Beyond Incremental Beam Search for Compositional Explanations of Neurons

Biagio La Rosa
University of California, Santa Cruz
bilarosa@ucsc.edu

Leilani H. Gilpin
University of California, Santa Cruz

Abstract

Compositional explanations describe the alignment between neuron activations and concepts through logical formulas, obtained by solving a search problem over combinations of concepts. Due to the combinatorial nature of this space, prior work relies on incremental beam search, which restricts the exploration but limits the expressiveness of explanations and may discard intermediate candidates needed to identify higher alignment. In this paper, we propose a non-incremental beam search framework that expands the search space while keeping the computation tractable. Our approach introduces a non-incremental beam expansion strategy and a functional-equivalence-aware beam search that leverages equivalence classes of formulas to improve both search quality and efficiency. Experiments on natural language and vision tasks show that our method finds explanations expressing a higher alignment than the incremental formulation while maintaining comparable runtime, and leads to a significant increase in the number of optimal explanations when these can be computed. Overall, these results show that expanding search beyond incremental formulations leads to more expressive and accurate explanations, making our method a promising replacement for standard incremental beam search.

1 Introduction

Compositional explanations [25] are a class of methods that describe the alignment between neuron activations and human-interpretable concepts through propositional logic formulas. These explanations capture how neurons respond to combinations of low-level (e.g., colors, textures) and high-level (e.g., objects, semantic attributes) concepts by identifying logical rules that best match the spatial or feature-level activation patterns of a neuron. Compositional explanations are typically obtained by solving a search problem over the space of possible concept combinations, where the objective is to identify the formula that maximizes alignment with the neuron activation mask.

A major problem in this setting is that the search space grows combinatorially with the number of concepts and the length of the explanation. As a result, prior work has adopted strong assumptions on the formula structure and leverages beam search to restrict the exploration to a subset of the state space [25, 11, 22, 21, 20]. In particular, the current literature enforces an incremental construction of formulas, where concepts are added sequentially. Although this design makes the search efficient, it also limits the expressiveness of the resulting explanations and may prevent the discovery of formulas that express higher alignment between activations and concepts (Figure 1). Indeed, intermediate candidates that are not selected by beam search at early stages cannot be revisited later, even if they would lead to better explanations when combined differently.

In this paper, **we take a first step towards addressing the limitations imposed by incremental beam search**. Specifically, we propose a **non-incremental beam search framework that expands the exploration beyond the incremental formulation while maintaining tractable computation**.

Preprint.

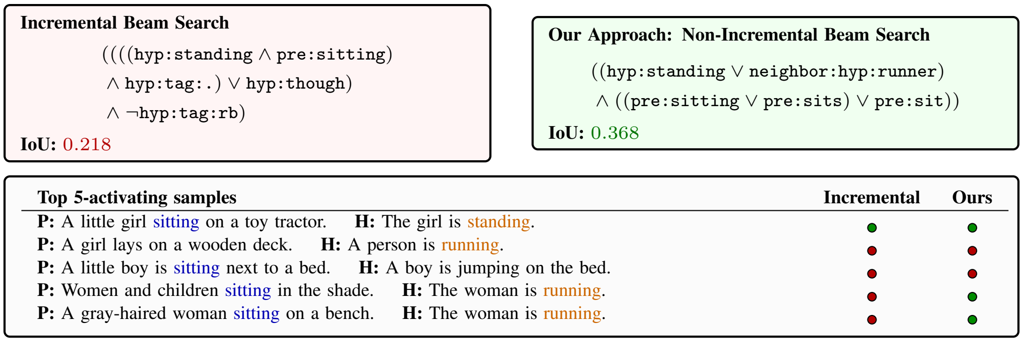

Figure 1: An example of explanations computed by incremental beam search and our non-incremental approach for a neuron that recognizes contradictions between sitting actions and standing or moving actions.

Our approach introduces a novel beam expansion strategy that generates candidates by combining the best candidate formulas previously explored and by inserting concepts within internal substructures, thus enabling the search to access regions of the state space that are unreachable under the incremental assumption. Additionally, we propose a functional-equivalence-aware beam search that identifies labels that induce the same activation mask and leverages this property to avoid redundant exploration and increase the number of combinations explored. We evaluate the proposed framework on both natural language inference and image classification tasks across 6 different models, following standard setups for this area of research.

In summary, our contributions are as follows:

- We introduce the notion of functional equivalence for compositional explanations, formalizing when different logical formulas induce the same activation mask.
- We propose a non-incremental beam search framework that supports the generation of non-incremental formulas and explicitly manages functional equivalence during beam expansion.
- We show that the proposed framework finds explanations with stronger alignment in natural language tasks and reduces the gap between optimal and non-optimal explanations in settings where optimal solutions can be computed, such as image classification.

We will release the code upon acceptance.

2 Related Work

Neuron explanations aim to understand what individual neurons learn during training. A wide range of methods has been proposed across both vision and natural language domains [23], including approaches that generate representative samples that maximize neuron activations [8, 29, 26], methods that associate neurons with concepts or textual descriptions through correlation-based analyses and foundational models [13, 27, 28, 15, 1, 31, 10], and approaches that study internal representations through probing or sparse decompositions [7, 9, 5].

Among these, compositional explanations [25] represent a specific family of methods that focus on **alignment between a neuron activation range and the locations of concepts**, expressing this relationship through propositional logic formulas. The seminal work in this area is Network Dissection [2, 3], which associates each neuron with the single concept that maximizes alignment. This approach was extended by Mu and Andreas [25] to consider combinations of concepts to model polysemantic behavior, where neurons respond to multiple, potentially unrelated concepts. Subsequent work has extended this formulation to capture richer relationships between concepts, including co-occurrence [2], exclusion [25], relative position [11], and hierarchy [22].

Key properties of compositional explanations are that they are **cross-domain** [25], they explicitly capture polysemanticity and they express complex logical relationships between concepts, while remaining grounded at the level of individual neurons and specific activation ranges [20]. While compositional explanations share the goal of understanding neuron behavior with other interpretability methods, they differ fundamentally along several dimensions. In particular, compositional explanations target alignment rather than correlation, operate at the level of individual neurons rather than aggregated representations or layer-level representations, and express knowledge through logical formulas rather than statistical associations. As a result, these approaches are typically considered complementary rather than directly comparable, as they capture different aspects of neuron behavior.

This paper focuses on addressing one of the limitations of compositional explanation methods: the incremental construction assumption. While this assumption is effective in practice, it limits the set of formulas that can be explored and may prevent the discovery of the best alignment. Our work builds on prior literature and addresses this limitation by proposing a non-incremental search framework that expands the beam state space while maintaining tractable computation, and by explicitly modeling functional equivalence between formulas during the search.

3 Method

This section introduces the compositional explanation problem, the notation, and terminology used in the area. Then, we present the concept of functional equivalence and describe the proposed non-incremental beam search algorithm.

Background and Terminology Let $\mathfrak{D} = \{x_1, \ldots, x_m\}$ be a probing dataset used to analyze the behavior of a trained model. Let $\Omega = \mathfrak{D} \times \{1, \ldots, d\}$ denote a set of locations. Each location $\omega = (x, j)$ identifies feature j in sample x (e.g., a pixel in vision or a token in NLP). Let $\mathfrak{L}^1 = \{c_1, \ldots, c_p\}$ be a set of atomic concepts. Each concept $c_k \in \mathfrak{L}^1$ is associated with a binary mask $M_k \subseteq \Omega$ representing the locations where the concept is present. We denote by $\mathbf{M} = \{M_k\}_{k=1}^m$ the collection of all concept masks. Let $\mathcal{O}$ be the set of all admissible logical connectives. In this work, in line with prior work, we consider as connectives $\wedge$, $\vee$, and $\neg$, although the framework can be applied to any admissible operator family satisfying the assumptions of the search space. Given a neuron z, we define its activation mask as $N \subseteq \Omega$ representing the locations where the neuron activates within a selected activation range $[\tau_1, \tau_2]$. The activation range can be identified by thresholding the highest activations [25] or by using clustering techniques [20]. Let $\mathfrak{L}^n$ denote the set of admissible logical formulas of arity at most n constructed from $\mathfrak{L}^1$. Each label $L \in \mathfrak{L}^n$ is associated with a semantic operator

$$\circ_L(\mathbf{M}) \subseteq \Omega \tag{1}$$

defined recursively by applying logical connectives to concept masks. In particular, for $L = c_k$, $\circ_L(\mathbf{M}) = M_k$, and for composite labels:

$$\circ_{L_1 \wedge L_2}(\mathbf{M}) = \circ_{L_1}(\mathbf{M}) \cap \circ_{L_2}(\mathbf{M}), \quad \circ_{L_1 \vee L_2}(\mathbf{M}) = \circ_{L_1}(\mathbf{M}) \cup \circ_{L_2}(\mathbf{M}), \quad \circ_{L_k \wedge \neg c_j}(\mathbf{M}) = \circ_{L_k}(\mathbf{M}) \setminus M_j$$

We denote $C(L)$ as the set of atomic concepts appearing in label L. Finally, let L be a label containing a subformula of the form $A \wedge \neg c$, where $c \in \mathfrak{L}^1$. The occurrence of $\neg c$ is *unverified* if $\circ_A(\mathbf{M}) \cap M_c = \emptyset$. Equivalently, $\circ_{A \wedge \neg c}(\mathbf{M}) = \circ_A(\mathbf{M})$. A label is *unverified* if it contains at least one unverified negation.

The goal of compositional explanations is to identify

$$L^* = \underset{L \in \mathfrak{L}^n}{\arg\max}\, IoU(N, \circ_L(\mathbf{M})), \tag{2}$$

where the IoU score is defined as

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

Previous compositional explanation methods typically restrict the search space by two assumptions. First, each concept may appear at most once in a label. Second, labels are constructed incrementally, by adding one concept at a time to the current formula. Under these assumptions, labels take the form $(((c_1 \oplus_1 c_2) \oplus_2 c_3) \ldots \oplus_{k-1} c_k)$, where $c_i \in \mathfrak{L}^1$ are distinct concepts and $\oplus_i \in \mathcal{O}$. Even with these restrictions, the search space grows rapidly with the vocabulary size. For this reason, prior work commonly relies on beam search to make the problem tractable. However, the combination of incremental construction and beam search can limit the complexity of the relationships captured by the explanations, thus limiting its expressiveness, and can discard intermediate labels that are necessary to later construct high-quality formulas. Our method addresses this limitation by expanding the beam search with additional non-incremental candidates and by explicitly accounting for functional equivalence among labels.

3.1 Functional Equivalence

We propose and define functional equivalence as an equivalence relation over labels based on their induced masks. Indeed, we observe that, in general, multiple syntactically different labels may induce the same mask over Ω. We formalize this notion through functional equivalence.

Definition 3.1 (Functional Equivalence)**.** Two labels $L_1, L_2 \in \mathfrak{L}^n$ are functionally equivalent if

$$\circ_{L_1}(\mathbf{M}) = \circ_{L_2}(\mathbf{M}) \tag{4}$$

Equivalently, $IoU(\circ_{L_1}(\mathbf{M}), \circ_{L_2}(\mathbf{M})) = 1$. Functional equivalence partitions $\mathfrak{L}^n$ into equivalence classes of labels that are indistinguishable under the dataset. We distinguish three types of equivalence, depending on their origin.

Type 1: Logical Equivalence

$$L_1 \equiv_{\text{logic}} L_2 \iff \forall \mathbf{M},\ \circ_{L_1}(\mathbf{M}) = \circ_{L_2}(\mathbf{M})$$

This equivalence is intrinsic to the logical structure of the formulas (e.g., commutativity) and holds independently of the dataset. In practice, checking logical equivalence is computationally expensive and, in prior work, is not explicitly enforced during search beyond the simplest cases.

Type 2: Dataset Equivalence

$$L_1 \equiv_{\mathfrak{D}} L_2 \iff \circ_{L_1}(\mathbf{M}) = \circ_{L_2}(\mathbf{M})\ \wedge\ \mathcal{C}(L_1) = \mathcal{C}(L_2)\ \wedge\ L_1 \not\equiv_{\text{logic}} L_2$$

This occurs when two labels use the same concepts but differ syntactically, yet produce identical masks due to dataset-specific properties (e.g., absence of certain co-occurrences).

Type 3: Dataset Ambiguity. Let $L_1 = L_{\text{comm}} \oplus L_a$ and $L_2 = L_{\text{comm}} \oplus L_b$ with $L_a \neq L_b$. Define the restriction $A \mid L := A \cap \circ_L(\mathbf{M})$. Then,

$$\circ_{L_a}(\mathbf{M}) \cap \circ_{L_{\text{comm}}}(\mathbf{M}) = \circ_{L_b}(\mathbf{M}) \cap \circ_{L_{\text{comm}}}(\mathbf{M})$$

This case captures situations where different concepts become indistinguishable within a restricted context. Unlike Type 2, this equivalence does not require the same concept set and reflects intrinsic ambiguity in the dataset. For example: $(A$ AND (NOT $B))$ and $(C$ AND (NOT $B))$ are functionally equivalent if the common context given by (NOT B) restricts the samples to a subset where the concepts A and C always co-occur, making them indistinguishable and leading to the same mask.

