JOINT LEARNING FOR ROBUST AND EFFICIENT NEURAL GRAMMAR INDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Grammar induction is the process of recovering the latent hierarchical structure of sentences without supervision, a notoriously unstable process, with different seeds converging to different grammars. While recent work on grounding probabilistic context-free grammar induction using vision as a fixed auxiliary signal showed modest improvement, it did not resolve the instability problem. Rather than treating vision as a fixed signal, we present a joint-learning approach that learns syntax and a shared image–text semantic space simultaneously, reducing variance across seeds while increasing the accuracy of the induced grammar. We test the robustness of our approach along three dimensions: across languages (English, French, and Chinese), across datasets (from Abstract Scenes to MS-COCO), and across tasks on a downstream compositional benchmark (COCO-Order task of ARO). The advantages of joint learning replicate across all three dimensions, where we highlight the importance of selecting the appropriate data to avoid collapse onto trivial grammars, introducing a harder negative-sampling regime. Furthermore, using the induced grammar model as an image–text encoder and scorer improves over CLIP on the ARO word-order task by roughly 25 points. Finally, we demonstrate that the quality of a model's induced grammar predicts its word-order accuracy, evidence that more linguistically aligned representations generalise better.

1 INTRODUCTION

Grammar induction is the task of recovering the latent hierarchical structure of sentences without any supervision (Klein & Manning, 2004; Kim et al., 2019). It is used to evaluate how much syntactic structure can be learnt from data alone, and to study the inductive biases and learning objectives that cause structure to emerge (Warstadt & Bowman, 2023). Conventionally, the standard objective maximises the log marginal likelihood of the observed text, summing over every tree (every hierarchical derivation) that could have produced it. However, this is a weak constraint as multiple different grammars can assign a sentence the same probability, therefore being indistinguishable in the optimisation landscape (Park & Kim, 2024). This drives different random seeds to converge to different grammars, some of which are trivial or unlike those we believe are implicitly learnt by humans (Williams et al., 2018). Current models that induce grammar are thus unstable, with reported parsing accuracy varying significantly across runs, further complicating both progress and evaluation. That accuracy is the overlap between the phrases of the induced tree and those of a gold reference parse (Figure 1).

One way to further constrain the space of learnable grammars comes from perception. Language is grounded in experience of the world, and much of that experience is visual (Bisk et al., 2020). Many constituents (spans that behave as a single grammatical unit, such as the noun phrase *the red ball*) name objects or scenes. In *the red ball is on the table*, the span *the red ball* should match a paired image more closely than a random span of the same sentence such as *ball is on*. That is a reason to prefer some trees over others that assign the sentence the same likelihood. Visually grounded grammar induction adds a contrastive image–text loss that rewards constituents that align well with the image (Zhao & Titov, 2020; Shi et al., 2019). Zhao & Titov (2020) employ a visually grounded PCFG (VC-PCFG) that treats vision as a fixed auxiliary signal: image features come from a frozen visual encoder, so the perceptual representation itself is not reshaped by language. This remains the strongest visually grounded induction model, however, the accuracy gains are modest, and the instability across seeds remains. We argue that visual grounding is most effective not as a fixed

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.

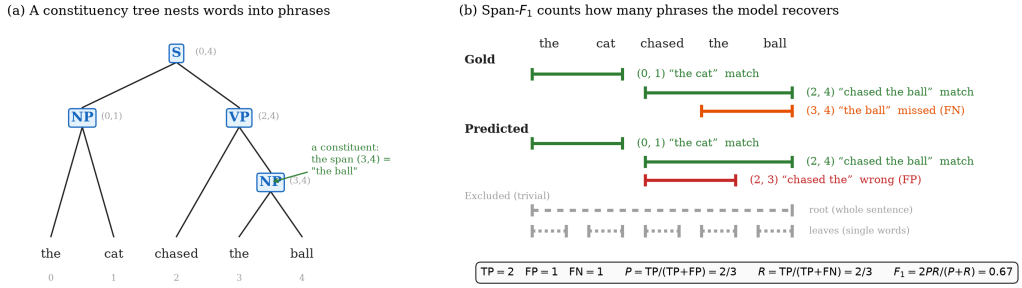

Figure 1: How we score an induced parse. **(a)** Nested phrases (constituents). **(b)** Span-F_1 compares predicted and gold phrase spans, ignoring labels and excluding single words and the whole sentence.

signal but as part of a joint-learning objective, in which syntax and a shared image–text semantic space are learnt together from the beginning of training. We therefore train a semantic encoder f_m that maps a fixed visual signal (akin to perception) into a shared semantic space that language (and therefore syntax) is allowed to reshape. Hence, there are two objectives mutually constraining one another rather than one merely regularising the other. We show that this joint approach, together with a harder negative-sampling regime, produces grammars that are both more accurate and robust across random seeds.

Prior work (Portelance et al., 2025) established these advantages of joint learning in a single language (English) and on a single synthetic dataset (Abstract Scenes; Zitnick & Parikh, 2013). If they are not an artifact of that setup, they should show in different languages, datasets, and downstream tasks. First, we extend the experiments to French and Chinese, successfully replicating the advantages and signatures of joint learning. Second, we move from simplified, synthetic images to naturalistic photographs by training on MS-COCO (Lin et al., 2014). To improve performance, we give the model a richer cross-situational learning signal, a notion we take from cognitive science: any one pairing is referentially ambiguous, and the correct mapping of words to images is the one that keeps co-occurring as competing scenes change (Yu & Smith, 2007; Smith & Yu, 2008). Thus, we introduce a harder negative-sampling regime: larger batches of visually similar images, so that distinguishing scenes requires recovering finer visual structure. This lets us test not only the objectives, but also how the data are presented, and how that data interacts with the objectives to constrain the space of induced grammars. Finally, we turn the induced grammar and the semantic encoder into a compositional image–text scorer and test it on an out-of-distribution task, the word-order split of ARO (Attribution, Relation, and Order) (Yuksekgonul et al., 2023). Whereas contrastive vision–language models such as CLIP (Radford et al., 2021) behave like bags of words (Yuksekgonul et al., 2023), leveraging the induced hierarchical structure improves over CLIP by roughly 25 points (46.8% to 71.5%). Notably, this uses only $\sim$0.6M image–caption pairs, nearly two orders of magnitude fewer than the $\sim$400M used to train CLIP, demonstrating that endowing the model with a structural bias is also more data-efficient. We also show a correlation between the accuracy of the induced grammar and ARO accuracy across multiple model configurations.

