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 remaining empirically tractable. Our approach introduces a non-incremental beam expansion strategy and a functional-equivalence-aware beam search that leverages equivalence classes of formulas to reduce redundant beam entries and improve search quality. 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. Additionally, these explanations exhibit higher perceived descriptive power while preserving understandability. Overall, these results show that expanding search beyond incremental formulations leads to more expressive and accurate explanations.

1 Introduction

Compositional explanations [26] 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 adopts strong assumptions on the formula structure and leverages beam search to restrict the exploration to a subset of the state space [26, 12, 23, 22, 21]. 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 remaining empirically tractable. Our approach

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

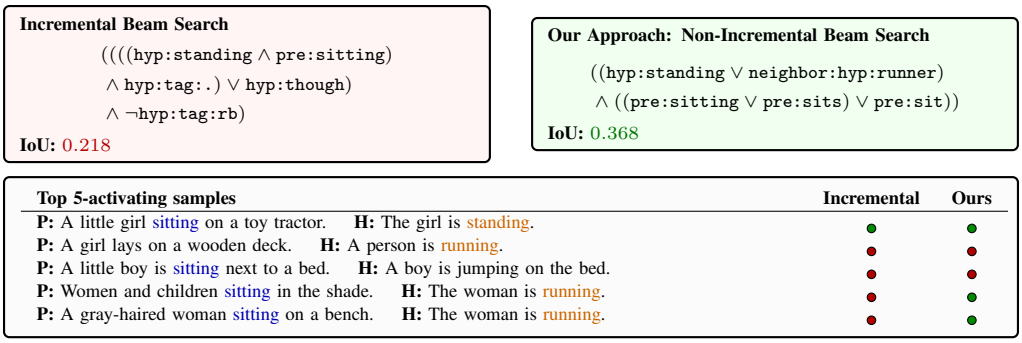

Figure 1: An example of explanations computed by incremental beam search and our non-incremental approach for a neuron that recognizes contradictions between sitting actions and standing or moving actions.

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 prevent redundant formulas from occupying multiple beam positions, thus increasing the diversity of beam candidates. We evaluate the proposed framework on both natural language inference and image classification tasks across six 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 improves explanation alignment across natural language and vision tasks, more frequently recovers optimal explanations, and does not introduce a loss in understandability while improving perceived descriptive power.

The code associated with this paper is available at https://github.com/aiea-lab/nonincremental_ce

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 [24], such as approaches that generate representative samples that maximize neuron activations [9, 30, 27], methods that associate neurons with concepts or textual descriptions through correlation-based analyses and foundational models [14, 28, 29, 16, 1, 34, 11], and approaches that study internal representations through probing or sparse decompositions [8, 10, 6].

Among these, compositional explanations [26] represent a specific family of methods that focus on the **alignment between a neuron activation range and the locations of concepts**, and that express this relationship through propositional logic formulas. The seminal work in this area is Network Dissection [3, 4], which associates each neuron with the single concept that maximizes alignment. This approach was extended by compositional explanations [26], which 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 [3], exclusion [26], relative position [12], and hierarchy [23], to capture behavior across multiple activation ranges [21], to broaden its applicability [19], and to provide formal guarantees about the optimality of the explanations [20].

This paper targets the incremental construction assumption of compositional explanation methods. 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 addresses this limitation by proposing a non-incremental search framework that expands the beam state space while remaining empirically tractable, and by explicitly modeling functional equivalence between formulas during the search. Our proposal shares some conceptual ideas with literature on combinatorial search over symbolic expressions. Similarly to our algorithm, in Inductive Logic Programming, logical hypotheses are explored through refinement operators [33] that progressively transform candidate clauses, while program synthesis methods can prune candidate programs that behave identically on the available examples through observational equivalence [2]. Related ideas also appear in genetic programming, where larger expressions are constructed by recombining previously generated subexpressions [17]. These approaches share with our work the need to search a large structured space while avoiding redundant candidates, but address different settings. Our contribution applies these ideas to the interpretability setting through the concepts of recombination and functional equivalence.

3 Method

This section introduces the compositional explanation problem, the notation, and the 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. For each sample $x \in \mathcal{D}$, let Ω_x denote the set of locations and $\Omega = \bigcup_{x \in \mathcal{D}} \Omega_x$. A location may correspond to the whole sample or to a specific element within the sample, depending on the considered setting. 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 $\mathbf{M} = \{M_k\}_{k=1}^{p}$ 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 $\wedge\neg$, where $\wedge\neg$ denotes conjunction with negation applied to an atomic concept. 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 [26] or by using clustering techniques [21]. When the dimensions of the activation space and input differ, we follow [26] and apply interpolation to project the activations into Ω. Let $\mathfrak{L}^n$ denote the set of admissible logical formulas of length at most n constructed from $\mathfrak{L}^1$, and let $\mathcal{C}(L)$ denote the set of atomic concepts appearing in the label L. 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:

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

Finally, let $L = A \wedge \neg c$, where $c \in \mathfrak{L}^1$. The occurrence of $\neg c$ is *unverified* if $\circ_{A \wedge \neg c}(\mathbf{M}) = \circ_A(\mathbf{M})$. A label is *unverified* if it contains at least one unverified negation.

The goal of compositional explanations is to identify

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

where the IoU score is defined as

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

and measures the overlap between the neuron activation mask and the mask induced by a candidate explanation, normalized by their union. The higher values indicate stronger alignment between the neuron activations and the explanation.