3.2 Non-Incremental Algorithm

Our proposed algorithm leverages heuristics [20, 19] to guide a beam search algorithm, consistent with prior research [25]. To move beyond incremental approaches, our algorithm proposes a novel design for beam expansion and a functional-equivalence-aware beam search.

3.2.1 Beam Expansion

Let $k+1$ be the current beam level, let B_k be the set of labels selected for expansion (i.e., the set of top b labels in the beam at the previous step k), and let $B_{\leq k} = \bigcup_{r=1}^{k} B_r$ be the set of labels selected for expansion up to and including level k. Given a label $L \in B_k$ and an atomic concept $c \in \mathfrak{L}^1$ not already appearing in L, our expansion procedure generates three families of candidates.

Incremental candidates. The first family corresponds to the standard incremental expansion used in prior beam-search approaches. For each admissible operator $\oplus \in \mathcal{O}$, the algorithm generates

$$L_{\text{inc}} = L \oplus c \tag{5}$$

Non-incremental candidates. The second family inserts the new concept inside one of the internal clauses of L. Let $\mathcal{S}_{\oplus}(L)$ denote the set of subformulas of L that are compatible with the operator $\oplus$. For example, for $\oplus = \vee$, the compatible subformulas are disjunctive clauses, while for $\oplus = \wedge$ and $\oplus = \wedge\neg$, the compatible subformulas are conjunctive clauses. Note that $\wedge\neg$ denotes conjunction with negation applied to a single concept. For each $S \in \mathcal{S}_{\oplus}(L)$, the algorithm produces a new label

$$L_{\text{ni}} = L[S \leftarrow S \oplus c] \tag{6}$$

where $L[S \leftarrow S \oplus c]$ denotes the formula obtained by replacing the subformula S in L with $S \oplus c$. This operation allows the search to construct formulas that cannot be reached by right-incremental expansions within the same beam tree.

Pairwise recombination candidates. The third family is obtained by recombining labels selected for expansion at previous beam levels, but only when the resulting label has arity equal to the current level. Formally, at level $k+1$, given two labels $L_1, L_2 \in B_{\leq k}$, the algorithm considers their pairwise recombination only if

$$|C(L_1) \cup C(L_2)| = k+1 \tag{7}$$

where $C(L)$ denotes the set of atomic concepts appearing in L. Given two labels L_1 and L_2, these combinations are computed using the following rules:

Rule 1: disjoint recombination. If L_1 and L_2 do not share atomic concepts, then the algorithm generates

$$L_1 \vee L_2 \qquad \text{and} \qquad L_1 \wedge L_2 \tag{8}$$

We do not generate $L_1 \wedge \neg L_2$ in this step because the heuristic (and prior research) assumes that negation is applied only to atomic concepts. For example, from $(A \vee B)$ and $(C \wedge D)$, the rule generates $(A \vee B) \vee (C \wedge D)$ and $(A \vee B) \wedge (C \wedge D)$

Rule 2: same-type overlapping recombination. Assume that L_1 and L_2 are both clauses formed using the same operator $\oplus_s \in \{\wedge, \vee\}$ and that they share at least one concept. Let $C_{\text{comm}} = C(L_1) \cap C(L_2)$ and $C_{\text{diff}} = C(L_1) \triangle C(L_2)$, where $C(L)$ is the set of atomic concepts appearing in L and $\triangle$ denotes symmetric difference. Let L_{comm} and L_{diff} be the clauses obtained by chaining the concepts in C_{comm} and C_{diff}, respectively, with operator $\oplus_s$. The algorithm generates

$$L_{\text{comm}} \oplus_s L_{\text{diff}} \qquad \text{and} \qquad L_{\text{comm}} \widehat{\oplus_s} L_{\text{diff}} \tag{9}$$

where $\widehat{\wedge} = \vee$ and $\widehat{\vee} = \wedge$. For example, from $(A \wedge B \wedge C)$ and $(A \wedge C \wedge D)$, the rule identifies A and C as common concepts and generates the labels $((A \wedge C) \wedge B) \wedge D$ and $(A \wedge C) \wedge (B \vee D)$

Rule 3: specialization of overlapping clauses. Assume that one label is disjunctive and the other is conjunctive, and that they share at least one concept. The algorithm specializes the shared concept in the disjunctive clause by replacing it with the conjunctive clause. For example, from $(A \vee B)$ and $(A \wedge C)$, the rule generates $((A \wedge C) \vee B)$

Overall, the candidate set at level k is

$$\mathcal{BC}_{k+1} = \mathcal{E}_{\text{inc}}(B_k) \cup \mathcal{E}_{\text{ni}}(B_k) \cup \mathcal{R}_{k+1}(B_{\leq k}) \tag{10}$$

where $\mathcal{E}_{\text{inc}}$ and $\mathcal{E}_{\text{ni}}$ denote incremental and non-incremental expansion, and $\mathcal{R}_{k+1}$ denotes the set of combinations of arity $k+1$ obtained through pairwise recombination, respectively. This design increases the expressive power of the search while preserving the efficiency of beam search. The beam expansion algorithm is included in the appendix in Algorithm 1.

3.2.2 Functional-Equivalence-Aware Beam Search

The candidate labels in $\mathcal{BC}_{k+1}$ are ranked according to a heuristic estimate of their alignment with the neuron activation mask. Let $h(L)$ denote the heuristic score associated with a label L, and let us assume BC_{k+1} is sorted in decreasing order of $h(L)$. The algorithm scans BC_{k+1} progressively and computes the exact IoU only for candidates considered for insertion into the beam.

Let B_{k+1} denote the current (partial) beam at level $k+1$, with maximum size b. We denote by $L_{\min} = \arg\min_{L \in B_{k+1}} IoU\big(N, \circ_L(\mathbf{M})\big)$ the lowest-scoring label currently in the beam. A candidate L_j is eligible for insertion if

$$IoU\big(N, \circ_{L_j}(\mathbf{M})\big) \geq IoU\big(N, \circ_{L_{\min}}(\mathbf{M})\big) \tag{11}$$

Before inserting L_j, the algorithm checks whether it is functionally equivalent to any label already in the beam. The algorithm distinguishes three cases:

No equivalent label in the beam. If there is no $L_t \in B_{k+1}$ such that L_j is functionally equivalent to L_t, then L_j is inserted into the beam and the lowest-scoring label is removed:

$$B_{k+1} \leftarrow \big(B_{k+1} \setminus \{L_{\min}\}\big) \cup \{L_j\} \tag{12}$$

Type-1 and Type-2 equivalence. If L_j is functionally equivalent to a label $L_t \in B_{k+1}$ through Type-1 or Type-2 equivalence, then $C(L_j) = C(L_t)$ and the difference between the two labels is either purely syntactic or induced by dataset-specific redundancy. Retaining both labels would duplicate the same semantic candidate and reduce beam diversity. Therefore, L_j is discarded and L_t is kept, unless L_t is unverified and L_j is verified. In that case, L_t is replaced by L_j:

$$B_{k+1} \leftarrow \big(B_{k+1} \setminus \{L_t\}\big) \cup \{L_j\} \tag{13}$$

Type-3 equivalence. If L_j is functionally equivalent to a label $L_t \in B_{k+1}$ through Type-3 equivalence, the two labels induce the same mask but differ in their concept composition or internal structure. This situation reflects ambiguity in the dataset rather than redundancy in the formula. In this case, the algorithm inserts L_j into the beam:

$$B_{k+1} \leftarrow \big(B_{k+1} \setminus \{L_{\min}\}\big) \cup \{L_j\} \tag{14}$$

and augments the next candidate sets by recombining L_j and L_t:

$$\mathcal{BC}_t \leftarrow \mathcal{BC}_t \cup \mathcal{R}_t(\{L_j, L_t\}) \tag{15}$$

The process terminates when either all candidates in BC_{K+1} have been processed or when the next candidate $L_j \in BC_{k+1}$ is associated with an estimation $h(L_j) < IoU\big(N, \circ_{L_{min}}(\mathbf{M})\big)$. In this case, no remaining candidate can satisfy the insertion condition, and the current beam is final. The proposed non-incremental beam search is included in the appendix in Algorithm 2.

4 Experiments

In this section, we evaluate the proposed non-incremental compositional explanation framework across vision and natural language tasks. We compare our approach against the incremental beam search used in prior work, analyzing both explanation quality and computational cost. We first describe the experimental setup, evaluation metrics, and the baseline, and then present results on natural language inference and image classification benchmarks.

Setup In this paper, we follow the most common setup of compositional explanations. Specifically, we apply our framework across three models trained on Places365 [33] for image classification (ResNet18 [12], AlexNet [16], and DenseNet161 [14]) and three models trained or finetuned on natural language inference tasks (BiLSTM, BERT [6], and TinyLLaMA 1.1B [32]). These models correspond to those investigated in previous literature [25], with the addition of more recent architectures for NLP (see Section F for further details about the models). As discussed in [19], extending compositional explanations in vision to transformer-based architectures remains an open research question, and we leave this direction for future work.

As probing datasets, we use Broden [2] for image classification, in line with prior work, and SNLI [4] for the language models. For image classification, the concepts correspond to the segmentation masks annotated in the dataset. In the NLI task, we follow [25] and use as the concept set the 2000 most frequent words (tok), part-of-speech tags (tag), overlap between premise and hypothesis, and the neighbors of a word (neighbor), computed as the logical OR between the top 5 closest words in the word2vec space [24]. Following previous approaches, we probe all the neurons in the penultimate layer of each model (the last MLP layer for transformer architectures) and define the neuron activation mask $N \subseteq \Omega$ by selecting the top 0.005 quantile of activations.

Baseline We consider the incremental beam search adopted in previous literature in this research area as a baseline. To the best of the authors' knowledge, all prior approaches in compositional explanations share the same incremental beam search backbone and therefore converge to the same explanations within the state space. The differences among these approaches mainly lie in the heuristic used to explore the state space [25, 20, 17], the stopping criterion [21], and the concept sets

Table 1: Average scores for explanation and search metrics associated with the computation of explanations in the natural language domain by our non-incremental beam search and the standard incremental beam search.

Method	Explanation Quality				Search Metrics		
	Better (%)	**IoU**	**DetAcc**	**ActCov**	**Visited**	**Esti.**	**s/u**
			BiLSTM				
Incremental	2	0.171	0.224	0.477	72	370570	37
Ours	58	**0.192**	**0.249**	**0.483**	42	575615	57
			BERT				
Incremental	4	0.144	0.189	0.411	54	369531	33
Ours	58	**0.156**	**0.202**	**0.435**	42	726259	62
			LLaMA				
Incremental	3	0.179	0.238	0.457	55	364649	34
Ours	65	**0.195**	**0.258**	**0.471**	42	648610	60

or relations considered [11, 22]. Our approach is orthogonal to these dimensions and can be applied to any of these variants without modification, as they all share the same underlying backbone. Among the available heuristics, we use the one proposed by [19] to estimate the IoU of candidate labels, as it is the only one applicable to both vision and NLP tasks. Finally, as discussed in Section 2, we align ourselves with prior literature and do not consider alternative families of neuron explanations (e.g., mechanistic interpretability or domain-specific approaches) as baselines, as they are generally complementary and do not simultaneously capture spatial alignment, polysemanticity, and complex relationships between concepts, and do not compute logical formulas.

Evaluation Metrics To evaluate the impact of non-incremental compositional explanations on both explanation quality and computational cost, we consider the following metrics, commonly adopted in this area [25, 21, 3, 20]: IoU, detection accuracy, activation coverage, visited nodes, estimated nodes, and seconds per unit. A higher IoU indicates a stronger alignment between the neuron activation mask N and the label mask $\circ_L(\mathbf{M})$, and therefore a better explanation. IoU scores in these settings are typically low, not all neurons are interpretable, and the variance across neurons is high. Therefore, even small differences are considered meaningful (see Section A for a discussion of this issue). Activation coverage measures the fraction of neuron activations explained by a label. Detection accuracy measures the fraction of concept annotations associated with a label that are recovered within the activation mask N. For further details on these metrics, we refer to Section B. Finally, we report the number of times each method is the top performer (i.e., it returns the explanation with the highest IoU), since the two algorithms explore different beam trees and neither strictly dominates the other.

4.1 Natural Language Inference

In this section, we explore the application of the proposed framework to the natural language inference task. In line with [25], we set the maximum explanation length to 5 and the beam width to 10. At this length, all components of our algorithm come into play, enabling the exploration of a state space that is inaccessible to the standard incremental formulation (i.e., the space including non-incremental logical formulas).

Table 1 shows the average scores over all the units (the standard deviation is reported and discussed in Section A). Across all models, our algorithm finds stronger alignments (i.e., higher IoU), and thus better explanations, in more than half of the cases, and produces explanations with higher quality scores across all dimensions. These results confirm the importance, in the NLP domain, of being able to express more complex relationships that go beyond incremental constructions. For example, many neurons in these settings recognize contradictions or coherence between semantics embedded in the premise and hypothesis sentences (e.g., Figure 1 or Section G). Similarities in semantics often correspond to differences at the syntactic level (e.g., *girl* vs. *woman* vs. *female* for gender). These types of relationships are difficult to express in an incremental formulation, since the scope of the last operator always covers the entire preceding formula, thus limiting the expressiveness of the

Table 2: Average scores for explanation and search metrics associated with the computation of explanations in the vision domain by our non-incremental beam search and the standard incremental beam search.