2 The Joint-Learning Model

Our model is also a VC-PCFG. The difference is that we train a semantic encoder f_m, rather than treating vision as a fixed grounding signal as in Zhao & Titov (2020). The model is built around the coupling between its two learning signals: the syntactic objective that induces a grammar, and the semantic objective that couples the constituents that grammar generates to a paired image (Portelance et al., 2025). The full pipeline is shown in Figure 2.

2.1 The Syntactic Objective: From Grammars to a Trainable Loss

A context-free grammar (CFG) is defined as $G = (S, \mathcal{N}, \mathcal{P}, \Sigma, \mathcal{R})$, where S is the start symbol ("a sentence"), $\mathcal{N}$ is a set of nonterminals (phrasal categories such as noun-phrases or verb-phrases), $\mathcal{P}$ is a set of preterminals (part-of-speech-like categories such as nouns), Σ is the vocabulary or set of terminals (the actual words), and $\mathcal{R}$ is a set of rules that define the transitions between symbols. In

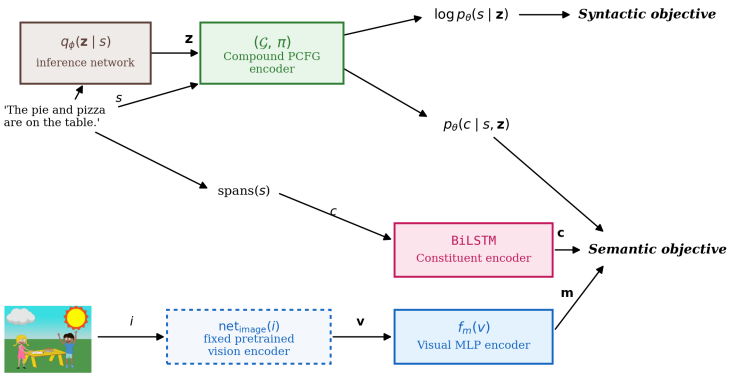

Figure 2: The joint model. A caption x is encoded by an inference network (a BiLSTM) that produces the per-sentence latent $\mathbf{z}$. The compound PCFG then yields $\log p_\theta(x \mid \mathbf{z})$ and a marginal constituent probability over spans. Each span is encoded by a BiLSTM and matched to the semantic embedding $\mathbf{m} = f_m(v)$ of the trainable semantic encoder, weighted by those marginals. Adapted from Portelance et al. (2025).

Chomsky normal form every rule has one of three shapes,

$$S \to A, \qquad A \to B\,C, \qquad T \to w,$$

with $A \in \mathcal{N}$, $B, C \in \mathcal{N} \cup \mathcal{P}$, $T \in \mathcal{P}$, $w \in \Sigma$. Note that the categories are latent and unlabeled, where the model learns multiple nonterminal classes but is never told which one is a "noun-phrase". It independently discovers these phrases and relationships without prior knowledge of linguistics. Following Kim et al. (2019) and Zhao & Titov (2020) we allow $|\mathcal{N}| = 30$ nonterminals and $|\mathcal{P}| = 60$ preterminals. The terminal vocabulary Σ is dataset dependent (§3).

A probabilistic CFG (PCFG) assigns to each of the three rule types a probability $\pi_{S\to A}$, $\pi_{A\to BC}$, or $\pi_{T\to w}$, with the constraint that the productions of every left-hand side form a valid distribution:

$$\sum_{A\in\mathcal{N}} \pi_{S\to A} = 1, \qquad \sum_{B,C} \pi_{A\to BC} = 1 \quad \text{for each } A \in \mathcal{N}, \qquad \sum_{w\in\Sigma} \pi_{T\to w} = 1 \quad \text{for each } T \in \mathcal{P}. \tag{1}$$

The probability of a parse tree t is then the product of the rules applied to build it, $p(t) = \prod_{r\in t} \pi_r$ (each r is one of $S \to A$, $A \to BC$, or $T \to w$), and the probability of a sentence x is the sum over every tree that yields it, $p(x) = \sum_{t:\,\text{yield}(t)=x} p(t)$. Grammar induction consists of recovering the rule probabilities (and hence the latent trees) that generated the observed sentences.

The classical approach stores one free scalar per rule, which struggles to recover linguistically meaningful structure (Kim et al., 2019). This arises from treating every rule as independent from each other, ignoring that related rules should behave similarly. Thus, we take the approach of Kim et al. (2019) where each symbol $N \in \{S\} \cup \mathcal{N} \cup \mathcal{P}$ is assigned a learned embedding $\mathbf{w}_N$, and small neural networks f_s, f_t (two-layer residual MLPs) map these embeddings to scores, where a softmax over each left-hand side turns these into a valid distribution:

$$\begin{aligned}
\pi_{S\to A} &= \frac{\exp\big(\mathbf{u}_A^\top f_s(\mathbf{w}_S)\big)}{\sum_{A'\in\mathcal{N}} \exp\big(\mathbf{u}_{A'}^\top f_s(\mathbf{w}_S)\big)}, \\
\pi_{A\to BC} &= \frac{\exp\big(\mathbf{u}_{BC}^\top \mathbf{w}_A\big)}{\sum_{B'C'} \exp\big(\mathbf{u}_{B'C'}^\top \mathbf{w}_A\big)}, \\
\pi_{T\to w} &= \frac{\exp\big(\mathbf{u}_w^\top f_t(\mathbf{w}_T)\big)}{\sum_{w'\in\Sigma} \exp\big(\mathbf{u}_{w'}^\top f_t(\mathbf{w}_T)\big)}.
\end{aligned} \tag{2}$$

It is from the application of the networks to every embedding that we can leverage distributed representations, where similar symbol embeddings map to similar rules, allowing for better generalization and extraction of useful features across symbols. This is the neural PCFG of Figure 3 (left).