Given the IoU objective and the fixed admissible space $\mathfrak{L}^n$, any L^* satisfying Equation (2) is an *optimal solution* [20]. Because computing such a solution is often impractical or prohibitively expensive [26], previous methods restrict the search space by two assumptions: each concept may appear at most once in a label, and 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 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. Our method addresses this limitation by expanding the beam search with additional non-incremental candidates and by explicitly accounting for functional equivalence among labels.

Running example As a running example throughout this and the following sections, consider a probing dataset containing five sentences: $\mathfrak{D} = \{x_1 :$ A funny child is sleeping on the couch., $x_2 :$ A funny man is standing in a quiet room., $x_3 :$ A dog is sleeping outdoors., $x_4 :$ A child is sleeping outdoors in a quiet park., $x_5 :$ A woman is sleeping on the sofa.$\}$. Consider the atomic concepts $F = \texttt{funny}$, $Q = \texttt{quiet}$, $S = \texttt{sleeping}$, and $O = \texttt{outdoors}$. Their concept masks are:

$$M_F = [1, 1, 0, 0, 0] \quad M_Q = [0, 1, 0, 1, 0] \quad M_S = [1, 0, 1, 1, 1] \quad M_O = [0, 0, 1, 1, 0]$$

since, for example, `funny` appears in x_1 and x_2 but not in the remaining samples. Similarly, if the neuron activates within the considered range for x_1 and x_3, we can write the activation mask as:

$$N = [1, 0, 1, 0, 0]$$

In this example, a formula such as $L_{e_1} = (F \vee Q) \wedge S$ induces a mask $\circ_{L_{e_1}}(\mathbf{M}) = [1, 0, 0, 1, 0]$, and therefore an alignment of $IoU(N, \circ_{L_{e_1}}(\mathbf{M})) = 1/3 \approx 0.33$.

Additionally, we provide a visual overview of the full pipeline in Figure 3 in the appendix.

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 $\circ_{L_1}(\mathbf{M}) = \circ_{L_2}(\mathbf{M})$ or, equivalently, $IoU(\circ_{L_1}(\mathbf{M}), \circ_{L_2}(\mathbf{M})) = 1$.

Functional equivalence partitions $\mathfrak{L}^n$ into equivalence classes. We distinguish three mutually exhaustive cases: logical equivalence (Type 1), dataset equivalence (Type 2), and dataset ambiguity (Type 3). The first two arise between formulas sharing the same concepts, while the third involves formulas differing in their concepts.

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). For example, in the running example, consider $L_{e_2} = F \vee (Q \wedge S)$ and $L_{e_3} = (F \wedge S) \vee Q$. The two labels contain the same concepts but are not logically equivalent. However, both induce the mask $[1, 1, 0, 1, 0]$ on the probing dataset, and are therefore Type-2 equivalent.

Type 3: Dataset Ambiguity.

$$L_1 \equiv_{\text{amb}} L_2 \iff \circ_{L_1}(\mathbf{M}) = \circ_{L_2}(\mathbf{M}) \ \wedge\ \mathcal{C}(L_1) \neq \mathcal{C}(L_2)$$

Unlike Type 2, this case involves different concept sets and therefore captures ambiguity induced by the probing dataset. This ambiguity can be global, when different concepts or formulas induce the same mask over the dataset, or context-dependent, when they become indistinguishable only within a shared subformula. For example, in the running example, O and $S \wedge O$ are functionally equivalent since $\circ_O(\mathbf{M}) = M_O = [0, 0, 1, 1, 0] = M_S \cap M_O = \circ_{S \wedge O}(\mathbf{M})$.

3.2 Non-Incremental Algorithm

To move beyond incremental approaches, we propose 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. For example, in the running example, $B_{\leq k}$ may contain $L_{e_1} = (F \vee Q) \wedge S \in B_3$, $L_{e_4} = F \vee Q \in B_2$, and $L_{e_5} = S \wedge O \in B_2$. 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

$$L_{\text{inc}} = L \oplus c \tag{4}$$

For example, incrementally adding O to L_{e_1} through $\wedge$ produces $C_{\text{inc}} = ((F \vee Q) \wedge S) \wedge O$, whose induced mask in the running example is $[0, 0, 0, 1, 0]$.

Non-incremental candidates. The second family changes the scope of a newly added concept by inserting it 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. For each $S \in \mathcal{S}_{\oplus}(L)$, the algorithm produces a new label

$$L_{\text{ni}} = L[S \leftarrow S \oplus c] \tag{5}$$

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. For example, inserting O into the internal disjunctive clause $F \vee Q$ in L_{e_1} produces $C_{\text{ni}} = (F \vee Q \vee O) \wedge S$, with induced mask $[1, 0, 1, 1, 0]$. Thus, the incremental and internal expansions of the same label produce different masks and different IoUs with N.

Pairwise recombination candidates. The third family provides a limited form of backtracking by recombining labels selected for expansion at previous beam levels when the resulting label has length equal to the current level. 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:

$$L_1 \vee L_2 \qquad \text{and} \qquad L_1 \wedge L_2 \tag{6}$$

We do not generate $L_1 \wedge \neg L_2$ in this step because compositional explanations assume that negation is applied only to atomic concepts. For example, since in the running example $L_{e_4} = F \vee Q$ and $L_{e_5} = S \wedge O$ have disjoint concept sets, their recombination produces $C_{\text{rec}_1} = (F \vee Q) \wedge (S \wedge O)$ and $C_{\text{rec}_2} = (F \vee Q) \vee (S \wedge O)$.