Method	**Explanation Quality**					**Search Metrics**		
	Optimal	**Better (%)**	**IoU**	**DetAcc**	**ActCov**	**Visited**	**Esti.**	**s/u**
				ResNet18				
Incremental	0.92	0	0.073	0.188	0.133	27	33446	122
Ours	**0.96**	**5**	0.076	0.181	0.137	71	48154	131
				AlexNet				
Incremental	0.74	0	0.050	0.155	0.089	61	34054	127
Ours	**0.90**	**16**	0.053	0.152	0.089	399	46847	204
				DenseNet161				
Incremental	0.61	0.33	0.037	0.094	0.079	64	36324	129
Ours	**0.76**	**17**	0.038	0.083	0.081	392	49036	180

explanation. Conversely, the possibility of combining candidates and breaking incrementality during beam expansion allows our algorithm to better capture these relationships.

Regarding the search metrics, the additional combinations introduced in the beam search lead to a trade-off in runtime and in the exploration of the search space. However, this trade-off is limited. Specifically, despite the considerably larger number of nodes estimated by the heuristic, the algorithm visits fewer nodes during beam search, and the runtime remains within the same order of magnitude ($\sim$ 30 sec vs. $\sim$ 1 min). Summarizing, since in less than 5% of the cases the incremental beam search produces better explanations, our algorithm is the preferred option in these settings.

4.2 Image Classification

In image classification, we consider the settings adopted in prior work [25, 20], fixing the maximum length to 3 and the beam width to 5. Under these settings, it is possible to compute optimal explanations over the state space $\mathfrak{L}^n$ in a reasonable time (see Section C for a discussion on the inapplicability of this approach to NLP). Moreover, any explanation retrieved by our proposed algorithm is still contained in $\mathfrak{L}^n$, since all non-incremental formulas can be transformed into incremental ones by reordering when the length is less than or equal to 3. In this context, optimal explanations represent the ground truth toward which beam-based explanations aim to converge at a reduced computational cost ($\sim$ 2 minutes instead of 1-3 hours). Therefore, we also compare the incremental and non-incremental approaches in terms of optimality.

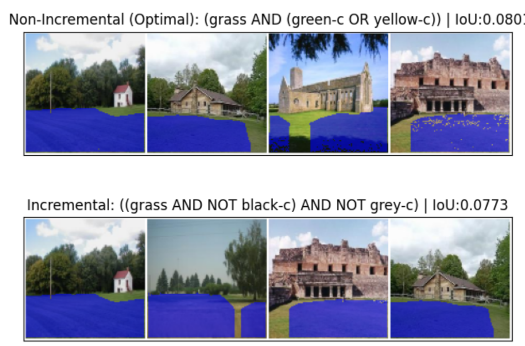

Figure 2: An example of cases where our algorithm reaches optimality while the incremental one does not (Unit 91 of AlexNet). The blue area indicates the activation captured by the explanation.

Table 2 shows the average scores over all units (the standard deviation is reported in Section A). In this domain, our approach significantly increases the number of optimal explanations across all models, representing a substantial step toward closing the gap between fast beam-based explanations and optimal explanations computed at higher cost. In particular, our algorithm is especially effective at reducing the number of non-optimal explanations in cases where beam and optimal explanations achieve similar scores because they share similar concepts but differ in their structure. In these cases, optimal explanations typically return more specific and specialized alignments [19]. Here, the ability to access previous combinations, together with the functional-equivalence-aware beam search, allows the algorithm to correct such cases, avoiding unverified explanations and identifying more specialized combinations of the same concepts (see Figure 2 and Section G).

Regarding runtime, the time required to compute explanations remains close to that of the incremental variant. Although the average is slightly higher, the median time is similar (Section A), suggesting that our algorithm is more sensitive to outliers. Finally, note that, across models, in less than 0.5% of the cases, the incremental beam search outperforms our approach. Therefore, our algorithm can be considered a drop-in replacement for the incremental variant in these settings.

5 Analysis and Limitations

This section discusses additional insights and limitations of the proposed algorithm, derived from its design and from further experiments (e.g., ablation studies) reported in Section D.

Dependency on heuristics The design of the algorithm assumes the availability of a heuristic to guide the search process. The heuristic is fundamental to filter out non-promising combinations and mitigate the impact of the larger state space. Therefore, the effectiveness of the approach depends on the quality of the heuristic. A bad heuristic may negatively impact runtime by increasing the number of visited nodes, which represent the most expensive computational component. We used the heuristic proposed by [18], as it does not assume any specific structure in the data and is broadly applicable. However, in cases where this or other heuristics are ineffective, the proposed algorithm may become too expensive to be applied in practice.

Scalability and Memory Footprint For the standard settings adopted in prior work and reasonable values of the maximum explanation length and beam width, we do not observe large runtime differences between the baseline and the non-incremental variant. However, in extreme settings, the proposed approach may become less effective due to the larger search space and the additional recombinations (Section D). Regarding the memory footprint, functional equivalence requires access to beam masks during the search process. Two strategies can be considered: recomputing masks on demand or caching them in memory. In the first case, the additional memory cost is negligible with respect to the overall setup. In the second case, the memory required to store a single beam is doubled. However, in the setups considered in this paper, we do not observe a significant runtime difference between the two strategies, since functional equivalence checks are relatively infrequent.

Human Interpretability The proposed algorithm slightly modifies the explanation structure to capture more complex relationships. The improvements in explanation quality are primarily technical, as the method is designed to identify better solutions in the search space. However, capturing more complex relationships may affect human interpretability. We argue that these explanations are typically used by researchers to analyze and improve models, in which case the difference between incremental and non-incremental formulations is negligible. Nevertheless, further extensions in this direction, supporting increasingly complex relationships, may hinder interpretability. We leave the study of the trade-off between expressiveness and human interpretability to future work.

6 Conclusion

In this paper, we address the limitations introduced by incremental beam search in compositional explanations. Specifically, we introduced the notion of functional equivalence and proposed a non-incremental beam search framework that expands the beam search space while maintaining tractable computation. We showed that the proposed approach improves the quality of compositional explanations and significantly increases the number of optimal explanations in settings where optimal solutions can be computed. These results demonstrate the importance of exploring a broader class of formulas when searching for explanations. Future work could investigate alternative non-incremental search strategies, extend the framework to other logic operators, and explore additional constraints to further improve efficiency and quality of explanations.

References

[1] Y. H. Ahn, H. B. Kim, and S. T. Kim. Www: a unified framework for explaining what where and why of neural networks by interpretation of neuron concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10968–10977, 2024.

[2] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[3] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907375117. URL `https://www.pnas.org/content/early/2020/08/31/1907375117`.

[4] S. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642, 2015.

[5] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423.

[7] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[8] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. 2009.

[9] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=tcsZt9ZNKD`.

[10] Y. Gur-Arieh, R. Mayan, C. Agassy, A. Geiger, and M. Geva. Enhancing automated interpretability with output-centric feature descriptions. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5757–5778, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.288. URL `https://aclanthology.org/2025.acl-long.288/`.

[11] R. Harth. *Understanding Individual Neurons of ResNet Through Improved Compositional Formulas*, pages 283–294. Springer International Publishing, 2022. ISBN 9783031092824. doi: 10.1007/978-3-031-09282-4_24.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[13] E. Hernandez, S. Schwettmann, D. Bau, T. Bagashvili, A. Torralba, and J. Andreas. Natural language descriptions of deep features. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=NudBMY-tzDr`.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] N. Kalibhat, S. Bhardwaj, C. B. Bruss, H. Firooz, M. Sanjabi, and S. Feizi. Identifying interpretable subspaces in image representations. In *International Conference on Machine Learning*, pages 15623–15638. PMLR, 2023.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[17] B. La Rosa and L. H. Gilpin. Open vocabulary compositional explanations for neuron alignment. Nov. 2025. doi: 10.48550/ARXIV.2511.20931.

[18] B. La Rosa and L. H. Gilpin. Guaranteed optimal compositional explanations for neurons. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=MHiiwC3oFR`.

[19] B. La Rosa and L. H. Gilpin. Guaranteed optimal compositional explanations for neurons. In *Forty-third International Conference on Machine Learning (ICML 2026)*, 2026. URL `https://openreview.net/forum?id=MHiiwC3oFR`.

[20] B. La Rosa, L. H. Gilpin, and R. Capobianco. Towards a fuller understanding of neurons with clustered compositional explanations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=51PLYhMFWz`.

[21] S. M. Makinwa, B. La Rosa, and R. Capobianco. Detection accuracy for evaluating compositional explanations of units. In *AIxIA 2021 - Advances in Artificial Intelligence*, pages 550–563. Springer International Publishing, 2022. doi: 10.1007/978-3-031-08421-8_38.

[22] R. Massidda and D. Bacciu. Knowledge-driven interpretation of convolutional neural networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 356–371. Springer International Publishing, 2023. doi: 10.1007/978-3-031-26387-3_22.

[23] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] J. Mu and J. Andreas. Compositional explanations of neurons. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[26] A. Nguyen, J. Yosinski, and J. Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *Visualization for Deep Learning workshop, ICML 2016*, Feb. 2016.

[27] T. Oikarinen and T.-W. Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=iPWiwWHc1V`.

[28] T. Oikarinen and T.-W. Weng. Linear explanations for individual neurons. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=WIbntm28cM`.

[29] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11), nov 2017. doi: 10.23915/distill.00007.

[30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. Feb. 2023. doi: 10.48550/ARXIV.2302.13971.

[31] A. Wang, W.-N. Lee, and X. Qi. Hint: Hierarchical neuron concept explainer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10254–10264, June 2022.

Table 3: Average and standard deviation explanation scores computed by our non-incremental beam search and the standard incremental beam search on natural language inference.

Method	Better (%)	IoU	DetAcc	ActCov
		BiLSTM		
Incremental	2	0.171 $\pm$ 0.055	0.224 $\pm$ 0.082	0.477 $\pm$ 0.181
Ours	**58**	**0.192** $\pm$ 0.066	**0.249** $\pm$ 0.091	**0.483** $\pm$ 0.143
		BERT		
Incremental	4	0.144 $\pm$ 0.026	0.189 $\pm$ 0.042	0.411 $\pm$ 0.115
Ours	**58**	**0.156** $\pm$ 0.030	**0.202** $\pm$ 0.044	**0.435** $\pm$ 0.118
		LLaMA		
Incremental	3	0.179 $\pm$ 0.046	0.238 $\pm$ 0.070	0.457 $\pm$ 0.128
Ours	**65**	**0.195** $\pm$ 0.050	**0.258** $\pm$ 0.075	**0.471** $\pm$ 0.126

Table 4: Average and standard deviation search scores computed by our non-incremental beam search and the standard incremental beam search on natural language inference.

Method	**Visited**		**Estimated**	**sec/unit**	
	Mean	Median		Mean	Median
			BiLSTM		
Incremental	72 $\pm$ 115	52	370570 $\pm$ 11364	37 $\pm$ 5	37
Ours	42 $\pm$ 3	41	575615 $\pm$ 176084	57 $\pm$ 15	56
			BERT		
Incremental	54 $\pm$ 16	50	369531 $\pm$ 8855	33 $\pm$ 3	33
Ours	42 $\pm$ 3	40	726259 $\pm$ 197093	62 $\pm$ 12	60
			LLaMA		
Incremental	55 $\pm$ 16	50	364649 $\pm$ 10608	35 $\pm$ 3	35
Ours	42 $\pm$ 3	40	648610 $\pm$ 186510	60 $\pm$ 12	59

[32] P. Zhang, G. Zeng, T. Wang, and W. Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[33] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A Complete Results

This section reports the standard deviation of the experiments presented in the main text, as well as the median values for the number of visited states and the runtime. Specifically, Table 3 and Table 5 report the scores for the explanation metrics, while Table 4 and Table 6 report the search-related metrics.

It is important to emphasize that the explanations and the metrics used to evaluate their quality are not expected to exhibit low variance. As observed by Bau et al. [2] and Mu and Andreas [25], not all neurons are aligned with specific concepts, leading to high variability in the degree of alignment. This effect is particularly pronounced for metrics computed over pixel-level data, where the alignment between neuron activations and labeled concepts can fluctuate significantly.

In addition to the observations reported in the main text, we note that the median runtime of the proposed algorithm is close to that of incremental beam search in the vision case. This highlights the sensitivity of the algorithm to outliers, where outliers correspond to neurons for which multiple candidate explanations achieve similar IoU scores (e.g., neurons with overall low alignment). At the same time, this indicates that, for a large portion of cases, the proposed algorithm runs in a time comparable to the incremental version in the vision case, which is a desirable property.

Table 5: Average and standard deviation explanation scores computed by our non-incremental beam search and the standard incremental beam search on image classification.