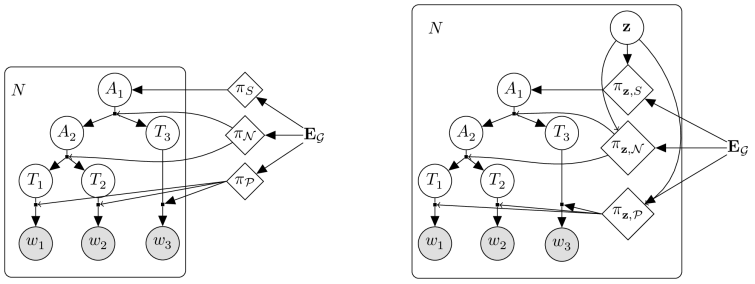

Figure 3: View of the neural PCFG (left) and compound PCFG (right) for one tree, adapted from Kim et al. (2019). A_i are nonterminals, T_i preterminals, w_i words, and $\mathbf{E}_{\mathcal{G}}$ are the learned symbol embeddings that parameterise the rules. The sentence latent $\mathbf{z}$ feeds every rule distribution $\pi_{\mathbf{z}}$.

However, the neural PCFG still assumes that constituents that are not directly linked as parent and child are independent of one another (the context-free independence assumption). This is a strong simplification, since the right way to expand a phrase often depends on the rest of the sentence. Kim et al. (2019) relax the assumption with a compound parameterisation: a single continuous vector $\mathbf{z}$ is associated with each sentence, and every rule probability is conditioned on it. Concretely, the symbol embedding $\mathbf{w}_N$ in equation 2 is replaced by its concatenation with $\mathbf{z}$, $[\mathbf{w}_N; \mathbf{z}]$, so that all rule decisions share the same per-sentence context. A choice made in one part of the tree can then inform choices elsewhere (if *watched* is used as a verb, its sibling is more likely to be a noun phrase such as *a movie*). The neural compound PCFG (C-PCFG) is shown in Figure 3 (right).

Given a sentence x, we need the distribution over the $\mathbf{z}$ that could have produced it (the posterior over the latent). However, computing it exactly is intractable as it would mean integrating over every possible $\mathbf{z}$, $\log p_\theta(x) = \log \int_{\mathbf{z}} \sum_{t \in \mathcal{T}(x)} p_\theta(t \mid \mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$. Thus, we amortise it with an inference network $q_\phi(\mathbf{z} \mid x)$ that embeds the words of x with a single-layer BiLSTM, max-pooling its hidden states into one sentence vector, and then outputting a Gaussian mean $\boldsymbol{\mu}$ and log-variance $\log \boldsymbol{\sigma}^2$. We subsequently apply the reparameterisation trick to sample $\mathbf{z}$,

$$[\,\boldsymbol{\mu};\, \log \boldsymbol{\sigma}^2\,] = W \operatorname{maxpool}\big(\mathrm{BiLSTM}(x)\big), \qquad \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

with $\mathbf{z} \in \mathbb{R}^{64}$. Cleverly, conditioning on $\mathbf{z}$, the compound PCFG is an ordinary PCFG, so the sentence probability is given by the inside algorithm, a dynamic program that sums over all parse trees $\mathcal{T}(x)$ in $O(n^3)$ time (Baker, 1979): $\log p_\theta(x \mid \mathbf{z}) = \log \sum_{t \in \mathcal{T}(x)} p_\theta(t \mid \mathbf{z})$. Again, because the outer integral over $\mathbf{z}$ is intractable, in place of the exact marginal we maximise the evidence lower bound (ELBO),

$$\log p_\theta(x) \;\geq\; \mathrm{ELBO}(x; \phi, \theta) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|x)}\big[\log p_\theta(x \mid \mathbf{z})\big]}_{\text{fit the sentence}} - \underbrace{\mathrm{KL}\big[q_\phi(\mathbf{z} \mid x) \,\|\, p(\mathbf{z})\big]}_{\text{stay near the prior}}. \tag{4}$$

The first term rewards latents $\mathbf{z}$ that make the observed sentence likely, and the second is a regulariser that pulls the inferred posterior toward the prior. Both terms are differentiable, the first through the reparameterisation trick of equation 3, and the KL in closed form for two Gaussians (Kingma & Welling, 2014), so we can apply gradient descent. The syntactic loss over the corpus $\mathcal{C}$ is

$$\mathcal{L}_{\text{syntax}}(\mathcal{C}) = -\sum_{x \in \mathcal{C}} \mathrm{ELBO}(x; \phi, \theta). \tag{5}$$

2.2 The semantic objective: grounding constituents in images

The semantic objective is a multimodal contrastive loss applied at the level of constituents rather than the whole sentence (Figure 2).

On one side, an image i is first passed through a fixed vision encoder, a pretrained network we never update, yielding a perceptual feature vector v that is independent of language. A separate trainable semantic encoder f_m then projects v into a shared image–text semantic space, $\mathbf{m} = f_m(v)$. It is this trained projection into a shared space (and not a fixed grounding signal) that forms a latent semantic representation that language is allowed to reshape during learning.

On the other side, each candidate span c (a contiguous word sequence) is encoded independently by a single-layer BiLSTM, $\mathbf{e}_c = \mathrm{enc}(c)$. A hinge then pushes matching span–image pairs above mismatched ones drawn from other examples in the batch,

$$h(\mathbf{e}_c, \mathbf{m}) = \big[\,\mathrm{sim}(\mathbf{e}_{c'}, \mathbf{m}) - \mathrm{sim}(\mathbf{e}_c, \mathbf{m}) + \epsilon\,\big]_+ + \big[\,\mathrm{sim}(\mathbf{e}_c, \mathbf{m}') - \mathrm{sim}(\mathbf{e}_c, \mathbf{m}) + \epsilon\,\big]_+, \quad (6)$$

where sim is cosine similarity, ϵ a margin, $[\cdot]_+ = \max(0, \cdot)$, and $\mathbf{e}_{c'}, \mathbf{m}'$ are negatives. A match is a span of caption x with that caption's image, whereas a mismatch is the same span with another image in the batch, or another caption's span with this image. That match is then weighted by the grammar's marginal that c is a constituent,

$$\ell_{\text{pair}}(x, \mathbf{m}) = \sum_{c \in \mathrm{spans}(x)} p_\theta(c \mid x, \mathbf{z})\, h(\mathbf{e}_c, \mathbf{m}), \qquad \mathcal{L}_{\text{semantics}} = \sum_j \ell_{\text{pair}}\big(x^{(j)}, f_m(v^{(j)})\big). \quad (7)$$

The two objectives now become coupled: the grammar decides which spans are worth grounding, while the grounding signal pushes the grammar toward trees whose constituents best match the image. This marginal $p_\theta(c \mid x, \mathbf{z})$ is computed by the same inside algorithm, so the semantic loss is also fully differentiable.