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 $\mathcal{C}_{\text{comm}} = \mathcal{C}(L_1) \cap \mathcal{C}(L_2)$ and $\mathcal{C}_{\text{diff}} = \mathcal{C}(L_1) \triangle \mathcal{C}(L_2)$, where $\triangle$ denotes symmetric difference. Let L_{comm} be the clause obtained by chaining the concepts in $\mathcal{C}_{\text{comm}}$ using $\oplus_s$. We construct two clauses from $\mathcal{C}_{\text{diff}}$: L^s_{diff}, obtained by chaining its concepts using $\oplus_s$, and L^c_{diff}, obtained using the dual operator $\widehat{\oplus_s}$, where $\widehat{\wedge} = \vee$ and $\widehat{\vee} = \wedge$. The algorithm generates:

$$L_{\text{comm}} \oplus_s L^s_{\text{diff}} \qquad \text{and} \qquad L_{\text{comm}} \oplus_s L^c_{\text{diff}} \tag{7}$$

For example, in our running example, consider the two disjunctive labels $L_{e_4} = F \vee Q$ and $L_{e_6} = F \vee S$. They share the concept F, while Q and S form the symmetric difference. Rule 2 therefore generates $F \vee (Q \vee S)$ and $F \vee (Q \wedge S)$. The two resulting masks are $[1, 1, 1, 1, 1]$ and $[1, 1, 0, 1, 0]$, respectively.

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, considering $L_{e_4} = F \vee Q$ and the conjunctive label $L_{e_7} = F \wedge S$, this rule specializes F in the disjunctive label with $F \wedge S$, producing $(F \wedge S) \vee Q = L_{e_3}$, which induces the mask $[1, 1, 0, 1, 0]$ in the running example.

Overall, the candidate set at level $k+1$ is given by:

$$\mathcal{BC}_{k+1} = \mathcal{E}_{\text{inc}}(B_k) \cup \mathcal{E}_{\text{ni}}(B_k) \cup \mathcal{R}_{k+1}(B_{\leq k}) \tag{8}$$

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 length $k+1$ obtained through pairwise recombination. The complete pseudo-code of the beam expansion is included in the appendix in Algorithm 1. We also provide a visual overview of the full pipeline in the appendix in Figure 3.

3.2.2 Functional-Equivalence-Aware Beam Search

Our proposed algorithm leverages heuristics [21, 20] to guide a beam search algorithm, consistent with prior research [26]. The candidate labels in $\mathcal{BC}_{k+1}$ are ranked according to an admissible heuristic estimate (i.e., upper bound) of their alignment with the neuron activation mask. As a heuristic, we use the one proposed by [20], which derives the upper bound from properties of the atomic concept and activation masks. Specifically, it distinguishes locations that are unique to one concept from those shared by multiple concepts, and whether these locations fall inside or outside the neuron activation mask. It then combines optimistic bounds on these quantities to overestimate the intersection and underestimate the contribution to the union, yielding an admissible upper bound (see Section E for more details about the heuristic). Let $h(L)$ denote the heuristic score associated with a label L, and let us assume $\mathcal{BC}_{k+1}$ is sorted in decreasing order of $h(L)$. The algorithm scans $\mathcal{BC}_{k+1}$ progressively and computes the exact IoU only for candidates whose upper bound is sufficiently high to enter the beam.

Let B_{k+1} denote the current 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 $|B_{k+1}| < b$ or $IoU\big(N, \circ_{L_j}(\mathbf{M})\big) \geq IoU\big(N, \circ_{L_{\min}}(\mathbf{M})\big)$. 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. In this case L_j is inserted into the beam and, if the beam is full, the lowest-scoring label is removed:

$$B_{k+1} \leftarrow \begin{cases} B_{k+1} \cup \{L_j\} & |B_{k+1}| < b \\ \big(B_{k+1} \setminus \{L_{\min}\}\big) \cup \{L_j\} & |B_{k+1}| = b \end{cases} \tag{9}$$

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 $\mathcal{C}(L_j) = \mathcal{C}(L_t)$ 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:

$$B_{k+1} \leftarrow \big(B_{k+1} \setminus \{L_t\}\big) \cup \{L_j\} \tag{10}$$

For example, in the running example, C_{inc} and C_{rec_1} are logically equivalent: $((F \vee Q) \wedge S) \wedge O \equiv (F \vee Q) \wedge (S \wedge O)$. Both therefore induce $[0, 0, 0, 1, 0]$. Once one has entered B_{k+1}, the other is identified as Type-1 equivalent and discarded.

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. In this case, the algorithm (i) inserts L_j into the beam, (ii) removes $L_{\min}$ if the beam is full,

$$B_{k+1} \leftarrow \begin{cases} B_{k+1} \cup \{L_j\} & |B_{k+1}| < b \\ \big(B_{k+1} \setminus \{L_{\min}\}\big) \cup \{L_j\} & |B_{k+1}| = b \end{cases} \tag{11}$$

and (iii) augments the next candidate sets by recombining L_j and L_t:

$$\mathcal{BC}_t \leftarrow \mathcal{BC}_t \cup \mathcal{R}_t(\{L_j, L_t\}) \tag{12}$$

where t is the length of the formula obtained by the recombination. For example, in the running example, consider $L_t = F \vee S$ and $L_j = Q \vee S$ at the same beam level. Both induce the mask $[1, 1, 1, 1, 1]$, but $\mathcal{C}(L_t) = \{F, S\} \neq \{Q, S\} = \mathcal{C}(L_j)$, and are therefore Type-3 equivalent. Unlike in Type-1/Type-2 equivalence, L_j is not discarded as redundant. Since the two labels share S, their recombination generates $S \vee (F \vee Q)$ and $S \vee (F \wedge Q)$, which are added as candidates for the next beam level.

