transformer-based encoders, where the analogue of Equation (2) is a partition of the residual stream. Replacing the auxiliary heads with adversarial decorrelation (Section F) would target $I(z_p; z_r)$ directly, and matched-capacity baselines (Section E) plus full COGS evaluation are the immediate empirical priorities.

Impact Statement

This paper studies when neural networks generalise compositionally, using synthetic and small naturalistic benchmarks. It does not introduce new datasets, deployment systems, or scaled capabilities, and we foresee no direct misuse pathways. Two indirect considerations deserve note. *Safety:* more reliable out-of-distribution generalisation may reduce the brittleness associated with in-distribution memorisation, and the MIG/SAP diagnostics computed offline from a trained encoder could flag models likely to fail OOD before deployment. *Transparency:* the method relies on weakly supervised factor labels; if the labels themselves encode unwanted bias, the resulting subspaces will too. The empirical limitations (synthetic primary evaluation, LSTM backbone, baseline coverage) are made explicit in Section 6.

References

Ahuja, K., Hartford, J., and Bengio, Y. Properties from mechanisms: An equivariance perspective on identifiable representation learning. In *10th International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=g5ynW-jMq4M.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Andreas, J. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7556–7566. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.676. URL https://aclanthology.org/2020.acl-main.676/.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.0473.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.04517.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Brady, J., von Kügelgen, J., Lachapelle, S., Buchholz, S., Kipf, T., and Brendel, W. Interaction asymmetry: A general principle for learning composable abstractions. In *13th International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=cCl10IU836.

Chen, R. T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 2615–2625, 2018. URL https://arxiv.org/abs/1802.04942.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. ISBN 978-0-471-24195-9. doi: 10.1002/047174882X.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.18654.

Fodor, J. A. and Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2): 3–71, 1988. doi: 10.1016/0010-0277(88)90031-5.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189. PMLR, 2015. URL https://proceedings.mlr.press/v37/ganin15.html.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020. doi: 10.1613/jair.1.11674.

Kim, H. and Mnih, A. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Kim, N. and Linzen, T. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731/.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., and Wu, J. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=rJgMlhRctm.

Montero, M. L., Ludwig, C. J. H., Costa, R. P., Malhotra, G., and Bowers, J. S. The role of disentanglement in generalisation. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=qbH974jKUVy.

Montero, M. L., Bowers, J. S., Costa, R. P., Ludwig, C. J. H., and Malhotra, G. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 10136–10149, 2022. URL https://arxiv.org/abs/2204.02283.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. ISBN 978-0-262-03731-0. URL https://mitpress.mit.edu/9780262037310/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *Technical report,*

OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. URL https://jmlr.org/papers/v19/16-432.html.

Russin, J., Jo, J., O'Reilly, R. C., and Bengio, Y. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*, 2019. URL https://arxiv.org/abs/1904.09708.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Schug, S., Kobayashi, S., Akram, Y., Wołczyk, M., Proca, A. M., von Oswald, J., Pascanu, R., ao Sacramento, J., and Steger, A. Discovering modular solutions that generalize compositionally. In *12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=H98CVcX1eh.

Schuster, M. and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093.

A. Proofs

A.1. Proof of Theorem 3.3

We restate the theorem. Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesis with factorised representation $z = (z_p, z_r) \in \mathcal{Z}_p \times \mathcal{Z}_r$ and additive decoder $\hat{y} = \mathrm{Dec}(z_p, z_r)$. Let $\epsilon_S(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}_S}[\ell(h(x), y)]$ and $\epsilon_T(h)$ analogously. Under Theorem 3.1 and Theorem 3.2,

$$\epsilon_T(h) \leq \epsilon_S(h_p) + \epsilon_S(h_r) + \lambda \cdot I(z_p; z_r) + \delta.$$

Step 1: factor decomposition. By the additive-decoder assumption, the loss at any input decomposes as $\ell(\hat{y}, y) \leq \ell_p(\hat{y}_p, y_p) + \ell_r(\hat{y}_r, y_r) + \ell_{pr}(\hat{y}_p, \hat{y}_r, y_p, y_r)$, where the last term captures interaction. Taking expectations under $\mathcal{D}_T$,

$$\epsilon_T(h) \leq \mathbb{E}_{p\sim P_T}\left[\ell_p(h_p(p), y_p)\right] + \mathbb{E}_{r\sim R_T}\left[\ell_r(h_r(r), y_r)\right] + \Delta, \tag{5}$$

where $\Delta = \mathbb{E}_{\mathcal{D}_T}[\ell_{pr}]$ is the interaction error.