2.3 The joint-learning model and its staged ablations

The joint-learning model optimises both objectives together and from the start of training, $\mathcal{L}_{\text{joint}} = \alpha_1 \mathcal{L}_{\text{syntax}} + \alpha_2 \mathcal{L}_{\text{semantics}}$, with $\alpha_1 = \alpha_2 = 1$.

To isolate the effect of jointly learning both vision and language, we compare this model against two staged ablations that initially receive only one of the two objectives, switching to both losses halfway through training. The semantics-first model begins with only the semantic loss active for the first half of training. Thus, it initially has no structural information about the constituents and reduces to a whole-caption, CLIP-style matching loss. The syntax-first model begins by optimising only the syntactic objective without any grounding from vision (hence doing ungrounded neural grammar induction), before switching to both losses at the halfway point. Comparing the joint model against these two ablations will allow us to isolate the effect of joint learning from the start on both the accuracy of induced grammars and the learning efficiency of our model.

3 Data and languages

The first dataset we use is Abstract Scenes (Zitnick & Parikh, 2013), clip-art illustrations paired with simple, complete sentences. It contains 10,020 images, each paired with six sentences, for a total of 60,120 image-sentence pairs. Because this dataset contains simpler and limited vocabulary intended for children, the vocabulary Σ is set to the 2,000 most frequent word types. Images are encoded with a self-supervised SimCLR model (Chen et al., 2020) trained from scratch so as to avoid any linguistic labels or biases from leaking into the model.

Three different languages were evaluated on Abstract Scenes: English, French, and Chinese. English is a right-branching language, meaning that it has a tendency towards parses such as "(*the cat [on the mat [by the door]]*)", thus an always "attach right" heuristic already obtains a high score. Chinese contains more mixed branching than English and thus this will allow us to observe whether the model generalises to more complex branching structures. French's branching structure sits somewhere in between these two languages and will further validate our results. We report right and left branching baselines, so the gains are read relative to the branching priors of each language.

For Abstract Scenes we obtain French sentences with an open-source instruction-tuned LLM (Qwen2.5-7B-Instruct) and Traditional-Chinese sentences with DeepSeek-V3.2, queried to preserve the style of the dataset's simple children's-book vocabulary. In both cases, a random subset (n=600) of the machine translations was verified by a native speaker and found to be accurate in over 98% of cases. Gold trees for evaluation come from Benepar (Kitaev & Klein, 2018) for English and French, and from HanLP (He & Choi, 2021) for Chinese.

The second dataset used is MS-COCO (Lin et al., 2014; Chen et al., 2015), naturalistic images paired with free-form captions, which do not necessarily have to be complete sentences. This is the dataset used in prior work by Shi et al. (2019) and Zhao & Titov (2020) to test their grammar induction

model on images and phrases closer to those encountered in reality. MS-COCO 2014 comprises 123,287 images with five captions each for a total of $\approx 617k$ image–caption pairs. Because this dataset contains a higher diversity of scenarios and therefore more words employed, we raise the vocabulary Σ to the $10{,}000$ most frequent word types. The main experiments use two ResNet vision encoders: self-supervised DINO-RN50 (Caron et al., 2021), which has never seen linguistic labels, and supervised ResNet-152 (He et al., 2016), which matches the backbone of Zhao & Titov (2020). We additionally evaluate two Vision-Transformer backbones that share the same architecture (ViT-B/16) and differ only in whether pretraining used label supervision, reported in Appendix A.1.

4 Robustness across languages

Prior work found that learning syntax and grounding jointly improves both the accuracy and the stability of the induced grammar in English (Portelance et al., 2025). Here we test whether these effects are robust to language shift, across typologically distinct languages. We train the joint, semantics-first and syntax-first models, five seeds each, and evaluate the mean span-F_1 of the most likely parse against the gold trees. For each sentence, span-F_1 is the harmonic mean of precision and recall over the predicted constituent spans, scored against the gold-standard parse (excluding the leaf and root spans). We report the mean over sentences (macro span-F_1), as illustrated in Figure 1. Further examples showing the gold vs induced trees for the same sentence in all three languages, and for a naturalistic MS-COCO caption, are in Appendix A.3. We also report the left and right branching baselines.

Figure 4 reports macro span-F_1 over training for all three languages. The main trend holds across all three languages, with the joint model obtaining the highest accuracy (English 0.89, French 0.82, Chinese 0.86) and less variance than its staged ablations, as observed by the shaded area. Joint-learning outperforms the right-branching baselines for all three languages, with the semantics-first model reaching high accuracy only after both objectives are introduced halfway through training.[1] The fact that this jump appears in the semantics-first model but not the syntax-first one is a signature of semantic bootstrapping, the idea that prior semantic knowledge can be leveraged to acquire syntax (Pinker, 1984). The semantics-first model first learns to align whole captions with images, and this semantic representation then bootstraps the induction of syntactic constituents once both objectives are optimised jointly.

In contrast, the syntax-first model always seems to enter a region of the induced grammar's landscape that it cannot later escape, even after the semantic objective is introduced. This is because the syntactic objective encouraged the model to learn a grammar that arbitrarily maximized the likelihood of observing a sentence with constituents that had no grounding in perception. Whereas the semantics-first model, in the first half of training, faced no pressure to make the sentence likely and only sought to align sentence embeddings with images. This discrepancy between models demonstrates that it is not only the presence of both objectives that matters, but that they are learnt jointly from the start. Only then do the two objectives mutually constrain the space of possible grammars, yielding both the highest accuracy and almost no variance across all three languages.

5 Robustness across scenes and across visual encoders

Abstract Scenes is built on synthetic clip-art, which places relatively constrained distributions over both the linguistic space (a smaller vocabulary and simpler descriptions) and the visual space (a limited set of clip-art scenes), limiting the scope of our conclusions. We therefore turn to MS-COCO, where the images and captions are more realistic and diverse, allowing us to make more general claims about the robustness and advantages of joint learning.