Method	Optimal	Better (%)	IoU	DetAcc	ActCov
			ResNet18		
Incremental	0.92	0	0.073 $\pm$ 0.042	0.188 $\pm$ 0.108	0.133 $\pm$ 0.090
Ours	**0.96**	**5**	0.076 $\pm$ 0.044	0.181 $\pm$ 0.106	0.137 $\pm$ 0.088
			AlexNet		
Incremental	0.74	0	0.050 $\pm$ 0.024	0.155 $\pm$ 0.101	0.089 $\pm$ 0.066
Ours	**0.90**	**16**	0.053 $\pm$ 0.024	0.152 $\pm$ 0.093	0.089 $\pm$ 0.057
			DenseNet161		
Incremental	0.61	0.33	0.037 $\pm$ 0.022	0.094 $\pm$ 0.058	0.078 $\pm$ 0.072
Ours	**0.76**	**17**	0.038 $\pm$ 0.022	0.083 $\pm$ 0.073	0.081 $\pm$ 0.052

Table 6: Average standard deviation search scores computed by our non-incremental beam search and the standard incremental beam search on image classification.

Method	Visited		Estimated	s/u	
	Mean	Median		Mean	Median
		ResNet18			
Incremental	27 $\pm$ 40	14	33446 $\pm$ 2628	122 $\pm$ 12	117
Ours	71 $\pm$ 204	17	48154 $\pm$ 5721	131 $\pm$ 34	122
		AlexNet			
Incremental	61 $\pm$ 48	40	34054 $\pm$ 2390	127 $\pm$ 8	126
Ours	399 $\pm$ 624	102	46847 $\pm$ 7639	204 $\pm$ 134	137
		DenseNet161			
Incremental	64 $\pm$ 47	60	36324 $\pm$ 2812	129 $\pm$ 12	126
Ours	392 $\pm$ 742	104	49036 $\pm$ 8169	180 $\pm$ 125	129

B Metrics

This section reports the Detection Accuracy and Activation Coverage definitions, as introduced in La Rosa et al. [20]. In the following definitions, $N \subseteq \Omega$ denotes the neuron activation mask and $\circ_L(\mathbf{M}) \subseteq \Omega$ denotes the mask induced by label L over the probing dataset.

Detection Accuracy quantifies the fraction of the label mask that is captured by the neuron activation mask. A high value indicates that most of the regions associated with the label are detected by the neuron within the considered activation range.

$$DetAcc(L, N, \mathbf{M}) = \frac{|N \cap \circ_L(\mathbf{M})|}{| \circ_L (\mathbf{M})|} \tag{16}$$

Activation Coverage measures the fraction of neuron activations that fall within the regions associated with the label. A high value indicates that the label covers a large portion of the activation mask.

$$ActCov(L, N, \mathbf{M}) = \frac{|N \cap \circ_L(\mathbf{M})|}{|N|} \tag{17}$$

C Optimal Explanations

As discussed in the main text, in the vision setting and for explanations of maximum length 3, it is possible to compute ground-truth optimal explanations in a reasonable time using the algorithm proposed by La Rosa and Gilpin [18]. This algorithm requires between 1 and 3 hours per neuron and assumes an incremental formulation. As noted in their work, the algorithm relies on the presence

Algorithm 1 Non-Incremental Beam Expansion

```
Input: Beam $B_k$, previous beams $B_{\leq k}$, concept set $\mathfrak{L}^1$, operator set $\mathcal{O}$, target arity $k+1$
Return: Candidate set $\mathcal{B}_{k+1}$
$\mathcal{B}_{k+1} \leftarrow \emptyset$
for each $L \in B_k$ do
    for each $c \in \mathfrak{L}^1$ such that $c \notin C(L)$ do
        for each $\oplus \in \mathcal{O}$ do
            $L_{\text{inc}} \leftarrow L \oplus c$
            $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{k+1} \cup \{L_{\text{inc}}\}$
        end for
        for each $\oplus \in \mathcal{O}$ do
            Identify the set $\mathcal{S}_{\oplus}(L)$ of subformulas of $L$ compatible with $\oplus$
            for each $S \in \mathcal{S}_{\oplus}(L)$ do
                $L_{\text{ni}} \leftarrow L[S \leftarrow S \oplus c]$
                if $|C(L_{\text{ni}})| = k+1$ then
                    $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{k+1} \cup \{L_{\text{ni}}\}$
                end if
            end for
        end for
    end for
end for
for each pair $(L_1, L_2) \in B_{\leq k} \times B_{\leq k}$ do
    if $|C(L_1) \cup C(L_2)| = k+1$ then
        $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{k+1} \cup \mathcal{R}_{k+1}(L_1, L_2)$
    end if
end for
Return: $\mathcal{B}_{k+1}$
```

of *unique elements* in the dataset, defined as elements associated with a single annotation, and its performance is sensitive to the length of the explanation.

We attempted to apply this algorithm to the NLP setting considered in this paper. However, in this case, the required assumptions do not hold. In particular, there are no unique elements, since each word is associated with at least two concepts (e.g., token and part-of-speech tag). Moreover, since activations are computed at the sentence level, many words are treated as *common* elements despite being distinct. Given these properties, together with the larger maximum explanation length considered in NLP, we observe that the optimal algorithm does not run in a feasible time in this setting. This limitation further motivates the need for approximate or search-based methods in settings where optimal computation is infeasible.

Finally, note that while any explanation of length 3 can be reordered into an incremental form and thus compared with the optimal solution, this is not the case for explanations of length greater than 3. Therefore, even if the optimal algorithm were applied to longer explanations in the vision setting, the results would not be directly comparable. In this direction, future work should explore the notion of optimality in non-incremental settings.

D Additional Experiments

In this section, we analyze the proposed algorithm under several configurations. As the probed model, we use BERT on the natural language inference task.

D.1 Ablation Study

Table 7 shows the difference in performance when ablating individual components of the proposed approach. In particular, we consider four configurations: *full*, which includes all components; *w/o recombinations*, which disables the generation of pairwise recombination candidates during beam expansion as well as the combinations induced by functional equivalence; *w/o Functional Equivalence*, which removes the checks for functional equivalence among promising candidates;

Algorithm 2 Functional-Equivalence-Aware Beam Selection at Level k

Input: Ordered candidate list BC_k, activation mask N, concept masks $\mathbf{M}$, beam width b
Return: Beam B_{k+1}, recombination candidates $\mathcal{B}^{FE}$
$B_{k+1} \leftarrow \emptyset$
for each $L_j \in BC_k$ **do**
  if $|B_{k+1}| = b$ **then**
    $L_{\min} \leftarrow \arg\min_{L \in B_{k+1}} IoU(N, \circ_L(\mathbf{M}))$
    if $h(L_j) < IoU(N, \circ_{L_{\min}}(\mathbf{M}))$ **then**
      break
    end if
  end if
  Compute $IoU(N, \circ_{L_j}(\mathbf{M}))$
  if $|B_{k+1}| < b$ **then**
    $B_{k+1} \leftarrow B_{k+1} \cup \{L_j\}$
    continue
  end if
  $L_{\min} \leftarrow \arg\min_{L \in B_{k+1}} IoU(N, \circ_L(\mathbf{M}))$
  if $IoU(N, \circ_{L_j}(\mathbf{M})) < IoU(N, \circ_{L_{\min}}(\mathbf{M}))$ **then**
    continue
  end if
  Find $L_t \in B_{k+1}$ such that $\circ_{L_j}(\mathbf{M}) = \circ_{L_t}(\mathbf{M})$, if it exists.
  if no such L_t exists **then**
    $B_{k+1} \leftarrow (B_{k+1} \setminus \{L_{\min}\}) \cup \{L_j\}$
  else if L_j and L_t have Type-1 or Type-2 equivalence **then**
    if L_t is unverified and L_j is verified **then**
      $B_{k+1} \leftarrow (B_{k+1} \setminus \{L_t\}) \cup \{L_j\}$
    end if
  else if L_j and L_t have Type-3 equivalence **then**
    $B_{k+1} \leftarrow (B_{k+1} \setminus \{L_{\min}\}) \cup \{L_j\}$
    $\mathcal{B}^{FE}_{k+1} \leftarrow \mathcal{B}^{FE}_{k+1} \cup \mathcal{R}_t(L_j, L_t)$
  end if
end for
Return: B_{k+1}

Table 7: Average scores and standard deviation for explanation metrics when individual components of our framework are removed.

Method	Better (%)	IoU	DetAcc	ActCov
w/o Recombinations	0.64	0.145 $\pm$ 0.021	0.183 $\pm$ 0.030	**0.442** $\pm$ 0.121
Full	**3.18**	**0.151** $\pm$ 0.026	**0.197** $\pm$ 0.027	0.417 $\pm$ 0.124
w/o Functional Equivalence	4	0.135 $\pm$ 0.016	0.174 $\pm$ 0.027	0.397 $\pm$ 0.086
Full	**60**	**0.147** $\pm$ 0.021	**0.187** $\pm$ 0.041	**0.445** $\pm$ 0.109
w/o Non-Incremental Expansion	4	0.137 $\pm$ 0.017	0.182 $\pm$ 0.033	0.382 $\pm$ 0.092
Full	**66**	**0.147** $\pm$ 0.020	**0.190** $\pm$ 0.042	**0.435** $\pm$ 0.111

and *w/o Non-Incremental Expansion*, which removes non-incremental candidates from the beam expansion while retaining recombinations.

Similarly to the main text, we generate explanations for both the ablated and full configurations and compare the explanations that differ pairwise. From Table 7, we observe that functional equivalence and non-incremental expansion are the most impactful components, as the full version finds better alignments in more than 60% of the cases than the ablated versions without these specific components and achieves higher scores across all metrics. Conversely, recombinations have a smaller impact but still contribute to improved overall performance.

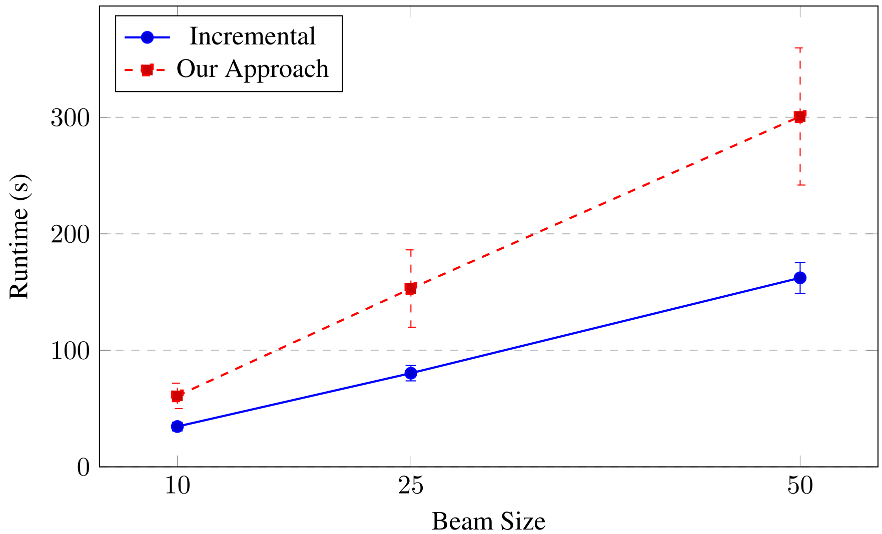

Figure 3: Runtime vs Beam Size. Error bars denote standard deviation.

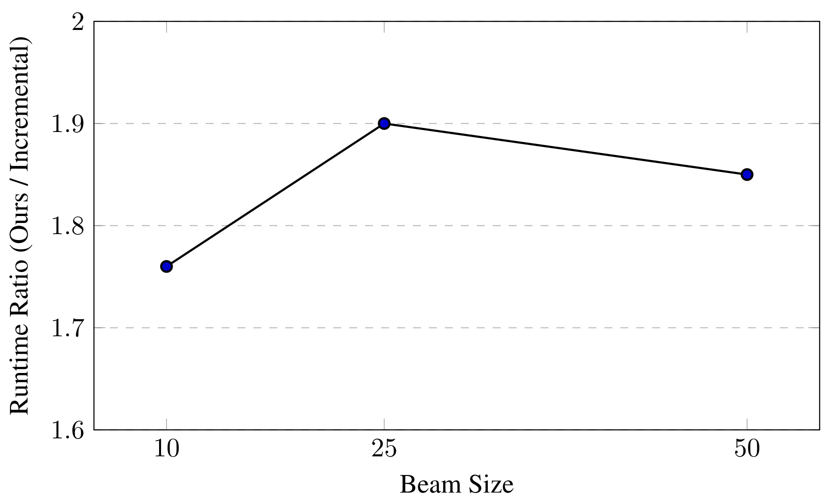

Figure 4: Runtime ratio between our approach and incremental beam search across beam sizes.

D.2 Sensitivity to Beam Size and Explanation Length

In this section, we study the sensitivity of the proposed algorithm to variations in the hyperparameters related to beam width and explanation length. In practice, both parameters are usually kept relatively small to preserve interpretability and tractable runtime. However, due to the inclusion of recombinations and the larger search space explored by the proposed method, we investigate whether the runtime gap between the incremental and non-incremental variants grows linearly or exponentially with respect to these parameters.