The process terminates when either all candidates in $\mathcal{BC}_{k+1}$ have been processed or when the next candidate $L_j \in \mathcal{BC}_{k+1}$ is associated with an estimate $h(L_j) < IoU\big(N, \circ_{L_{\min}}(\mathbf{M})\big)$ and the beam is full. Since $h(L_j)$ upper-bounds the IoU and the remaining candidates have equal or lower heuristic scores, none of them can obtain an IoU sufficiently high to enter the beam. The current beam is therefore final. The pseudo-code of the non-incremental beam search is included in the appendix in Algorithm 2.

4 Experiments

In this section, we evaluate the proposed 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.

Setup In this paper, we follow the most common setup of compositional explanations. Specifically, we apply our framework across three models trained on Places365 [36] for image classification (ResNet18 [13], AlexNet [18], and DenseNet161 [15]) and three models trained or finetuned on natural language inference tasks (BiLSTM, BERT [7], and TinyLLaMA 1.1B [35]). These models correspond to those investigated in previous literature [26, 31]. As probing datasets, we use Broden [3] for image classification, in line with prior work, and SNLI [5] for the language models. For image classification, the concepts correspond to the segmentation masks annotated in the dataset. In the NLI task, we follow [26] 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 [25]. Following previous approaches, we probe all the neurons in the penultimate layer of each model (the last MLP layer for transformer architectures), define the neuron activation mask $N \subseteq \Omega$ by selecting the top 0.005 quantile of activations, and discard neurons that activate less than 0.5% of the time.

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 [26, 21, 19], the stopping criterion [22], and the concept sets or relations considered [12, 23]. 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.

Evaluation Metrics To evaluate our proposal, we consider the following metrics, commonly adopted in this area [26, 22, 4, 21]: IoU, detection accuracy, activation coverage, visited nodes, estimated nodes, and seconds per unit. IoU measures the alignment between the neuron activation mask N and the label mask $\circ_L(\mathbf{M})$. Detection accuracy measures the fraction of concept annotations associated with a label that are recovered within the activation mask N. Activation coverage measures the fraction of neuron activations explained by a label. *Estimated nodes* are candidates for which the heuristic bound $h(L)$ is computed, whereas *visited nodes* are candidates for which the exact IoU is evaluated. For further details on these metrics, we refer to Section B. Finally, we report the value Better (%) as the percentage of neurons for which a beam-based method returns an explanation associated with a strictly higher IoU than the other beam-based method, since the two algorithms explore different beam trees and neither strictly dominates the other.

Table 1: Average scores for explanation and search metrics associated with the computation of explanations in the NLP tasks by non-incremental and incremental beam search. Explanation scores are averaged over units where the search solutions differ.

Method	Explanation Quality				Search Metrics		
	Better (%)	IoU	DetAcc	ActCov	Visited	Esti.	s/u
			BiLSTM				
Incremental	3	0.171	0.477	0.222	40	370246	29
Ours	58	**0.191**	**0.481**	**0.248**	41	576139	49
			BERT				
Incremental	4	0.144	0.414	0.189	40	369144	27
Ours	57	**0.156**	**0.435**	**0.202**	42	724188	55
			LLaMA				
Incremental	4	0.185	0.466	0.244	40	366149	28
Ours	67	**0.202**	**0.476**	**0.267**	41	651615	54

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 [26], we set the maximum explanation length to 5 and the beam width to 10.

Table 1 shows the average search and explanation scores (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). 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 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 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 [26, 21], 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 A.3 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

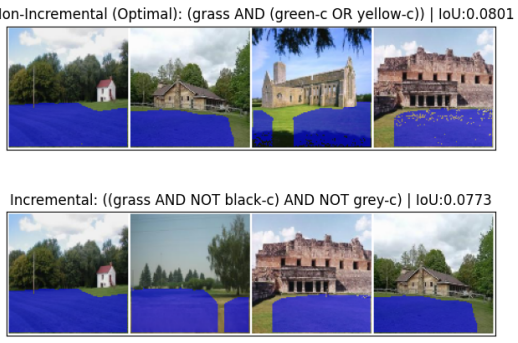

Figure 2: An example of a case where our algorithm reaches optimality while the incremental one does not (Unit 91 of AlexNet). The blue area indicates the activation captured by the explanation.

Table 2: Average scores for explanation and search metrics associated with the computation of explanations in the vision domain by our non-incremental beam search, the standard incremental beam search, and the optimal algorithm. Explanation scores are averaged over units where the search solutions differ.