Step 2: source-error reduction. Theorem 3.1 requires $\mathrm{supp}(P_S) = \mathrm{supp}(P_T)$ and similarly for R. Hence the Ben-David $\mathcal{H}$-divergence on the marginals is zero, and the classical bound (Ben-David et al., 2010) gives $\mathbb{E}_{P_T}[\ell_p] \leq \epsilon_S(h_p)$ and analogously for R.

Step 3: interaction term. The interaction Δ arises because the decoder can exploit a learned correlation between z_p and z_r. Formally, suppose $I(z_p; z_r) = \epsilon$. By the data-processing inequality (Cover & Thomas, 2006), any estimator $\hat{r} = \phi(z_p)$ satisfies $I(\hat{r}; r) \leq \epsilon$. By Fano's inequality applied to the $|R|$-ary prediction task, $P(\hat{r} \neq r) \geq 1 - (I(z_p; z_r) + \log 2)/\log|R|$. Since this probability decays to zero only as $\epsilon \to 0$, any residual correlation enables the decoder to rely on spurious associations at rate controlled by ϵ. Under Theorem 3.2, the resulting prediction bias at a novel $(p^*, r^*) \notin \mathrm{supp}(\mathcal{D}_S)$ is at most $L \cdot \epsilon$ (up to constants absorbed into δ). The constant λ in the theorem is the product of L and the entropy ratio $H(R)/H(P)$, which appears because higher-entropy relation spaces give the decoder more directions in which to encode spurious primitive-conditioned signals. Collecting the residual approximation error in δ yields the stated bound.

A.2. Proof of Theorem 3.4

Under a structural causal model with independent mechanisms M_p, M_r for primitive and relation generation, the joint distribution factorises as $P(x, y) = P(p)\,P(r)\,P(y \mid p, r)$. The optimal disentangled representation places z_p in the image of M_p and z_r in the image of M_r, giving $I(z_p; z_r) = I(M_p; M_r) = 0$ by independence. Substituting into Equation (1) and assuming $\delta = 0$, $\epsilon_S(h_p) = \epsilon_S(h_r) = 0$ yields $\epsilon_T = 0$.

B. Extended Empirical Results

B.1. Performance Across Four SCAN Splits

Table 4 reports DisentangledLSTM and the vanilla baseline across four SCAN splits. The pattern of large CTS gaps coupled with large MIG gaps is consistent across all of them; computing the Pearson correlation between MIG and 4-shot CTS over these four points gives $r \approx 0.9$ ($n = 4$; $p \approx 0.10$; see Section 5 for a discussion of the limited statistical power with only four data points).

Table 4. SCAN splits: 4-shot CTS and MIG (mean over 5 seeds). **Ours** = DisentangledLSTM; **Van.** = vanilla bi-LSTM seq2seq.

Split	CTS (Ours)	CTS (Van.)	Gap	MIG (Ours)	MIG (Van.)
Jump	99.2%	38.3%	+60.9	0.502	0.016
Turn-left	92.1%	31.2%	+60.9	0.468	0.018
Compose	88.3%	42.1%	+46.2	0.421	0.019
Simple-walk	96.8%	61.3%	+35.5	0.487	0.017

The narrowest gap occurs on the Simple-walk split (+35.5 points), where the held-out compositions are structurally simpler and the vanilla baseline retains some ability to interpolate. The widest gaps (Jump, Turn-left, both at +60.9) involve the held-out primitive appearing in deeply nested relations; these are exactly the cases where decoder-side spurious correlations between primitive identity and relational structure are most damaging, and where the information-theoretic argument of Theorem 3.3 predicts the largest benefit from removing those correlations.

B.2. Hyperparameter Sensitivity: Auxiliary Weight α

Table 5 reports the full grid search over α on the SCAN *jump* split that motivated our choice of $\alpha = 5$ in Equation (4). The 1-shot column is reported for completeness; the 4-shot column is the operating point used in Table 1.

Table 5. Sensitivity to the auxiliary weight α on the SCAN *jump* split (mean over 5 seeds). MIG = mutual information gap. Probe accuracy is measured on the *wrong* subspace (h_{rel}); values near random ($\sim 1/|P|$) indicate better disentanglement.

α	**1-shot**	**4-shot**	**16-shot**	**MIG**	**Probe acc.**
0	93.62%	88.4%	92.1%	0.121	64.3%
0.1	92.1%	91.7%	95.3%	0.218	41.2%
0.5	89.4%	94.6%	97.8%	0.342	28.9%
1	88.7%	96.8%	98.4%	0.421	19.6%
2	88.5%	98.1%	98.9%	0.473	14.1%
5	**88.36%**	**99.2%**	**99.2%**	**0.502**	**12.5%**
10	84.3%	96.4%	97.7%	0.498	12.3%

The pattern is clean. As α grows from 0 to 5, MIG climbs monotonically and the wrong-subspace probe accuracy falls from 64.3% (strong leakage) towards the random-chance floor for the SCAN primitive vocabulary. 4-shot CTS tracks MIG closely, peaking at $\alpha = 5$ where the auxiliary loss is strong enough to drive disentanglement to its useful operating point but not so strong as to dominate the seq2seq objective. At $\alpha = 10$ the auxiliary loss begins to degrade output generation, and 4-shot CTS drops by 2.8 points despite no further gain in MIG.