Zhao & Titov (2020) visually ground a compound PCFG by aligning constituents with a paired image, using frozen ResNet-152 features rather than a jointly trained semantic encoder f_m. We reran that model (same hyperparameters) on our data and gold trees. It reaches only 47.5 ± 9.1 macro

[1] English semantics-first underperforms because one of five seeds collapsed onto a trivial grammar, whereas the other four finished near the joint model. That single-seed collapse is the instability this paper later addresses.

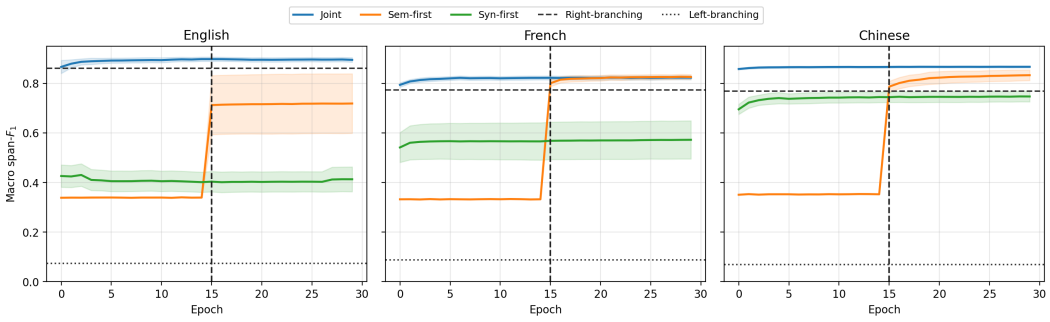

Figure 4: Macro span-F_1 over training on Abstract Scenes for English, French, and Chinese. Shading is SD over five seeds, dashed/dotted horizontals are right-/left-branching baselines, and the vertical dashed line is the staged-model switch at epoch 15.

span-F_1 and collapses to different grammars across seeds (Table 1).[2] We attribute this instability to a grounding signal that is too weak to constrain the grammar: when the semantic signal is weak, the under-constrained syntactic objective dominates and many grammars equally maximise the sentence likelihood (Park & Kim, 2024). Why, then, is the learning signal from vision weaker here than in Abstract Scenes? This can be attributed to a combination of two factors. First, whereas all the images from Abstract Scenes are relatively similar, this is not the case for MS-COCO, where random in-batch negatives are trivially distinguishable. Therefore, the vision encoder is not forced to extract meaningful features to distinguish between images, and can rely on co-occurrence statistics. Secondly, only using five image-caption pairs per batch (the batch size used by Zhao & Titov (2020)) can be insufficient and lead to a noisy gradient update, leading the model to a region of grammar space from which it cannot later escape.

In order to address these two issues we perform an ablation on three different batch sizes (5, 20, 42) and harder contrastive pairs, each drawn at random from that image's 64 nearest neighbors in DINO-RN50 feature space (a fixed similarity index, used for every visual encoder). As observed in Figure 8 (Appendix A.2), using harder negative pairs improves both stability and performance. Increasing batch size makes the learning signal less noisy and also leads to increased performance, without any seed collapsing when using twenty hard-negative pairs. This gain in performance is a reminder of the importance of appropriate data selection during training, even if the objective and optimisation functions are suitable. Another way to think about it is that the benefits of grounding and joint learning are related to the relative richness of the input. On MS-COCO, the captions are more diverse and rich in content, and at the same time the set of random images does not provide a rich visual signal. Therefore, the model previously had little to no gain from paying attention to the images and defaulted to learning grammar merely from the observed sentences.

We re-ran both ResNet encoders (the self-supervised DINO-RN50 and the supervised ResNet-152) under the best regime (batch size $= 20$, hard negatives). The results can be observed in Table 1, with a clear increase in performance and reduction in variance for both encoders. Our replication of Zhao and Titov's model under the best regime improves performance (57.1 ± 9.6), but underperforms compared to the joint model and still suffers from relatively high variance even with harder negatives. We also ran the semantics-first and syntax-first variants under this regime: the training trajectories repeat the semantic-bootstrapping signature of Figure 4 on natural images (Figure 5). The joint model is highest and most stable, and only the semantics-first variant recovers after both objectives are introduced. To ensure that this is not an artifact of ResNets, we repeated the evaluation with two Vision-Transformer backbones that share the ViT-B/16 architecture and differ only in pretraining: supervised (ImageNet labels) vs. self-supervised (DINO, no labels). This isolates the effect of linguistic biases leaking in through supervised pretraining, as we hold the architecture fixed. The joint model again outperforms the others, with the same trajectories (Appendix A.1, Table 3 and Figure 7). This supports our central hypothesis: learning syntax and a shared semantic space jointly from the start, yields more robust grammars that align better with the gold structure.

[2]Zhao & Titov (2020) report 59.4 ± 8.3 sentence-level F_1 on MS-COCO. We obtain 47.5 ± 9.1 under our right-binarized gold trees, averaging all five seeds with none dropped. We were unable to match their published score. A later re-evaluation (Li et al., 2024) reports the top four of ten seeds in order to replicate Zhao & Titov (2020)'s results. We do not apply that filter.

Model	**Encoder**	**Superv.**	**Random** (bs 5)	**Hard-neg** (bs 20)
Joint (ours)	DINO-RN50	self-sup.	38.6 ± 24.3	65.2 ± 0.5
Joint (ours)	ResNet-152	sup.	46.0 ± 21.5	67.3 ± 3.6
Right-branching baseline			55.0	
Zhao (our repl.)	ResNet-152	sup.	47.5 ± 9.1	57.1 ± 9.6

Table 1: Macro span-F_1 (mean $\pm$ SD, five seeds, none dropped) on MS-COCO. **Random**: random negatives, batch size 5. **Hard-neg**: 64 nearest DINO-RN50 neighbors, batch size 20. Joint rows are our model (trainable f_m). Zhao is our replication of Zhao & Titov (2020) on the same gold trees.