Method	Explanation Quality					Search Metrics		
	Optimal	**Better (%)**	**IoU**	**DetAcc**	**ActCov**	**Visited**	**Esti.**	**s/u**
				ResNet18				
Incremental	0.93	0	0.073	0.133	0.188	28	32621	67
Ours	**0.96**	**5**	0.076	0.137	0.181	71	48142	73
Optimal	1.00	–	0.077	0.139	0.180	–	–	> 3600
				AlexNet				
Incremental	0.78	0	0.049	0.082	0.155	57	34086	70
Ours	**0.92**	**15**	0.052	0.084	0.151	398	46866	130
Optimal	1.00	–	0.052	0.084	0.155	–	–	> 3600
				DenseNet161				
Incremental	0.76	0	0.046	0.088	0.115	50	36810	67
Ours	**0.90**	**14**	0.048	0.099	0.109	267	51130	108
Optimal	1.00	–	0.049	0.100	0.105	–	–	> 3600

minutes instead of 1-3 hours). Therefore, we compare the two beam-search approaches both in terms of explanation quality and their ability to recover the optimal solution, and we also report for reference the explanation quality obtained by the optimal solutions retrieved using the exact non-beam method [20]. Note that the optimal method maximizes (and is defined for) exclusively the IoU.

Table 2 shows the average scores (the standard deviation is reported in Section A). As expected, the guaranteed-optimal method provides the highest IoU, but at substantially higher computational cost. 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 [20]. 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 I).

Regarding runtime, the time required to compute explanations remains close to that of the incremental variant. Although the average is higher, the median time is similar (Section A), suggesting that our algorithm is more sensitive to outliers.

4.3 Human Evaluation

To evaluate whether the additional structural complexity introduced by non-incremental explanations affects their interpretability, we conducted a human study in the NLP setting, where the two formulations can produce substantially different logical structures. We recruited 65 qualified participants, each evaluating six pairs of incremental and non-incremental explanations, for a total of 390 judgments per criterion. In each comparison, participants were shown eight highly activating samples and two anonymized explanations, with their order randomized and balanced. Participants evaluated *understandability*, i.e., which explanation was easier to understand as a description of when the unit activates, and *descriptive power*, i.e., which explanation better described the patterns associated with the unit's activations. Further study details are reported in Section C.

For understandability, 45.4% of judgments rated the formulations as equivalent, 30.0% preferred the non-incremental explanation, and 24.6% preferred the incremental one. At the participant level, we

observed no significant directional preference (two-sided exact sign test, $p = .488$). A non-inferiority analysis with a 10% margin rejects a loss in understandability for the non-incremental formulation ($p < .001$).

For descriptive power, 42.8% of judgments preferred the non-incremental formulation, 24.9% rated the formulations as equivalent, and 32.3% preferred the incremental one. At the participant level, the preference for the non-incremental formulation was statistically significant (two-sided exact sign test, $p = .018$). Overall, these results provide evidence that the proposed non-incremental structure does not introduce a loss in understandability while improving perceived descriptive power.

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 A.4.

Dependency on heuristics The proposed algorithm relies on an admissible heuristic to guide the search process. As discussed in Section 3, admissibility guarantees that the stopping criterion cannot discard a candidate whose exact IoU could still enter the beam. In this context, the tightness of the heuristic instead affects the computational cost, since a loose but admissible heuristic may substantially increase the number of visited nodes and runtime.

Scalability and Computational Complexity Let p be the number of concepts, b the beam width, n the maximum explanation length, and o the number of operators. Standard incremental beam search generates at most $O(bpo)$ candidates per level, and thus $O(nbpo)$ overall. Our method additionally considers internal expansions and pairwise recombinations over the beam history, resulting in $O(bpo\,n^2 + b^2n^2)$ generated candidate formulas in the worst case. Thus, the proposed search has a higher polynomial overhead than incremental beam search, although exploration remains constrained by the beam width. A complete derivation and comparison with guaranteed-optimal search are reported in Section D. In the standard configurations evaluated in this work, this overhead remains moderate, but degenerate cases with many competitive candidates, as reflected by the larger gap between mean and median runtime observed in the vision experiments, can increase runtime.

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.

6 Conclusion

In this paper, we addressed some of 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 remaining empirically tractable. 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. Additionally, we showed that these explanations exhibit higher perceived descriptive power while preserving understandability. 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.

7 Funding Transparency Statement

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0149.

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] A. Albarghouthi, S. Gulwani, and Z. Kincaid. Recursive program synthesis. In *International conference on computer aided verification*, pages 934–950. Springer, 2013.

[3] 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.

[4] 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`.

[5] 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.

[6] 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.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

[8] 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.

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

[10] 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`.

[11] 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/`.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] J. R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2):87–112, 1994.

[18] 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.

[19] B. La Rosa and L. H. Gilpin. Open vocabulary compositional explanations for neuron alignment. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=iS38vzTMdd.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] R. Srinivas, A. Kartikeya, L. H. Gilpin, and B. La Rosa. Tracking concept-level representation changes under network pruning. *BayLearn*, 2026.

[32] 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.

[33] P. R. van der Laag and S.-H. Nienhuys-Cheng. Completeness and properness of refinement operators in inductive logic programming. *The Journal of Logic Programming*, 34(3):201–225, 1998.

[34] 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.

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

[36] 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.