Figures 3 and 4 show that, for both beam width and explanation length, the ratio between the runtime of the incremental and non-incremental variants remains approximately constant, indicating similar scaling behavior. However, considering the absolute values reported in Figures 3 and 5, it is clear that, in extreme configurations involving very large beam widths and long explanations, the proposed algorithm may become impractical to apply.

Nevertheless, we argue that these settings are rare in practice. Indeed, it has already been observed that, for complex datasets such as Broden, beam widths larger than 20 become impractical even for

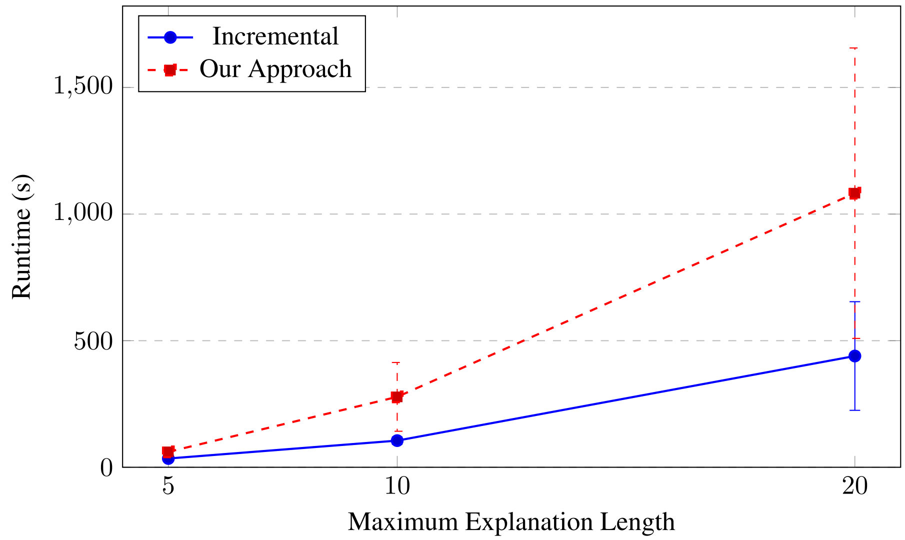

Figure 5: Runtime as a function of the maximum explanation length. Error bars denote standard deviation.

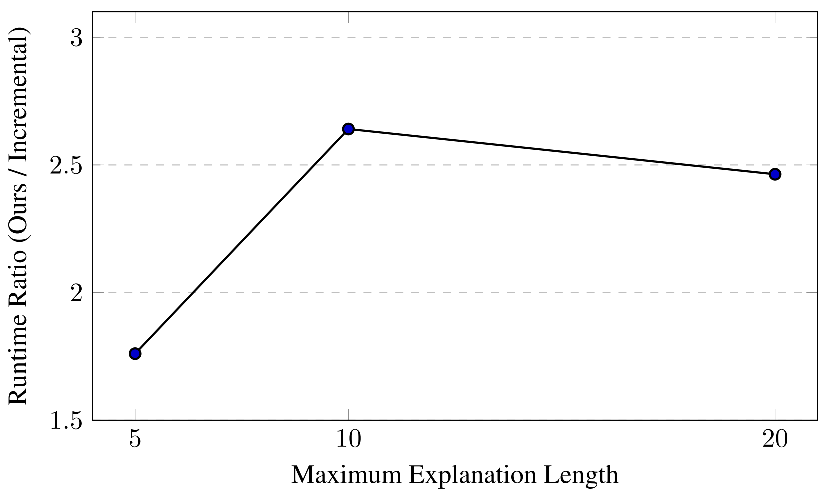

Figure 6: Runtime ratio between our approach and incremental beam search as a function of the maximum explanation length.

incremental beam search, due to the combination of large mask sizes and the high number of closely related formulas in the search space. Still, we believe this limitation is important to highlight.

A natural follow-up question is **whether the gains in explanation quality are simply due to the relatively small beam sizes** typically adopted in prior work, and whether the incremental formulation could recover similar explanations by increasing the beam width. Our experiments suggest that the answer is negative. Table 8 reports a comparison of the explanation metrics when the beam width is fixed to 50 on 50 randomly extracted units. We observe that the gap between the incremental and non-incremental approaches remains consistent with the one observed in the main text, highlighting that the improvements are not tied to a specific beam configuration. These results suggest that the gains are primarily due to the expanded search space enabled by the non-incremental formulation, rather than to differences in beam width alone.

Table 8: Average and standard deviation explanation scores computed by our non-incremental beam search and the standard incremental beam search on BERT using a beam width equal to 50.

Method	Better (%)	IoU	DetAcc	ActCov
Incremental	0	0.144 $\pm$ 0.026	0.190 $\pm$ 0.044	0.407 $\pm$ 0.109
Ours	**69**	**0.154** $\pm$ 0.022	**0.198** $\pm$ 0.037	**0.434** $\pm$ 0.096

E Broader Impact Statement

The contributions of this work are mainly experimental; they are not deployed in downstream applications and refer mainly to algorithmic aspects. Moreover, this work does not alter sensitive components of the explanation pipeline. As such, we believe that this work doesn't improve or degrade the broader impact of explainability methods in this area of research.

F Reproducibility

To ensure full reproducibility, we will release the complete codebase and all scripts required to reproduce the results presented in this paper upon acceptance. In the meantime, this section serves as a brief summary and documentation of the experimental setup used by our framework, along with the resources required.

F.1 Dataset, Models, and Explanations

In this section, we provide the repository, dataset, explanations, and model information, versions, their corresponding licenses, download links, and a brief description of the modifications required to ensure compatibility with our framework.

Probing Datasets

- Broden
  - Accessible at: `https://github.com/CSAILVision/NetDissect`
  - License: MIT
- SNLI v.1.0
  - Accessible at: `https://nlp.stanford.edu/projects/snli/`
  - License: CC-BY-SA

Probed Models

- AlexNet
  - Accessible at: `http://places2.csail.mit.edu/models_places365/alexnet_places365.pth.tar`
  - License: MIT
- ResNet18
  - Accessible at: `http://places2.csail.mit.edu/models_places365/resnet18_places365.pth.tar`
  - License: MIT
- DenseNet161
  - Accessible at: `http://places2.csail.mit.edu/models_places365/whole_densenet161_places365_python36.pth.tar`
  - License: MIT
- BiLSTM
  - Accessible at: `http://downloads.cs.stanford.edu/nlp/data/muj/bowman_snli/6.pth`

– License: MIT

To these models, we add BERT and LLaMA. The pretrained models are respectively accessible at `https://huggingface.co/google-bert/bert-base-uncased` and `https://huggingface.co/knowledgator/Llama-encoder-1.0B`, both associated with Apache License 2.0. For both, we follow the structure proposed in [25]. Specifically, we replace the LSTM encoder with a BERT encoder and a LLaMA-based encoder, respectively, and adjust the input dimension of the MLP according to the hidden size of the encoder. We train BERT and LLaMA using the AdamW optimizer for 5 epochs with a learning rate of $2\mathrm{e}{-5}$, following the training setup included in `https://github.com/jayelm/compexp`. All models converge to an accuracy greater than 80%, which matches the performance of the BiLSTM baseline. We will release the weights of both models upon acceptance.

Explanations Our framework generates explanations through a heuristic search guided by the heuristic proposed in [18]. We reimplemented the heuristic and adapted it to the NLP domain by taking inspiration from the compositional explanation repository at `https://github.com/jayelm/compexp`. Regarding the hyperparameters, we follow the setup proposed in [25], analyzing the top 0.005 quantile of activations, setting the explanation length to 3 (vision) and 5 (nlp), setting the beam size to 5 (vision) and 10 (nlp), and employing the AND, OR, AND NOT operators, and considering the same concepts set defined in [25].

F.2 Resources

The computational resources required by the proposed algorithm are comparable to those of the incremental variant of compositional explanations. In terms of memory, two strategies can be considered: caching the beam masks for functional equivalence checks or recomputing them on demand. When masks are recomputed on demand, the additional memory cost introduced by our modifications is negligible with respect to the overall setup. When masks are cached, the memory footprint increases, approximately doubling the memory required to store a single beam. Note that, in the experiments reported in this paper, we do not observe a large difference in runtime between the two strategies. However, in general settings, recomputing the masks may become a significant bottleneck.

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU, 8 CPU cores, and 64 GB of RAM. The runtimes reported in the main text and in this appendix are based on this setup. However, the experiments can be replicated on significantly lighter configurations.

The overall computational cost depends on (i) the data (i.e., domain), (ii) the number of concepts, and (iii) the probed model. These factors are independent of the proposed method. The data type influences the representation of concept masks, neuron activations, and the information required by the heuristic. In vision settings, these are typically matrices, resulting in a heavier memory footprint (~ 10 GB). In the NLP setting analyzed in this paper, they are vectors, leading to a significantly lighter setup (~ 1 GB). The number of concepts is also related to the data domain: a larger concept set increases memory usage. As a reference, the ~ 10 GB estimate corresponds to a concept dataset with approximately 1200 concepts in the vision domain.

Finally, the probed model impacts the computation only in the preprocessing stage. The model is used to compute neuron activations but is not required during the explanation phase. Therefore, once the activations are computed, the model can be discarded. In this context, the required memory depends on the specific model, ranging from a few MB for the BiLSTM (Bowman) model to approximately ~ 20 GB for LLaMA.

G Visualization of Explanation Differences

This section reports examples of explanations computed by incremental and non-incremental beam search. Specifically, we focus on cases where the explanations differ. Note that **compositional explanations are computed over a much larger set of samples than the one visualized**. The visualizations are provided only to contextualize the highlighted differences between explanations and should not be interpreted as a direct comparison method. More precisely, the images shown do not represent the only and specific samples explained by the labels, but rather a limited representative subset of the state space over which the explanations are computed. As a consequence, due to the sample selection procedure, the visualized samples associated with different explanations or different methods for the same neuron may differ. However, all the visualized samples belong to the activation range considered by both methods.

Finally, as discussed in the main text, the **improvements are not related to visualization properties or human interpretability**, but rather to the problem of identifying the combination of concepts that achieves the highest alignment with the neuron activations.

G.1 Vision

For vision, we follow the setup used in [20]. Specifically, for each visualization, we select the samples (i) associated with the highest intersection between the explanation area and the activation area (to ease interpretation), (ii) where the explanation holds, and (iii) where the neuron is active within the considered activation range. The examples (Figures 7 to 12) are not cherry-picked and correspond to the first six differing explanations for the last convolutional layer of each explained model (ResNet18 [12], AlexNet [16], and DenseNet [14]). We indicate with a check mark whether one of the two explanations is optimal. Note that the $-s$ included in some of the concepts refer to scene concepts, which extend over the whole image.

G.2 Natural Language Inference

For NLP, we follow a setup similar to the one adopted for the vision case. Specifically, for each visualization, we select the samples (i) associated with the highest activations, (ii) for which the explanation holds, and (iii) for which the neuron is active within the considered activation range. Note that this visualization differs from the one used in Figure 1, since we filter out samples for which the explanation is not active. This choice is motivated by the same considerations discussed in the vision setting: simply visualizing the top-k activating samples would often result in examples where none of the explanations hold, since explanations are computed over a much larger sample set and visualizing all of them is impractical.

As in the vision setting, the examples (Figures 13 to 29) are not cherry-picked and correspond to the first six units where the explanations differ for the MLP layer of each analyzed model (BiLSTM [25], BERT [6], and LLaMA [30]).