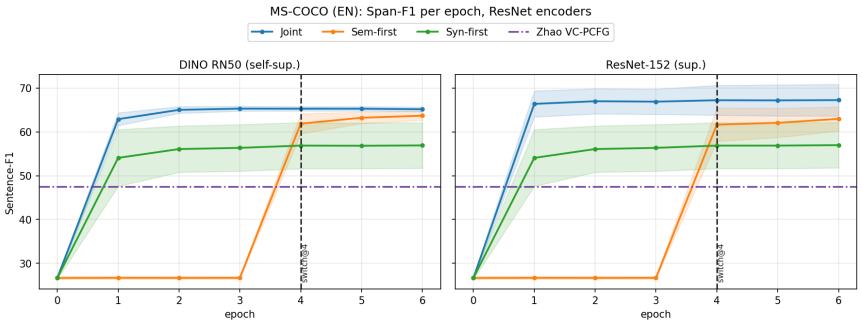

Figure 5: MS-COCO macro span-F_1 over training for DINO-RN50 (self-supervised) and ResNet-152 (supervised). Shading is SD over five seeds, the vertical dashed line is the staged-model switch at epoch 4, and the horizontal one is the Zhao VC-PCFG score (47.5, their hyperparameters).

6 Generalisation and efficient learning on a downstream compositional encoding task

Prior grounded-grammar induction work limits itself to evaluating parsing accuracy and does not really test this model outside this domain. Therefore, we ask whether explicit access to hierarchical representations of language will improve performance on downstream tasks. Specifically, we test on the COCO-Order task of ARO (Yuksekgonul et al., 2023), which evaluates the word-order sensitivity of image–caption matching: given an image, the model must rank the true caption above four word-order perturbations of it (e.g. *a dog sitting on a couch* vs. *a couch sitting on a dog*).

Contrastive vision–language models such as CLIP are known to behave like bags of words (Yuksekgonul et al., 2023), scoring near chance when a caption's words are reordered. We look into three different ways of turning our induced grammar into a matching image-caption score. For semantic embedding $\mathbf{m} = f_m(v)$, span embedding $\mathbf{e}_k$ with marginal weight p_k and similarity $s_k = \cos(\mathbf{m}, \mathbf{e}_k)$, the marginal score is the per-span similarities weighted by the grammar's marginal probability that the span is a constituent, $\sum_k p_k s_k / \sum_k p_k$. Then, the Viterbi score is the mean similarity over the spans of the single most likely tree, and the uniform score weights every per-span similarity uniformly. This will allow us to isolate the contribution of utilizing the grammar's induced knowledge, and we will compare against CLIP as a further baseline.

Table 2 shows the results on the joint DINO-ResNet-50 run, averaged over five seeds. The marginal scorer successfully reaches 71.5%, and the Viterbi scorer 63.1%, both clearing the CLIP baseline (46.8%). Importantly, they outperform the uniform scorer, which scores 7.2%, below chance. The uniform scorer uses the same span embeddings but ignores the induced grammar, weighting every span equally ("grammar-off" in Table 2). That this falls below chance indicates that discarding the induced tree is not merely uninformative but actually harmful. We hypothesize that during training the learned semantic vector is shaped to align with the few constituents that have high marginal probability, and to decrease alignment with spans the grammar treats as ungrammatical. However, the spans of a perturbed sentence are out-of-distribution to what the encoder encountered during training, so the alignment of these spans to the image is mainly neutral. Uniformly averaging over all spans could therefore drag the correct caption below a neutral baseline, thus scoring below chance.

Scorer	P@1 (%)
Marginal	71.5 ± 1.8
Viterbi (single best tree)	63.1 ± 4.7
Uniform (grammar-off)	7.2 ± 1.1
CLIP (Radford et al., 2021)	46.8
Chance	20.0

Table 2: COCO-Order P@1 (%) on the joint DINO-ResNet-50 model (mean $\pm$ SD over five seeds). Marginal and Viterbi use the induced grammar, whereas Uniform ignores it (grammar-off).

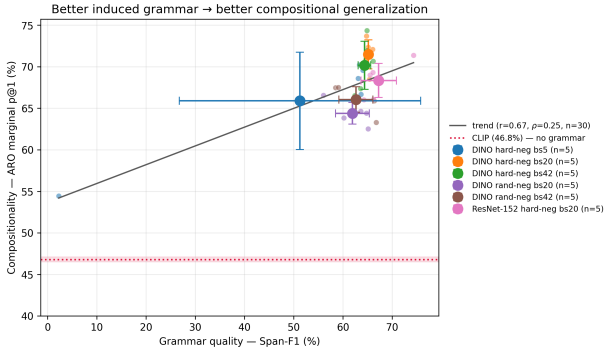

Figure 6: Span-F_1 vs. COCO-Order P@1 under the marginal scorer (n=35 runs across DINO-RN50 and ResNet-152). The dotted line is CLIP (46.8%); Pearson r=0.64, Spearman ρ=0.48.

Figure 6 plots the ARO performance against parsing accuracy for different training configurations and encoders under the marginal scorer. As observed, there is a correlation between better induced grammars and better compositional abilities (Pearson $r = 0.64$, Spearman $\rho = 0.48$), indicating that more linguistically-aligned representations generalise better to this out-of-distribution task. This suggests that the right inductive biases are also more data-efficient. Our model is trained on only $\sim$0.6M image–caption pairs, significantly fewer than the $\sim$400M used to train CLIP, while outperforming it at encoding structural information such as word order. Encoding hierarchical structure into the model, rather than relying on scale alone, is what enables this compositional generalisation.

7 Discussion

These experiments suggest that the benefits of jointly learning syntax and a shared semantic space are not an artifact of one language or one dataset, and that it indeed yields grammars that are both more accurate and more stable across languages (§4), naturalistic scenes (§5), and a downstream word-order task (§6). However, this occurs only when the visual signal is rich enough to constrain the grammar. That richness depends on a cross-situational learning regime, in which any single image–caption pairing underdetermines meaning, and the correct mapping is recovered from which word–scene pairings consistently co-occur as competitors change (Yu & Smith, 2007; Smith & Yu, 2008). We implemented that regime with harder negatives, so each contrast is ambiguous and the model is forced to extract richer semantic features. On MS-COCO, Zhao and Titov's VC-PCFG with fixed vision embeddings leaves the grammar under-constrained (47.5 ± 9.1 macro span-F_1), whereas our joint model with a trainable semantic encoder raised performance significantly (67.3 ± 3.6). This induced grammar was also shown to efficiently generalise to a downstream task, outperforming CLIP by roughly 25 points on COCO-Order. These results suggest that syntax and a shared semantic space need to be learnt together so that each can reshape the other, rather than treating vision as a fixed signal that merely regularises an independently induced grammar.