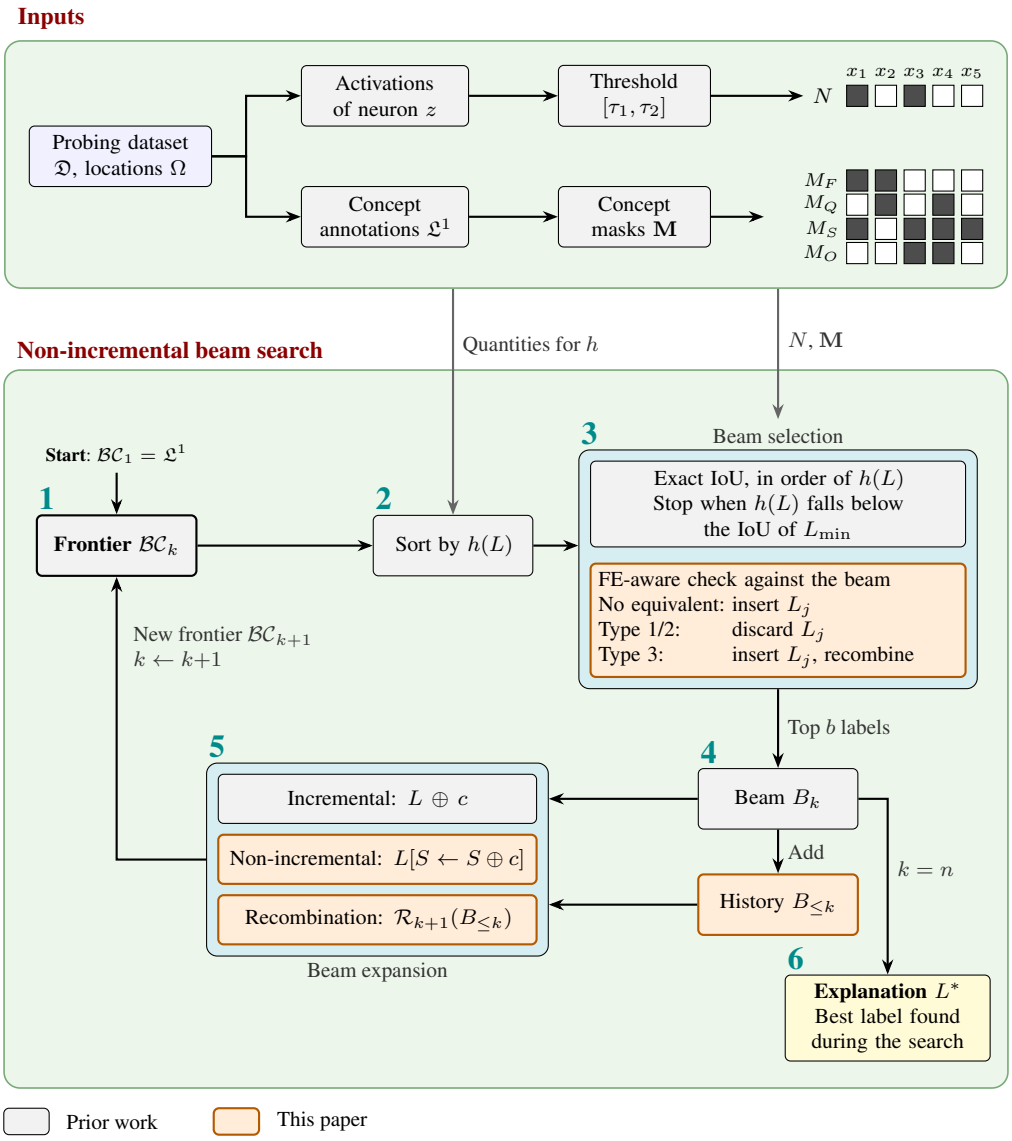

Figure 3: Overview of compositional explanations computed with the proposed non-incremental beam search. In the pre-processing step, the probing dataset is used to compute the activation mask N of neuron z by thresholding its activations within $[\tau_1, \tau_2]$. Similarly, the algorithm computes the concept masks $\mathbf{M}$ over the locations Ω. **Non-incremental beam search**: starting from the atomic concepts ($\mathcal{BC}_1 = \mathfrak{L}^1$), the frontier (1) is sorted by the admissible heuristic $h(L)$ (2). Candidates are then evaluated by exact IoU in this order and filtered by the functional-equivalence-aware check (3). The top b labels form the beam B_k (4), which is also added to the history $B_{\leq k}$. The beam and the history are expanded through incremental, non-incremental, and recombination candidates (5), producing the next frontier $\mathcal{BC}_{k+1}$. When the maximum length n is reached, the best label found during the search is returned as the explanation (6).

A Complete and Additional Results

A.1 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

Table 3: Average and standard deviation of explanation scores computed by our non-incremental beam search and the standard incremental beam search on natural language inference. The scores are averaged over units where the search solutions differ.

Method	Better (%)	IoU	DetAcc	ActCov
		BiLSTM		
Incremental	3	0.171 ± 0.055	0.477 ± 0.180	0.222 ± 0.081
Ours	**58**	**0.191** ± 0.066	**0.481** ± 0.144	**0.248** ± 0.091
		BERT		
Incremental	4	0.144 ± 0.026	0.414 ± 0.119	0.189 ± 0.043
Ours	**57**	**0.156** ± 0.030	**0.435** ± 0.118	**0.202** ± 0.044
		LLaMA		
Incremental	4	0.185 ± 0.049	0.466 ± 0.125	0.244 ± 0.074
Ours	**67**	**0.202** ± 0.053	**0.476** ± 0.119	**0.267** ± 0.076

Table 4: Average and standard deviation of 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	40 ± 0	40	370246 ± 11377	29 ± 4	29
Ours	41 ± 3	40	576139 ± 176192	49 ± 14	48
			BERT		
Incremental	40 ± 0	40	369144 ± 8826	27 ± 5	26
Ours	42 ± 3	40	724188 ± 197174	55 ± 14	53
			LLaMA		
Incremental	40 ± 0	40	366149 ± 9856	28 ± 5	28
Ours	41 ± 3	40	651615 ± 181464	54 ± 14	51

report the scores for the explanation metrics for units where the search solutions differ, 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. [3] and Mu and Andreas [26], 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.