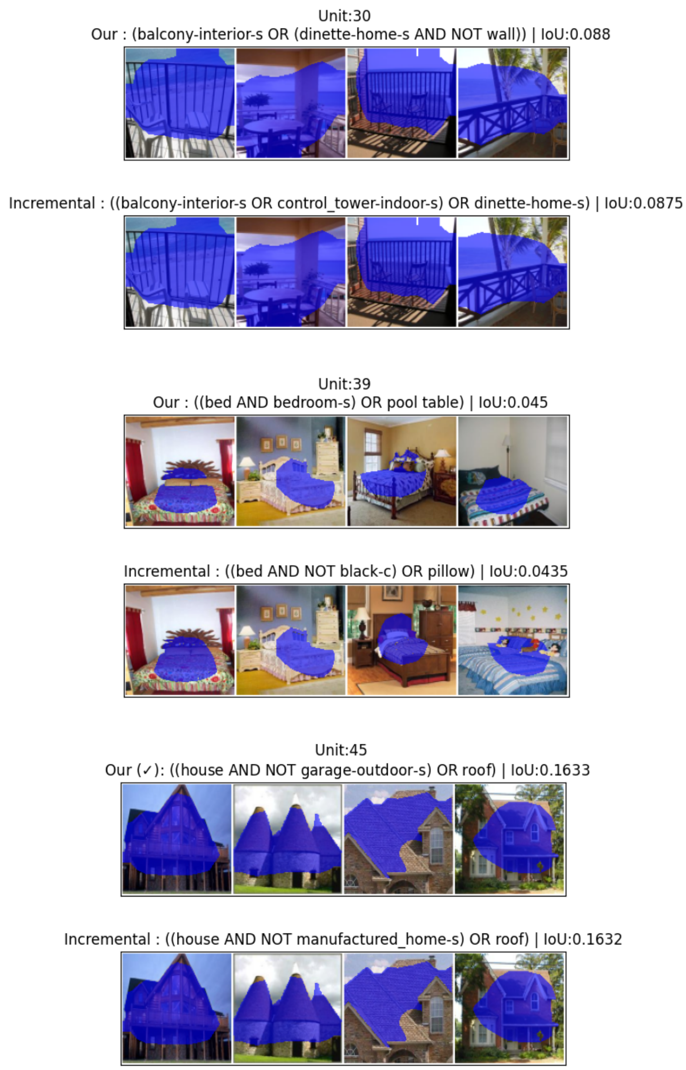

Figure 7: Alignment detected incremental and non-incremental beam-search methods on ResNet18. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

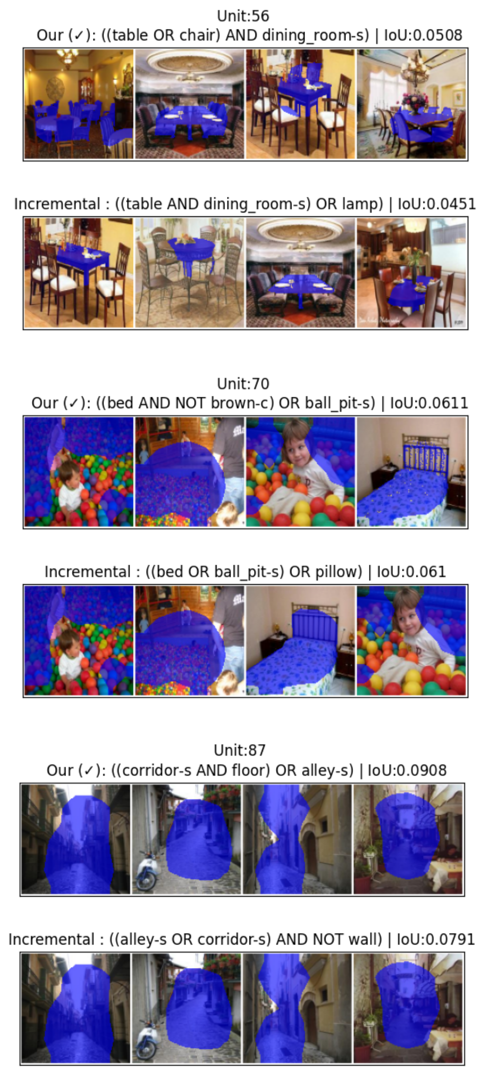

Figure 8: Alignment detected incremental and non-incremental beam-search methods on ResNet18. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

Figure 9: Alignment detected incremental and non-incremental beam-search methods on AlexNet. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

Figure 10: Alignment detected incremental and non-incremental beam-search methods on AlexNet. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

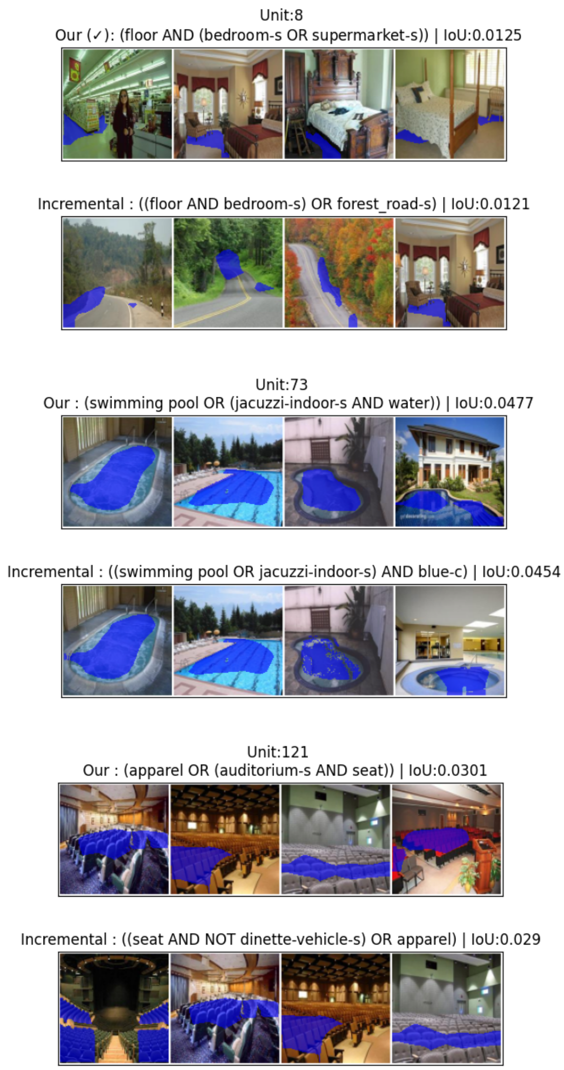

Figure 11: Alignment detected incremental and non-incremental beam-search methods on DenseNet161. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

Figure 12: Alignment detected incremental and non-incremental beam-search methods on DenseNet161. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

Unit 0 – Incremental Beam Search

IoU: 0.136

Explanation

$$((((\texttt{hyp:tok:alone} \lor \texttt{pre:tok:sleeves}) \land \neg\texttt{hyp:tag:rb}) \lor \texttt{hyp:tok:sleeps}) \land \texttt{neighbor:pre:tok:with})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: On a sunny day, in a lush, green field, one small girl inspects a rock while her friend goes off in another direction. **H:** A girl sleeps	🔴
P: A woman dressed in a blue long-sleeved shirt and wearing a hairnet. **H:** The woman is naked and alone in the bathroom.	🟢
P: A girl in a purple sweater and a shoulder bag is looking to the side with a lot of people behind her. **H:** A boy sleeps on the couch.	🟢
P: A man with a black shirt holds a baby while a blue shirted individual faces them. **H:** A man sits all alone on a park bench, arms folded.	🟢
P: Two women dressed for cold weather in jackets and gloves look at something on a cellphone screen. **H:** The woman is alone in her bathroom.	🟢

Unit 0 – Our Non-Incremental Search

IoU: 0.149

Explanation

$$(((\texttt{pre:tok:sleeves} \lor \texttt{hyp:tok:alone}) \lor \texttt{hyp:tok:sleeps}) \land (\texttt{pre:tag:vbn} \lor \texttt{neighbor:pre:tok:individual}))$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A woman dressed in a blue long-sleeved shirt and wearing a hairnet. **H:** The woman is naked and alone in the bathroom.	🟢
P: A band is performing on stage for people. **H:** A band practicing alone in their parents' basement.	🔴
P: A girl in a purple sweater and a shoulder bag is looking to the side with a lot of people behind her. **H:** A boy sleeps on the couch.	🟢
P: A man with a black shirt holds a baby while a blue shirted individual faces them. **H:** A man sits all alone on a park bench, arms folded.	🟢
P: Two women dressed for cold weather in jackets and gloves look at something on a cellphone screen. **H:** The woman is alone in her bathroom.	🟢

Figure 13: Comparison between incremental and non-incremental compositional explanations for Unit 0 in the BiLSTM model.

Unit 15 – Incremental Beam Search

IoU: 0.246

Explanation

$$((((\texttt{hyp:tok:sitting} \land \texttt{neighbor:pre:tok:around}) \land \neg\texttt{pre:tok:sitting}) \land \neg\texttt{pre:tag:rb}) \land \neg\texttt{neighbor:pre:tok:face})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: A female dancer does a center leap across a stage. **H:** The woman is sitting in the audience.	🟢
P: A little girl in pink boots runs down the street. **H:** A little girl in pink boots is sitting in a chair.	🔴
P: A blond woman is holding 2 golf balls while reaching down into a golf hole. **H:** The blond woman is sitting down.	🔴
P: A group of people walking arm and arm across the green grass. **H:** The people are sitting on the ground.	🟢
P: A man jumps over another man doing a back flip. **H:** Two men are sitting down, watching the game.	🟢

Unit 15 – Our Non-Incremental Search

IoU: 0.275

Explanation

$$((\texttt{hyp:tok:sitting} \land ((\texttt{pre:tok:walking} \lor \texttt{neighbor:pre:tok:between}) \lor \texttt{hyp:tok:bench})) \land \neg\texttt{oth:overlap:overlap50})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A man is a safety suit walking outside while another man in a dark suit walks into a building. **H:** Men are sitting in the park.	🔴
P: A female dancer does a center leap across a stage. **H:** The woman is sitting in the audience.	🟢
P: A group of people walking arm and arm across the green grass. **H:** The people are sitting on the ground.	🟢
P: A man jumps over another man doing a back flip. **H:** Two men are sitting down, watching the game.	🟢
P: Three people walking down a hallway at the train station. **H:** A few people are sitting in a dining car of a train drinking coffee.	🟢

Figure 14: Comparison between incremental and non-incremental compositional explanations for Unit 15 in the BiLSTM Model.

Unit 16 – Incremental Beam Search

IoU: 0.129

Explanation

$((((\texttt{neighbor:hyp:tok:trick} \land \texttt{neighbor:hyp:tok:him})$
$\lor \texttt{hyp:tok:eye}) \land \neg\texttt{hyp:tok:man})$
$\lor \texttt{hyp:tok:mcdonald})$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: A guy on inline skates with a white hat is on a yellow rail. **H:** The guy on inline skates is trying to impress his girlfriend.	🟢
P: A young boy wearing safety swimming gear and goggles is in a pool. **H:** A kid uses eye protection while in the water.	🔴
P: A brown dog is trying to retrieve his orange ball that has fallen in a dirt hole. **H:** A dog is trying to get a ball back for his master.	🟢
P: Three men; one pedaling while playing drums, one playing piano and one both pedaling and steering, move a type of mobile band down a street. **H:** Three men are trying to attract a crowd and take them to a bar where they will be playing later.	🟢
P: Several men working to repair a McDonald's sign. **H:** Several men working to repair a McDonald's sign at their place of employment.	🟢

Unit 16 – Our Non-Incremental Search

IoU: 0.164

Explanation

$(((\texttt{neighbor:hyp:tok:trick} \lor \texttt{neighbor:hyp:tok:drummer})$
$\land (\texttt{neighbor:hyp:tok:him} \lor \texttt{neighbor:hyp:tok:after}))$
$\lor \texttt{hyp:tok:mcdonald})$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A guy on inline skates with a white hat is on a yellow rail. **H:** The guy on inline skates is trying to impress his girlfriend.	🟢
P: Children play on a sit-and-spin at the local playground while others slide in the background. **H:** Children are trying to enjoy the outside before a snowstorm.	🔴
P: A brown dog is trying to retrieve his orange ball that has fallen in a dirt hole. **H:** A dog is trying to get a ball back for his master.	🟢
P: Three men; one pedaling while playing drums, one playing piano and one both pedaling and steering, move a type of mobile band down a street. **H:** Three men are trying to attract a crowd and take them to a bar where they will be playing later.	🟢
P: Several men working to repair a McDonald's sign. **H:** Several men working to repair a McDonald's sign at their place of employment.	🟢

Figure 15: Comparison between incremental and non-incremental compositional explanations for Unit 16 in the BiLSTM model.

Unit 17 – Incremental Beam Search

IoU: 0.133

Explanation

$((((\texttt{hyp:tok:girls} \land \neg\texttt{neighbor:pre:tok:cheerleaders})$
$\land \texttt{neighbor:pre:tok:pair}) \lor \texttt{pre:tok:teen})$
$\land \neg\texttt{neighbor:pre:tok:brown})$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Four men are wearing red baseball caps and orange and yellow shirts. **H:** A group of little girls are dancing.	🟢
P: Three Indian boys in shorts play with one another. **H:** Three girls painting their toenails.	🟢
P: This is the image of two Asian men squatting near a mop-pad. **H:** Two Italian girls are squatting.	🟢
P: Three hikers are hiking in a mountain filled with trees and snow. **H:** They are girls there too.	🟢
P: There is a teen boy and a younger boy with the same shirts on in a room full of people. **H:** People are on a class trip.	🟢

Unit 17 – Our Non-Incremental Search

IoU: 0.137

Explanation

$((((\texttt{hyp:tok:girls} \land \neg\texttt{neighbor:pre:tok:cheerleaders})$
$\lor \texttt{hyp:tok:tv}) \land \texttt{neighbor:pre:tok:pair})$
$\lor \texttt{pre:tok:teen})$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Four men are wearing red baseball caps and orange and yellow shirts. **H:** A group of little girls are dancing.	🟢
P: Three Indian boys in shorts play with one another. **H:** Three girls painting their toenails.	🟢
P: This is the image of two Asian men squatting near a mop-pad. **H:** Two Italian girls are squatting.	🟢
P: Three hikers are hiking in a mountain filled with trees and snow. **H:** They are girls there too.	🟢
P: Two kids in winter clothing play near a lakefront. **H:** A man and woman eating TV dinners.	🔴

Figure 16: Comparison between incremental and non-incremental compositional explanations for Unit 17 in the BiLSTM model.