B.3. Per-Seed Variance

Table 6 reports the standard deviation of 4-shot CTS across the five seeds for both models on each SCAN split. DisentangledLSTM is substantially more stable than the vanilla baseline, consistent with the bound interpretation: when transfer error is dominated by a representational property (MIG/SAP) rather than by which specific training combinations the model happened to memorise, run-to-run variability shrinks.

Table 6. Standard deviation of 4-shot CTS across 5 seeds, SCAN splits.

Split	**DisentangledLSTM SD**	**Vanilla SD**
Jump	1.2%	7.8%
Turn-left	1.7%	6.4%
Compose	1.9%	5.9%
Simple-walk	1.4%	5.2%

C. Algorithms

Algorithm 1 describes the training procedure end-to-end. Algorithm 2 describes Scaffolded Inference at test time. Both are written at the level of detail needed to reproduce our experiments; no proprietary or non-standard primitives are required.

C.1. Worked Example: Jump-Split Inference

To make Algorithm 2 concrete, we walk through inference on the held-out command `"jump after walk"`, which never appears in source training because it contains `jump`. The grammar parser (Step 1) identifies two primitives (`walk`, `jump`) and the relation `after`. Step 2 runs the frozen encoder on each primitive separately: `"walk"` $\rightarrow$ `I_WALK`; `"jump"` $\rightarrow$ `I_JUMP`. Step 6 recomposes using the `after` relation, which in SCAN semantics means *first do the second primitive, then the first*, yielding `I_WALK I_JUMP`. Crucially, the encoder never observes the out-of-distribution composition `jump after walk`; it processes only `walk` and `jump` individually, both of which are in-distribution for `walk` and in the held-out set for `jump`. The compositional novelty is handled entirely by the deterministic parser–recomposer in Steps 1 and 6 of Algorithm 2.

Algorithm 1 DisentangledLSTM training

Require: Source dataset $\mathcal{D}_S = \{(x_i, y_i, y_i^p, y_i^r)\}$, hidden size d, auxiliary weight α, learning rate η, epochs E.
Initialise encoder Enc_θ, decoder Dec_ϕ, auxiliary heads W_p, W_r.
for epoch $= 1, \ldots, E$ **do**
  for minibatch $\{(x, y, y^p, y^r)\} \subset \mathcal{D}_S$ **do**
    $h \leftarrow \mathrm{Enc}_\theta(x)$ {forward + backward final states concatenated}
    $h_{\mathrm{prim}} \leftarrow h[1{:}d/2],\ h_{\mathrm{rel}} \leftarrow h[d/2{+}1{:}d]$ {hard split, Equation (2)}
    $\hat{y} \leftarrow \mathrm{Dec}_\phi(h)$ via attention
    $\hat{y}^p \leftarrow W_p\, h_{\mathrm{prim}},\ \hat{y}^r \leftarrow W_r\, h_{\mathrm{rel}}$ {auxiliary classifiers, Equation (3)}
    $\mathcal{L} \leftarrow \mathrm{CE}(\hat{y}, y) + \alpha\big[\mathrm{CE}(\hat{y}^p, y^p) + \mathrm{CE}(\hat{y}^r, y^r)\big]$ {composite loss, Equation (4)}
    Update θ, ϕ, W_p, W_r by Adam at rate η.
  end for
end for
Discard W_p, W_r for inference; retain θ, ϕ.
return $\mathrm{Enc}_\theta, \mathrm{Dec}_\phi$.

Algorithm 2 Scaffolded Inference

Require: Test command x^*, trained $\mathrm{Enc}_\theta, \mathrm{Dec}_\phi$, grammar $\mathcal{G}$.
$(p_1, r_1, p_2, r_2, \ldots, p_k) \leftarrow \mathrm{Parse}_{\mathcal{G}}(x^*)$ {symbolic decomposition, Section 4.2}
for $i = 1, \ldots, k$ **do**
  $h_i \leftarrow \mathrm{Enc}_\theta(p_i)$
  $\hat{y}_i \leftarrow \mathrm{Dec}_\phi(h_i)$ {independently decode each primitive's output}
end for
$\hat{y}^* \leftarrow \mathrm{Compose}(\hat{y}_1, r_1, \hat{y}_2, r_2, \ldots, \hat{y}_k)$ {recompose using parsed relations}
return $\hat{y}^*$.