A.2 Heuristics Comparison

For completeness, Table 7 compares the original MMESH heuristic [21] with the heuristic adopted [20] in this paper, in the same vision settings. Since both heuristics are upper bounds, they estimate the same candidates and return explanations with the same IoU on every unit. They differ only in how many candidates are evaluated exactly. Overall, under the same incremental beam-search formulation, the heuristic adopted in this work requires fewer exact evaluations and lower runtime than the original MMESH heuristic across all three vision models. These results are

Table 5: Average and standard deviation of explanation scores for units where the search solutions differ, associated with explanations computed by our non-incremental beam search, the standard incremental beam search, and the optimal algorithm on image classification.

Method	Optimal	Better (%)	IoU	DetAcc	ActCov
			ResNet18		
Incremental	0.93	0	0.073 ± 0.042	0.133 ± 0.090	0.188 ± 0.108
Ours	**0.96**	**5**	0.076 ± 0.044	0.137 ± 0.088	0.181 ± 0.106
Optimal	**1.00**	-	0.077 ± 0.044	0.139 ± 0.086	0.180 ± 0.108
			AlexNet		
Incremental	0.78	0	0.049 ± 0.021	0.082 ± 0.042	0.155 ± 0.103
Ours	**0.92**	**15**	0.052 ± 0.021	0.084 ± 0.039	0.151 ± 0.095
Optimal	**1.00**	-	0.052 ± 0.022	0.084 ± 0.038	0.155 ± 0.098
			DenseNet161		
Incremental	0.76	0	0.046 ± 0.035	0.088 ± 0.084	0.115 ± 0.083
Ours	**0.90**	**14**	0.048 ± 0.036	0.099 ± 0.092	0.109 ± 0.077
Optimal	**1.00**	-	0.049 ± 0.037	0.100 ± 0.094	0.105 ± 0.083

Table 6: Average and standard deviation of 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	28 ± 30	15	32621 ± 2810	67 ± 7	65
Ours	71 ± 203	17	48142 ± 5722	73 ± 35	65
			AlexNet		
Incremental	57 ± 45	45	34086 ± 2406	70 ± 8	68
Ours	398 ± 623	102	46866 ± 7628	130 ± 111	77
			DenseNet161		
Incremental	50 ± 42	37	36810 ± 3616	67 ± 6	66
Ours	267 ± 578	62	51130 ± 7666	108 ± 98	74

in line with and confirm the results provided in [20], where the reader can find a more extensive comparison between the two heuristics.

A.3 Optimal Explanations for NLP

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 [20]. 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 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.

Table 7: Search metrics for incremental beam search using the heuristic adopted in this work and the MMESH heuristic.

Model	Heuristic	Visited	s/u
ResNet18	Adopted	28	67
	MMESH	58	76
AlexNet	Adopted	57	70
	MMESH	111	84
DenseNet161	Adopted	50	67
	MMESH	122	86

Table 8: Average scores and standard deviation for explanation metrics when individual components of our framework are removed.

Method	**Better (%)**	**IoU**	**DetAcc**	**ActCov**
w/o Recombinations	0.39	0.147 $\pm$ 0.021	**0.447** $\pm$ 0.118	0.184 $\pm$ 0.029
Full	**3.12**	**0.154** $\pm$ 0.025	0.418 $\pm$ 0.116	**0.201** $\pm$ 0.029
w/o Functional Equivalence	2	0.141 $\pm$ 0.024	0.397 $\pm$ 0.121	0.186 $\pm$ 0.037
Full	**21**	**0.148** $\pm$ 0.025	**0.421** $\pm$ 0.123	**0.191** $\pm$ 0.035
w/o Non-Incremental Expansion	5	0.146 $\pm$ 0.027	0.410 $\pm$ 0.116	0.192 $\pm$ 0.043
Full	**53**	**0.156** $\pm$ 0.030	**0.438** $\pm$ 0.118	**0.202** $\pm$ 0.044

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.

A.4 Ablation Study and Sensitivity

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

A.4.1 Ablation Study

Table 8 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; 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 on all the units of BERT and compare the explanations that differ pairwise. From Table 8, we observe that all the components are important to achieve the highest results, since the full version achieves a higher average IoU in all the ablations. Among the components, the non-incremental expansion appears to be the most impactful one since, without it, the search returns a less aligned explanation in 53% of the units. Overall, the fact that ablated versions are sometimes able to achieve higher scores represents an interesting clue for future research, since it implies that there is still room for improvement in the search procedure and that these explanations are not all optimal.