Unit 19 – Incremental Beam Search

IoU: 0.109

Explanation

$$((((\texttt{hyp:tok:boys} \land \neg\texttt{pre:tok:boys}) \lor \texttt{pre:tok:blows}) \land \texttt{oth:overlap:overlap25}) \lor \texttt{hyp:tok:dryer})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: A blond woman talking on her cellphone as her hair blows. **H:** A woman talking on her cellphone.	🟢
P: Two young children in black clothing and padding perform a martial arts match where they compete against the other. **H:** Two young boys compete in martial arts.	🔴
P: Two girls are in line about to order at a Mexican eatery. **H:** Two boys are in line about to order at a Mexican eatery.	🔴
P: Four girls look at objects under a microscope. **H:** Four boys look at objects under a microscope.	🔴
P: A blond woman talking on her cellphone as her hair blows. **H:** A woman talking on her cellphone to her husband.	🟢

Unit 19 – Our Non-Incremental Search

IoU: 0.118

Explanation

$$((((\texttt{pre:tok:blows} \lor \texttt{neighbor:hyp:tok:atv}) \lor \texttt{neighbor:hyp:tok:inflatable}) \lor \texttt{neighbor:hyp:tok:cheer}) \land \texttt{neighbor:hyp:tok:male})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Two women stop between a red smart car and a group of photographers to examine their items in hand. **H:** Everyone pictured is male.	🔴
P: A blond woman talking on her cellphone as her hair blows. **H:** A woman talking on her cellphone.	🟢
P: A girls sports team putting their hands in a circle to do a bonding cheer. **H:** Male cheerleaders talking.	🔴
P: A child is playing with a green, white and red spinning toy that turning on the ground. **H:** A man is playing with a green, white and red spinning toy that turning on the ground.	🔴
P: Someone in a red and white uniform is driving a red and white motorcycle down a track with the front wheel in the air. **H:** A man rides an ATV.	🔴

Figure 17: Comparison between incremental and non-incremental compositional explanations for Unit 19 in the BiLSTM Model.

Unit 22 – Incremental Beam Search

IoU: 0.156

Explanation

$$((((\texttt{pre:tok:dress} \land \neg\texttt{neighbor:hyp:tok:female}) \land \texttt{neighbor:hyp:tok:men}) \lor \texttt{pre:tok:routine}) \land \neg\texttt{pre:tag:vbn})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Woman helps bride into dress. **H:** A man brushes his hair.	🟢
P: A girl in a pumpkin dress sitting at a table. **H:** There is a boy in a pumpkin dress sitting at a table.	🟢
P: A little girl in a red flowered dress with a tennis racket. **H:** A man in red.	🟢
P: A woman in a blue dress is sweeping the floor in a cobblestone road. **H:** A man mops the road.	🟢
P: Group of women in red swimsuits perform a dance routine in the sand at 2008 Beijing Olympics. **H:** The men enjoy dancing in the sand.	🔴

Unit 22 – Our Non-Incremental Search

IoU: 0.190

Explanation

$$(((\texttt{pre:tok:dress} \lor \texttt{neighbor:pre:tok:nun}) \land (\texttt{hyp:tok:man} \lor \texttt{hyp:tok:boy})) \land \neg\texttt{neighbor:hyp:tok:female})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A little girl and lady wearing face paint and costumes are among others holding orange balloons and an event. **H:** A man and his son put camouflage greasepaint on their faces before they go hunting.	🔴
P: A young girl washing a lime-green car. **H:** An old man washing a lime-green car.	🔴
P: Woman helps bride into dress. **H:** A man brushes his hair.	🟢
P: A girl in a pumpkin dress sitting at a table. **H:** There is a boy in a pumpkin dress sitting at a table.	🟢
P: A girl with a purple sweater and a green headband is doing classwork in a classroom. **H:** A boy is scratching his name into his desk.	🔴

Figure 18: Comparison between incremental and non-incremental compositional explanations for Unit 22 in the BiLSTM model.

Unit 0 – Incremental Beam Search

IoU: 0.136

Explanation

((((hyp:tok:dancing $\wedge$ neighbor:pre:tok:white) $\vee$ hyp:tok:everyone) $\vee$ hyp:tok:naked) $\wedge$ $\neg$neighbor:pre:tok:hold)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Four men are wearing red baseball caps and orange and yellow shirts. **H:** A group of little girls are dancing.	🟢
P: A well built black man stands in the subway, listening to headphones. **H:** A man is dancing for tips on the street.	🟢
P: A man in a blue windbreaker jacket, laying face down on a park bench, with a liquor bottle nestled by one of the bench's legs. **H:** A man is dancing at the club popping bottles.	🟢
P: A man and woman and two small children are sitting in the floor with wrapped Christmas presents. **H:** Everyone in the photo are related.	🟢
P: A man and woman hug. **H:** Everyone is standing apart.	🟢

Unit 0 – Our Non-Incremental Search

IoU: 0.141

Explanation

((((hyp:tok:dancing $\vee$ pre:tok:desk) $\wedge$ neighbor:pre:tok:white) $\vee$ hyp:tok:everyone) $\vee$ pre:tok:antique)

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Four men are wearing red baseball caps and orange and yellow shirts. **H:** A group of little girls are dancing.	🟢
P: A well built black man stands in the subway, listening to headphones. **H:** A man is dancing for tips on the street.	🟢
P: A man in a blue windbreaker jacket, laying face down on a park bench, with a liquor bottle nestled by one of the bench's legs. **H:** A man is dancing at the club popping bottles.	🟢
P: A man and woman and two small children are sitting in the floor with wrapped Christmas presents. **H:** Everyone in the photo are related.	🟢
P: A girl wearing a red shirt and jeans is sitting in a rocking chair at a cluttered desk, while another girl wearing a black t-shirt and jogging pants is sitting on the floor writing on a piece of paper with other papers arranged around her. **H:** Two girls flying a kite in the park.	🔴

Figure 19: Comparison between incremental and non-incremental compositional explanations for Unit 0 in the BERT model.

Unit 2 – Incremental Beam Search

IoU: 0.141

Explanation

$$((((\texttt{pre:tok:gloves} \lor \texttt{pre:tok:backs}) \land \texttt{neighbor:pre:tok:shirt}) \lor \texttt{pre:tok:!}) \lor \texttt{pre:tok:denim})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Man with dark shirt in work gloves and a helmet that has a yellow flashlight attached. **H:** A woman handing out water.	●
P: Making something tasty hope you like lemons! **H:** A lady is sewing up a hole in a dress.	●
P: A crowd of people with a lady with a bright orange shirt and another lady in black have their backs to the camera. **H:** Three girls try to buy candy in a grocery store.	●
P: Man with dark shirt in work gloves and a helmet that has a yellow flashlight attached. **H:** An rescue worker looking for injured people.	●
P: Making something tasty hope you like lemons! **H:** A woman is making a lemon meringue pie.	●

Unit 2 – Our Non-Incremental Search

IoU: 0.140

Explanation

$$((((\texttt{pre:tok:you} \lor \texttt{pre:tok:denim}) \land \neg\texttt{pre:tag:vbz}) \lor \texttt{pre:tok:backs}) \lor \texttt{hyp:tok:stay})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Making something tasty hope you like lemons! **H:** A lady is sewing up a hole in a dress.	●
P: A crowd of people with a lady with a bright orange shirt and another lady in black have their backs to the camera. **H:** Three girls try to buy candy in a grocery store.	●
P: Making something tasty hope you like lemons! **H:** A woman is making a lemon meringue pie.	●
P: A small dog wearing a denim miniskirt. **H:** A dog is wearing a skirt.	●
P: A woman walking, smiling and holding a shopping bag. **H:** Some women go shopping, while men stay at home.	●

Figure 20: Comparison between incremental and non-incremental compositional explanations for Unit 2 in the BERT model.

Unit 3 – Incremental Beam Search

IoU: 0.118

Explanation

$$((((\texttt{pre:tok:smile} \lor \texttt{hyp:tok:cheerleaders}) \lor \texttt{pre:tok:pretty}) \lor \texttt{neighbor:hyp:tok:upset}) \land \neg\texttt{hyp:tok:on})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Young brunette woman smiling in front of bushes. **H:** The man looks very upset.	●
P: A woman is smiling at a baby in a stroller in a meadow. **H:** A woman is angry at the baby.	●
P: A young girl in a pretty pink and white blouse and shorts is participating in a soccer game with other children. **H:** A man is making a candy sculpture out of peppermints.	●
P: A girls sports team putting their hands in a circle to do a bonding cheer. **H:** Male cheerleaders talking.	●
P: Two girls smile for the camera. **H:** Three men are smoking cigars.	●

Unit 3 – Our Non-Incremental Search

IoU: 0.120

Explanation

$$((((\texttt{pre:tok:smile} \land \neg\texttt{hyp:tag:nn}) \lor \texttt{hyp:tok:cheerleaders}) \lor \texttt{pre:tok:pretty}) \lor \texttt{neighbor:hyp:tok:upset})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Young brunette woman smiling in front of bushes. **H:** The man looks very upset.	●
P: A woman is smiling at a baby in a stroller in a meadow. **H:** A woman is angry at the baby.	●
P: A young girl in a pretty pink and white blouse and shorts is participating in a soccer game with other children. **H:** A man is making a candy sculpture out of peppermints.	●
P: A girls sports team putting their hands in a circle to do a bonding cheer. **H:** Male cheerleaders talking.	●
P: Two girls smile for the camera. **H:** Three men are smoking cigars.	●

Figure 21: Comparison between incremental and non-incremental compositional explanations for Unit 3 in the BERT model.

Unit 4 – Incremental Beam Search

IoU: 0.161

Explanation

$$((((\texttt{hyp:tok:which} \lor \texttt{hyp:tok:cola}) \lor \texttt{hyp:tok:crossed}) \lor \texttt{pre:tok:cold}) \land \texttt{neighbor:pre:tok:and})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: The side of a building next to a church is painted with a brightly colored Coca-Cola sign. **H:** There is a Coca-Cola sign on a church next to a building.	●
P: Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her. **H:** Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her and waving.	●
P: Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her. **H:** Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her.	●
P: A balding man wearing a brown jacket with his hands crossed and his head tilted down. **H:** A man wearing a brown jacket with his hands crossed and his head tilted down.	●
P: The side of a building next to a church is painted with a brightly colored Coca-Cola sign. **H:** There is a Coca-Cola sign on a building.	●

Unit 4 – Our Non-Incremental Search

IoU: 0.179

Explanation

$$((((\texttt{neighbor:hyp:tok:smiling} \lor \texttt{hyp:tok:jacket}) \land \texttt{oth:overlap:overlap75}) \land \texttt{neighbor:pre:tok:women}) \lor \texttt{hyp:tok:cola})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Man smiling outside in blue shirt. **H:** The man with the shirt is smiling.	●
P: The side of a building next to a church is painted with a brightly colored Coca-Cola sign. **H:** There is a Coca-Cola sign on a church next to a building.	●
P: Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her. **H:** Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her and waving.	●
P: Two girls on an amusement park ride smiling and laughing. **H:** On the amusement park ride there were two girls smiling and laughing.	●
P: Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her. **H:** Two young girls have fun on a trampoline, one of which appears to be falling down onto her back while the other poses with a smile behind her.	●

Figure 22: Comparison between incremental and non-incremental compositional explanations for Unit 4 in the BERT model.

Unit 5 – Incremental Beam Search

IoU: 0.132

Explanation

$$((((\texttt{hyp:tok:swimming} \wedge \texttt{hyp:tok:ocean}) \vee \texttt{pre:tok:crafts}) \vee \texttt{hyp:tok:pictured}) \vee \texttt{hyp:tok:swims})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Two women stop between a red smart car and a group of photographers to examine their items in hand. **H:** Everyone pictured is male.	🔴
P: A woman in a jacket and scarf looking a table that a man in a black shirt is cooking something on. **H:** A man and woman are swimming in the ocean.	🟢
P: People shopping at the marketplace. **H:** A woman is swimming in the ocean.	🟢
P: A young boy reaches for and touches the propeller of a vintage aircraft. **H:** A young boy swims in his pool.	🔴
P: Three women and one man are sitting on a ledge facing away towards a large window. **H:** People are swimming in the ocean.	🟢

Unit 5 – Our Non-Incremental Search

IoU: 0.185

Explanation

$$((\texttt{hyp:tok:swimming} \vee \texttt{hyp:tok:play}) \wedge ((\texttt{neighbor:hyp:tok:shore} \vee \texttt{neighbor:hyp:tok:alone}) \vee \texttt{pre:tok:walking}))$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A woman in a jacket and scarf looking a table that a man in a black shirt is cooking something on. **H:** A man and woman are swimming in the ocean.	🟢
P: Lady vendor selling her goods, one man standing on a post, and two men in a wholesale store. **H:** A dog sits alone next to a swimming pool.	🔴
P: People walking through dirt. **H:** People are swimming.	🔴
P: People shopping at the marketplace. **H:** A woman is swimming in the ocean.	🟢
P: Three women and one man are sitting on a ledge facing away towards a large window. **H:** People are swimming in the ocean.	🟢

Figure 23: Comparison between incremental and non-incremental compositional explanations for Unit 5 in the BERT model.