The advantages of jointly learning syntax and semantics are clear, but they are currently biased towards concepts with a concrete visual referent, and a static image remains an incomplete learning signal. Human perception is continuous and encodes the temporal and spatial structure of events. Video would give a richer learning signal for grounding verbs, relations and function words, and is the natural next test of whether the same joint principle extends beyond static images.

REPRODUCIBILITY STATEMENT

All data preprocessing, model, training, and evaluation code will be released in a public repository upon publication. The link is omitted from this submission for anonymization.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data (machine-translated Abstract Scenes captions), assist with translation, clean and reformat datasets, design and provide feedback on research methodology and experiments (including evaluation metrics), and implement methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative and thematic data analysis, or interpret results. The remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create and edit analysis code, create or modify scientific figures, identify relevant literature, format references, and edit the paper to improve readability. We have reviewed all AI-assisted work: machine translations were checked by a native speaker on a random subset as reported in §3, and AI-assisted code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

James K. Baker. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132, 1979. doi: 10.1121/1.2017061.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8718–8735, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers (DINO). In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations (SimCLR). In *International Conference on Machine Learning (ICML)*, 2020.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Han He and Jinho D. Choi. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders (HanLP). In *Proceedings of EMNLP*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition (ResNet). In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yoon Kim, Chris Dyer, and Alexander M. Rush. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder (Benepar). In *Proceedings of ACL*, 2018.

Dan Klein and Christopher D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL*, 2004.

Boyi Li, Rodolfo Corona, Karttikeya Mangalam, Catherine Chen, Daniel Flaherty, Serge Belongie, Kilian Q. Weinberger, Jitendra Malik, Trevor Darrell, and Dan Klein. Re-evaluating the need for visual signals in unsupervised grammar induction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1078–1093, 2024.

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 *European Conference on Computer Vision (ECCV)*, 2014.

Jinwook Park and Kangil Kim. Structural optimization ambiguity and simplicity bias in unsupervised neural grammar induction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Steven Pinker. *Language Learnability and Language Development*. Harvard University Press, 1984.

Eva Portelance, Siva Reddy, and Timothy J. O'Donnell. Reframing linguistic bootstrapping as joint inference using visually-grounded grammar induction models. *Journal of Memory and Language*, 145:104672, 2025. ISSN 0749-596X. doi: 10.1016/j.jml.2025.104672. URL https://www.sciencedirect.com/science/article/pii/S0749596X25000658.

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 (CLIP). In *International Conference on Machine Learning (ICML)*, 2021.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1842–1861, 2019.

Linda B. Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

Alex Warstadt and Samuel R. Bowman. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, 2023.

Adina Williams, Andrew Drozdov, and Samuel R. Bowman. Do latent tree learning models identify meaningful structure in sentences? *Transactions of the Association for Computational Linguistics (TACL)*, 6:253–267, 2018.

Chen Yu and Linda B. Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420, 2007.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023.

Yanpeng Zhao and Ivan Titov. Visually grounded compound PCFGs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4369–4379, 2020.

C. Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

A Appendix

A.1 Training trajectories on MS-COCO (ViT encoders)

Figure 7 repeats Figure 5 for the ViT encoders. The same semantic-bootstrapping pattern holds, and the self-supervised (DINO) and supervised encoders track each other closely. Table 3 reports the corresponding final Span-F_1 scores, showing performance does not depend on whether or not the vision encoder was pretrained with label supervision.

Model	Encoder	Superv.	Random (bs 5)	Hard-neg (bs 20)
Ours (joint)	DINO-ViT-B/16	self-sup.	58.5 ± 6.1	63.6 ± 1.2
	ViT-B/16	sup.	59.3 ± 6.0	62.6 ± 2.4

Table 3: Macro span-F_1 (mean $\pm$ SD over five seeds) on MS-COCO English for the Vision-Transformer encoders, using the same joint model and Random / Hard-neg columns as Table 1. Holding the architecture (ViT-B/16) fixed, only the supervision objective changes.

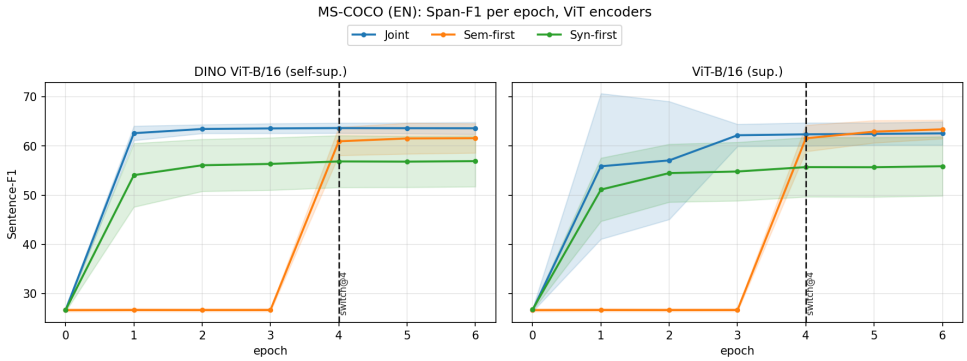

Figure 7: MS-COCO English macro span-F_1 over training for the ViT encoders: DINO-ViT-B/16 (self-supervised) and ViT-B/16 (supervised). Shading is SD over five seeds, and the vertical dashed line is the staged-model switch at epoch 4.

A.2 Disentangling negative hardness from negative count

Figure 8 separates how hard or similar the contrastive image pairs (negative pairs) are from how many of them sit in a batch. At every matched batch size, hard negatives sit above random in-batch negatives. Random batches of size 5 are unstable: some seeds collapse onto trivial grammars, and the median falls below the right-branching baseline (55.0).