D. Hyperparameters and Reproducibility

Architecture. Bidirectional LSTM encoder, hidden size $d = 256$ (split as $128 + 128$); LSTM decoder with Bahdanau attention (Bahdanau et al., 2015), hidden size 256; embedding dimension 128; dropout 0.5 on encoder outputs.

Optimisation. Adam (Kingma & Ba, 2015), learning rate 10^{-3}, batch size 128, gradient clipping at 1.0, 15 epochs of source training, 5 epochs of few-shot fine-tuning, auxiliary weight $\alpha = 5$ (selected by grid search over $\{0.1, 0.5, 1, 2, 5, 10\}$; full results in Table 5).

Few-shot protocol. For $K \in \{1, 4, 16\}$ we draw K examples uniformly from the held-out compositional split, fine-tune for 5 epochs at the same learning rate, and evaluate on the remaining held-out examples. Five seeds per configuration.

Compute. Single NVIDIA V100 GPU per run. Source training takes roughly two hours; few-shot fine-tuning under five minutes. Total compute for all results in this paper (including the α-sensitivity grid in Table 5 and the four splits in Table 4, each across five seeds) is approximately 160 GPU-hours.

Code and data. SCAN is available at `https://github.com/brendenlake/SCAN`; COGS at `https://github.com/najoungkim/COGS`. Our code, configuration files, and raw evaluation logs are available at `https://github.com/shivaay790/compositional-generalization`.

D.1. Reproducibility Checklist

- *Datasets and licenses.* SCAN (Lake & Baroni, 2018): BSD-3-Clause. COGS (Kim & Linzen, 2020): MIT. Both used unmodified.
- *Splits.* Standard *jump*, *turn-left*, *compose*, and *simple-walk* splits as released by the respective dataset authors. No new splits were created.
- *Hyperparameters.* Listed above. The full α-grid search is reported in Table 5; all other hyperparameters were fixed at standard SCAN-baseline values without per-split tuning.

- *Random seeds.* Five seeds per configuration ($\{0, 1, 2, 3, 4\}$). Seed values are passed to PyTorch and NumPy at the start of each run.
- *Compute.* Reported above.
- *Software.* PyTorch 2.x, Python 3.10. Code uses no proprietary libraries.
- *Reporting.* All numbers in the main paper are means; standard deviations are reported in Table 6.

E. Matched-Capacity Protocol (Future Work)

To causally isolate the contribution of disentanglement we propose a matched-capacity baseline. Because DisentangledLSTM with $d = 256$ allocates only 128 dimensions per factor, a capacity-fair comparison should run a non-split bi-LSTM with $d = 256$ against a DisentangledLSTM with $d = 512$ (so each subspace has 256 dimensions, matching the baseline). If the gap in Table 1 survives this rescaling, the disentanglement effect is not an artefact of effective capacity. This experiment is the highest-priority next step. A complementary control would hold the auxiliary supervision constant while removing only the hard split, training a $d = 256$ network with $\alpha = 5$ but no factorisation; the ablation in Table 3 (the "no split" row, 85.22% at 1-shot) gives a preliminary estimate, but the full 4-shot CTS with this configuration would tighten the causal claim.

F. Adversarial Decorrelation Variant (Future Work)

In domains where the auxiliary classification heads are insufficient to drive $I(z_p; z_r)$ low, the natural extension is to attach two adversarial classifiers via gradient-reversal layers (Ganin & Lempitsky, 2015): one trained to predict relations from h_{prim}, the other primitives from h_{rel}, both through a sign-flipped gradient. The full objective becomes

$$\mathcal{L}_{\text{adv}} \;=\; \mathcal{L}_{\text{seq2seq}} \;+\; \alpha(\mathcal{L}_p + \mathcal{L}_r) \;-\; \beta\big(\mathcal{L}_p^{\text{adv}} + \mathcal{L}_r^{\text{adv}}\big), \tag{6}$$

where the third term is maximised with respect to the encoder (gradient reversal) and minimised with respect to the adversarial heads. The sign-flipped third term in Equation (6) directly penalises any remaining mutual information between the two subspaces, targeting the $\lambda \cdot I(z_p; z_r)$ term in Equation (1) more directly than the standard auxiliary heads do. Calibrating β relative to α requires care: too-large β destabilises encoder training in the same way that too-aggressive domain-adversarial losses do (Ganin & Lempitsky, 2015). We anticipate $\beta \in [0.1, 1.0]$ as a reasonable starting range based on standard practice.