Unit 1 – Incremental Beam Search

IoU: 0.179

Explanation

$$((((\texttt{hyp:tok:moon} \vee \texttt{hyp:tok:empty}) \vee \texttt{pre:tok:types}) \wedge \neg\texttt{pre:tag:cd}) \vee \texttt{hyp:tok:nothing})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Young blond girl holding chopsticks looks with red eyes at camera while another girl in the background looks at food. **H:** The girl is alone in the middle of the ocean on top of the moon.	🟢
P: There is a room full of pictures all in the wall and a woman in a coat is looking back over her shoulders strangely. **H:** The room is empty.	🟢
P: A young man in a black shirt and an elderly woman in a pink shirt sitting on a train together. **H:** The man is asleep in the bathtub on the moon.	🟢
P: A lady in a business suit sits at a desk and types on her keyboard. **H:** The boy liked cowboys.	🔴
P: Interior scene of fairly full tour bus. **H:** The bus is empty.	🟢

Unit 1 – Our Non-Incremental Search

IoU: 0.185

Explanation

$$(((\texttt{hyp:tok:moon} \vee \texttt{hyp:tok:nothing}) \vee (\texttt{hyp:tok:empty} \wedge \neg\texttt{hyp:tag:in})) \vee \texttt{hyp:tok:straight})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Young blond girl holding chopsticks looks with red eyes at camera while another girl in the background looks at food. **H:** The girl is alone in the middle of the ocean on top of the moon.	🟢
P: There is a room full of pictures all in the wall and a woman in a coat is looking back over her shoulders strangely. **H:** The room is empty.	🟢
P: A young man in a black shirt and an elderly woman in a pink shirt sitting on a train together. **H:** The man is asleep in the bathtub on the moon.	🟢
P: Interior scene of fairly full tour bus. **H:** The bus is empty.	🟢
P: A woman in glittery face-paint and a sequined cap parades with several similarly-dressed individuals in shades of orange and brown. **H:** The woman is swimming alone in her indoor pool on the moon.	🟢

Figure 24: Comparison between incremental and non-incremental compositional explanations for Unit 1 in the LLaMA model.

Unit 2 – Incremental Beam Search

IoU: 0.148

Explanation

$$((((\texttt{hyp:tok:naked} \lor \texttt{hyp:tok:sun}) \land \lnot\texttt{neighbor:pre:tok:man}) \lor \texttt{pre:tok:airport}) \land \lnot\texttt{neighbor:pre:tok:at})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: The tops of a large number of balloons are visible behind a wall, with a large high rise building in the background. **H:** A man and woman watch the sun set over the ocean.	🔴
P: A person wearing a suit is walking through an airport. **H:** A woman sits in the lobby waiting on the doctor.	🟢
P: People dressed in colorful outfits are gathered together. **H:** The man and women walked naked along the secluded beach.	🔴
P: People dressed in costumes walking down a street. **H:** A naked man walking down the street.	🔴
P: Child inside of large snow fort. **H:** A boy sweating in the sun.	🔴

Unit 2 – Our Non-Incremental Search

IoU: 0.164

Explanation

$$((((\texttt{neighbor:pre:tok:snowboarder} \lor \texttt{pre:tok:person}) \land \texttt{neighbor:hyp:tok:man}) \land \lnot\texttt{pre:tag:cc}) \land \lnot\texttt{neighbor:pre:tok:men})$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A person performing a snowboarding trick on a high mountain. **H:** A young man is picking up his girlfriend for a date.	🔴
P: A person is climbing a cliff wall, over a rocky shore, using a rope. **H:** A man is swimming to his favorite beach.	🔴
P: A person in gray snowboarding down a hill. **H:** The snowboarder is a man.	🔴
P: A person wearing a suit is walking through an airport. **H:** A woman sits in the lobby waiting on the doctor.	🟢
P: A person is riding a colorful dirt bike on dirt hills. **H:** A boy plays with his remote control car in his living room.	🔴

Figure 25: Comparison between incremental and non-incremental compositional explanations for Unit 2 in the LLaMA model.

Unit 3 – Incremental Beam Search

IoU: 0.224

Explanation

$$((((\texttt{hyp:tok:picture} \land \texttt{hyp:tag:ex}) \lor \texttt{hyp:tok:humans}) \lor \texttt{hyp:tok:least}) \land \lnot\texttt{neighbor:pre:tok:into})$$

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Sidewalk vendor dressed in white shirt brown pants sells grilled corn. **H:** There is a person in this picture.	🟢
P: A group of people in costumes are holding their arms in the air in the manner of the finale of a musical number. **H:** There are some humans wearing outfits.	🟢
P: Seagulls are flying above the crashing waves. **H:** There are flying animals in this picture.	🟢
P: Many people have painted faces at night. **H:** There are at least 2 people.	🟢
P: Two people sit facing away in a downtown scene with a motorcycle parked in front of a pool. **H:** There are 2 people in this picture, as well as a two-wheeled vehicle.	🟢

Unit 3 – Our Non-Incremental Search

IoU: 0.236

Explanation

$$((\texttt{hyp:tok:least} \lor (\texttt{hyp:tok:humans} \land \texttt{pre:tag:jj})) \lor (\texttt{hyp:tok:picture} \land \texttt{hyp:tag:ex}))$$

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Sidewalk vendor dressed in white shirt brown pants sells grilled corn. **H:** There is a person in this picture.	🟢
P: A group of people in costumes are holding their arms in the air in the manner of the finale of a musical number. **H:** There are some humans wearing outfits.	🟢
P: Seagulls are flying above the crashing waves. **H:** There are flying animals in this picture.	🟢
P: Many people have painted faces at night. **H:** There are at least 2 people.	🟢
P: Two people sit facing away in a downtown scene with a motorcycle parked in front of a pool. **H:** There are 2 people in this picture, as well as a two-wheeled vehicle.	🟢

Figure 26: Comparison between incremental and non-incremental compositional explanations for Unit 3 in the LLaMA model.

Unit 4 – Incremental Beam Search

IoU: 0.159

Explanation

((((pre:tok:running ∧ ¬neighbor:hyp:tok:run)
∧ neighbor:hyp:tok:women) ∨ pre:tok:skydiving)
∨ hyp:tok:cook)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: A group of grimacing males is running on green grass, several of them are wearing Nike sneakers. **H:** The men are all wearing sandals.	🔴
P: A child and small dog running. **H:** A man eats a carrot.	🔴
P: The man in the black and white shirt is running with a ball. **H:** The man in the black and white shirt is sitting with a book.	🟢
P: Two yellow dogs are running side by side down a road. **H:** The 2 dogs cook pizza.	🔴
P: A girl in uniform running through the water fountain gushing water from the tiles. **H:** A girls is happy.	🟢

Unit 4 – Our Non-Incremental Search

IoU: 0.183

Explanation

(((pre:tok:running ∨ hyp:tok:boots)
∧ (neighbor:hyp:tok:boys ∨ hyp:tok:sitting))
∧ ¬neighbor:hyp:tok:runner)

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A girl in green is running behind a boy in red in a sports arena whilst a girl in pink is running in the background. **H:** Some kids sit on a rollercoaster.	🔴
P: A little girl in pink boots runs down the street. **H:** A little girl in pink boots is sitting in a chair.	🔴
P: Girl with helmet riding a brown horse. **H:** The girl is wearing tall boots.	🔴
P: The man in the black and white shirt is running with a ball. **H:** The man in the black and white shirt is sitting with a book.	🟢
P: A girl leaps through the air while running on the riverbank. **H:** The girl is wearing pants.	🔴

Figure 27: Comparison between incremental and non-incremental compositional explanations for Unit 4 in the LLaMA model.

Unit 8 – Incremental Beam Search

IoU: 0.135

Explanation

((((pre:tok:swimming ∨ pre:tok:wave)
∧ ¬oth:overlap:overlap50) ∧ ¬neighbor:hyp:tok:others)
∧ ¬neighbor:hyp:tok:playground)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: Five young boys in swim trunks jump into a swimming pool at the same time. **H:** A group of young boys studying at the library.	🔴
P: A person surfs through a wave as it crashes down. **H:** A person is outdoors.	🟢
P: A lady wearing a red and white polka dot swim cap is swimming in a pool. **H:** Woman sits on the sidelines of a basketball game.	🟢
P: A person surfs through a wave as it crashes down. **H:** A person is lying on the beach.	🟢
P: A brown dog and a black dog in the edge of the ocean with a wave under them boats are on the water in the background. **H:** The pets are sleeping on the grass.	🟢

Unit 8 – Our Non-Incremental Search

IoU: 0.159

Explanation

(((pre:tok:swimming ∧ ¬oth:overlap:overlap25)
∧ ¬pre:tag:rb)
∨ (pre:tok:wave ∧ pre:tag:prp))

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A person surfs through a wave as it crashes down. **H:** A person is outdoors.	🟢
P: A lady wearing a red and white polka dot swim cap is swimming in a pool. **H:** Woman sits on the sidelines of a basketball game.	🟢
P: A person surfs through a wave as it crashes down. **H:** A person is lying on the beach.	🟢
P: A brown dog and a black dog in the edge of the ocean with a wave under them boats are on the water in the background. **H:** The pets are sleeping on the grass.	🟢
P: Two little girls on a hot summer day playing in a swimming pool. **H:** The girl is alone in her bedroom asleep.	🟢

Figure 28: Comparison between incremental and non-incremental compositional explanations for Unit 8 in the LLaMA model.

Unit 9 – Incremental Beam Search

IoU: 0.342

Explanation

$$((((\texttt{hyp:tok:sleeping} \lor \texttt{hyp:tok:asleep}) \land \texttt{neighbor:hyp:tok:in}) \lor \texttt{hyp:tok:sleeps}) \land \neg\texttt{neighbor:hyp:tok:women})$$

Samples selected by this explanation

Premise / Hypothesis	Ours	Applies
P: A woman in a white hat and red dress holding hands with a man in a black jacket with a flower in his upper pocket. **H:** The woman is alone and asleep in the car on the moon.		●
P: Woman standing on a ladder fixing letters on a theater sign, while one woman watches and one man stands nearby. **H:** Woman sleeps on a couch.		●
P: The Raiders complete the pass while the Dolphin defense scrambles to catch him. **H:** The player is alone asleep in the bathtub.		●
P: Nine women in white robes with hoods walk on plush, green grass. **H:** The woman is alone and asleep in her bedroom.		●
P: A little kid is jumping off a brick balcony. **H:** A child is sleeping in bed.		●

Unit 9 – Our Non-Incremental Search

IoU: 0.342

Explanation

$$(((\texttt{hyp:tok:sleeping} \land \texttt{neighbor:pre:tok:kid}) \lor (\texttt{hyp:tok:asleep} \land \texttt{neighbor:hyp:tok:in})) \lor \texttt{hyp:tok:sleeps})$$

Samples selected by this explanation

Premise / Hypothesis	Inc.	Applies
P: A woman in a white hat and red dress holding hands with a man in a black jacket with a flower in his upper pocket. **H:** The woman is alone and asleep in the car on the moon.		●
P: Woman standing on a ladder fixing letters on a theater sign, while one woman watches and one man stands nearby. **H:** Woman sleeps on a couch.		●
P: The Raiders complete the pass while the Dolphin defense scrambles to catch him. **H:** The player is alone asleep in the bathtub.		●
P: Nine women in white robes with hoods walk on plush, green grass. **H:** The woman is alone and asleep in her bedroom.		●
P: A little kid is jumping off a brick balcony. **H:** A child is sleeping in bed.		●

Figure 29: Comparison between incremental and non-incremental compositional explanations for Unit 9 in the LLaMA model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim that our algorithm "finds explanations expressing a higher alignment than the incremental formulation while maintaining comparable runtime, and leads to a significant increase in the number of optimal explanations when these can be computed". The claim about higher alignment is supported by experiments in Table 1, while the ones about optimality are supported and demonstrated in Table 2. The comparable time claim is demonstrated in the same tables and in Section A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: the paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We will release the code upon acceptance to ensure full reproducibility, including all the scripts required to replicate the experiments reported in the paper. In addition, we provide comprehensive details for reproducing the results across multiple sections: Section 4, Section A, and Section F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We are not providing the data and code in the review phase, but **we are committed to releasing and providing open access to the full code and scripts to faithfully reproduce all the experimental results upon acceptance**.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe all the experimental details in Section 4 in the main text, and in Section F in the appendix. Our algorithm does not require training, and the hyperparameters for both the models and explanations are adopted from prior literature, as explained in the previously mentioned sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our proposed algorithm is deterministic and does not involve any training. However, we report and discuss the significance of the standard deviation for all experiments in Section A.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the time needed to run individual configurations in every experiment and discusses the resources needed to run all the experiments in section Section F.2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper and the research conducted in this paper conform with the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The contributions of this work are mainly experimental and refer mainly to algorithmic aspects. As such, we believe that this work doesn't improve or degrade the broader impact of explainability methods in this area of research. We discuss the broader impact in Section E.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: the paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We report in Section F the relevant paper citations, licenses, versions (if applicable), terms of use, and links to the corresponding code repositories. Citations are also included in the main text and in the relevant appendix sections.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.