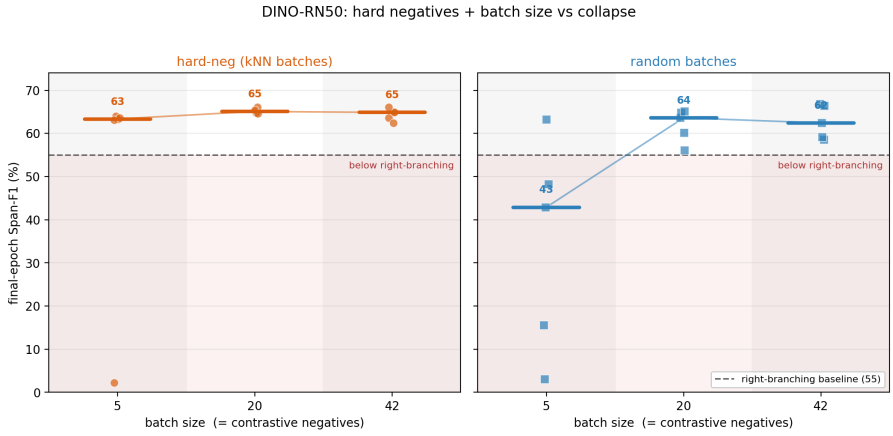

Figure 8: Hardness of contrastive images vs. batch size (DINO-RN50, joint model). Each dot is one seed's final Span-F_1. At matched batch size the hard-negative sampler (orange) sits uniformly above random batches (blue). The random sampler is finding multiple trivial grammars at low batch sizes and its median falls below the right-branching baseline (55.0).

A.3 Qualitative examples: induced vs. gold trees

To complement the aggregate span-F_1 numbers, we show the most-likely tree of a trained joint model against the gold parse for individual sentences. Green marks a phrase both trees share, orange a gold phrase the model missed, and red a phrase the model invented. Single-word leaves and the whole-sentence root are omitted, matching the metric in §4.

Figure 9 shows the induced parse of the same sentence, “the cat wants to eat the hot dog”, in English, French, and Chinese. In all three languages the model recovers the same nested predicate form: who is acting (the cat), then wanting, then eating, then the food. In English the mistake arises out of misclassifying “[[the hot] dog]” instead of “[the [hot dog]]”, but it still correctly treats “the hot dog” as a single constituent. In French, the error arises out of a failure to close the subject NP “[le chat]” at the very top of the tree, whereas in Chinese (F_1=1.00) the parse is exact.

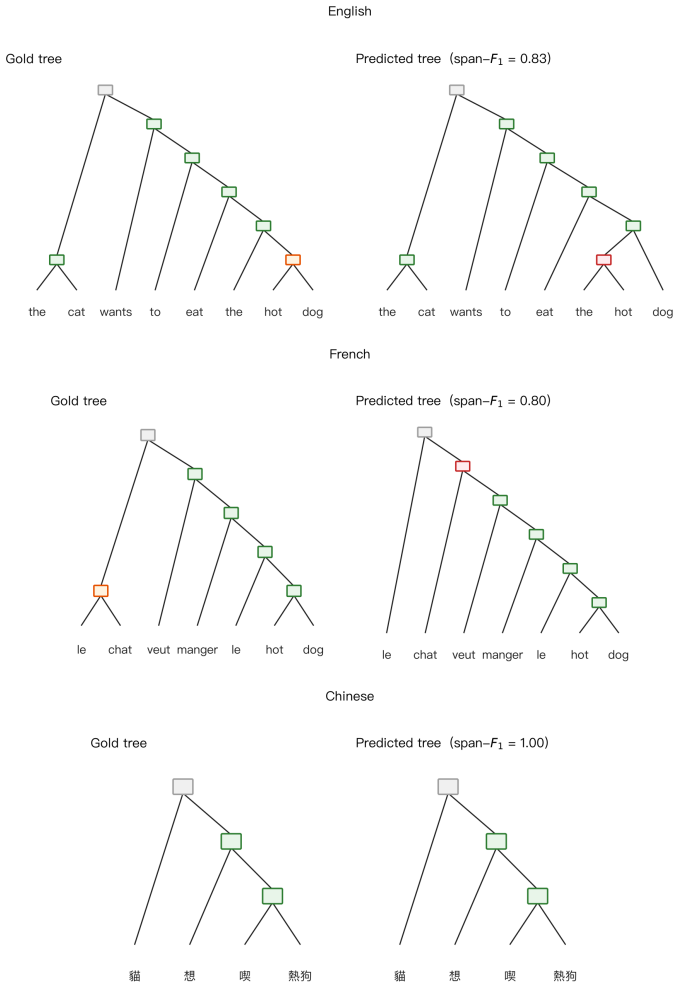

Figure 9: Gold vs. induced most-likely tree for the same sentence, “the cat wants to eat the hot dog”, in English (top), French (middle), and Chinese (bottom), from the joint model on Abstract Scenes. The nested “want → eat → [the hot dog]” structure is recovered in all three languages. The few errors are minor and interpretable, while Chinese is exact.

Figure 10 shows an exact parse (F_1=1.00) of the MS-COCO caption “a phone and a computer on a kitchen counter”. This is a non-right-branching structure, from which the model discovers three distinct linguistic phenomena. First, coordination, where “a phone” and “a computer” are grouped into parallel conjuncts joined by *and* to form a single higher noun phrase. Second, Prepositional Phrase (PP) attachment, where the PP “on a kitchen counter” attaches to the whole coordinated NP, giving the correct “two objects on the counter” reading. Third, compound nouns, with “kitchen counter” being treated as one unit. Recovering coordinated and PP-embedded structure on free-form naturalistic captions shows that the gains in Table 1 reflect real hierarchical structure.

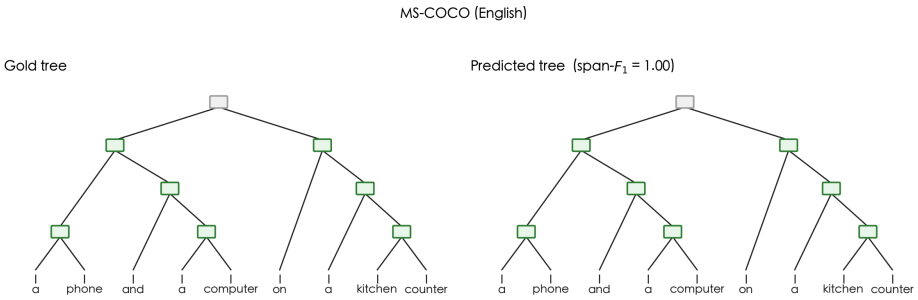

Figure 10: Gold vs. induced most-likely tree for the MS-COCO caption "a phone and a computer on a kitchen counter" (joint model, ResNet-152, hard negatives). The model exactly recovers coordination (*a phone and a computer*), attaches *on a kitchen counter* to both objects, and treats *kitchen counter* as one compound.