A.4.2 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 recombi-

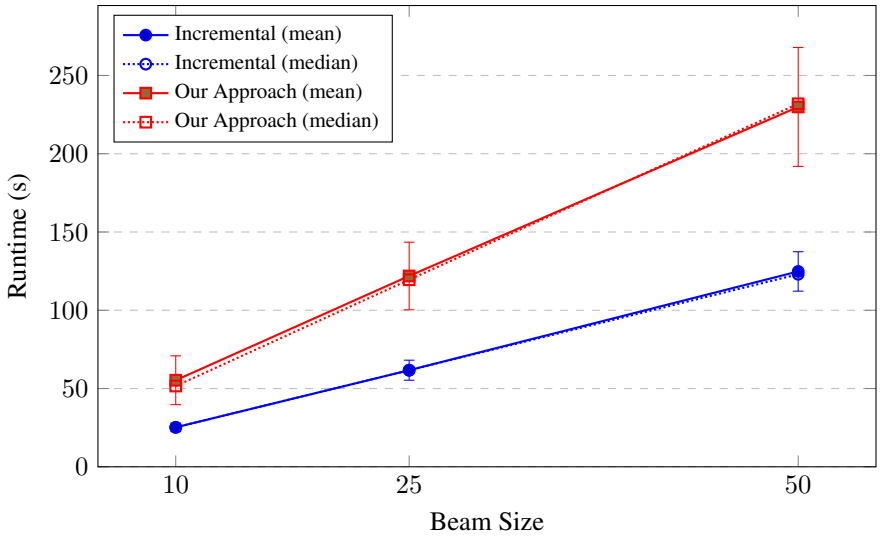

Figure 4: Runtime vs Beam Size on 50 randomly extracted units. Solid lines report the mean (error bars denote standard deviation), dotted lines the median.

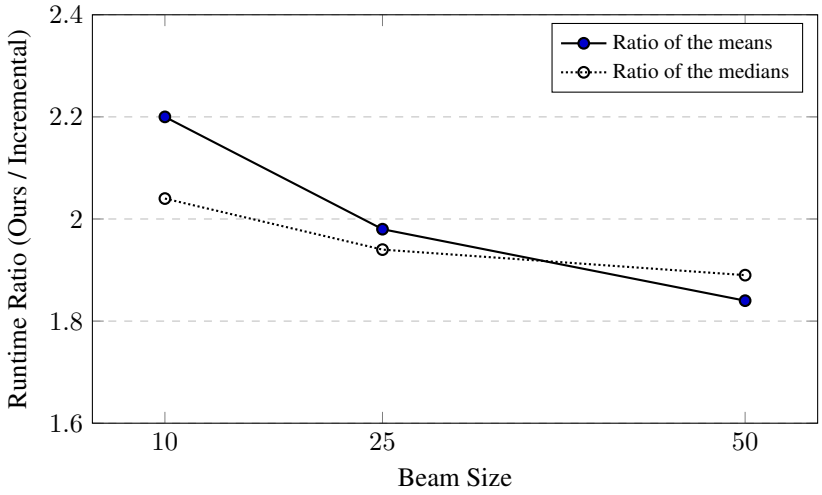

Figure 5: Runtime ratio between our approach and incremental beam search across beam sizes. The solid line indicates the ratio of the mean runtimes, while the dotted line indicates the ratio of the median runtimes.

nations 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. All configurations are evaluated on the same 50 randomly extracted units. Since a few units can be much slower than the others, we report both the mean and the median runtime of each search.

Figures 4 and 5 show that the ratio between the runtime of our approach and that of the incremental beam search remains approximately constant as the beam width grows, indicating similar scaling behavior. Conversely, the ratio grows with the maximum explanation length (Figures 6 and 7), in line with the complexity analysis of Section D. For the typical unit, this growth is moderate, as highlighted by the ratio of the medians. For a few units, however, the non-incremental expansions generate many more candidates, and the runtime for those units grows significantly. Consequently, in configurations involving very long explanations, the proposed algorithm may become expensive 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

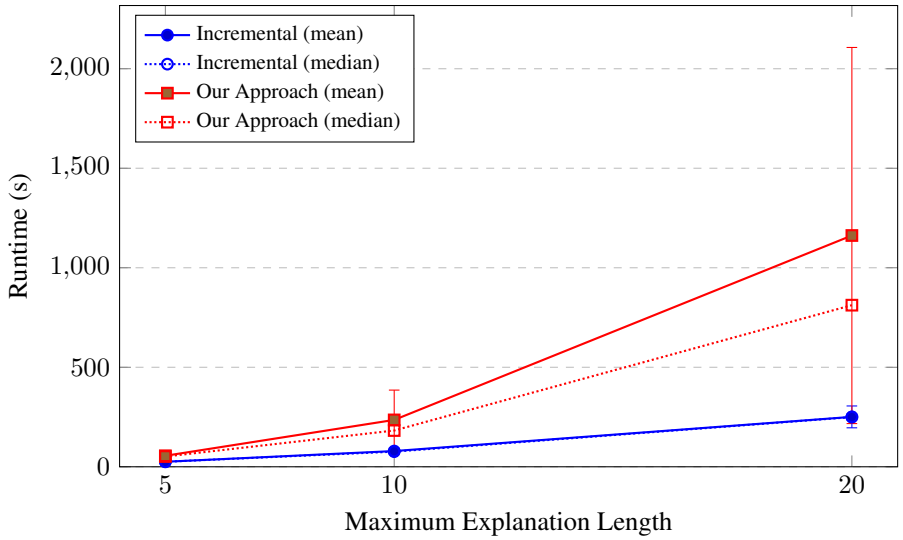

Figure 6: Runtime as a function of the maximum explanation length on 50 randomly extracted units. Solid lines report the mean (error bars denote standard deviation), dotted lines the median.

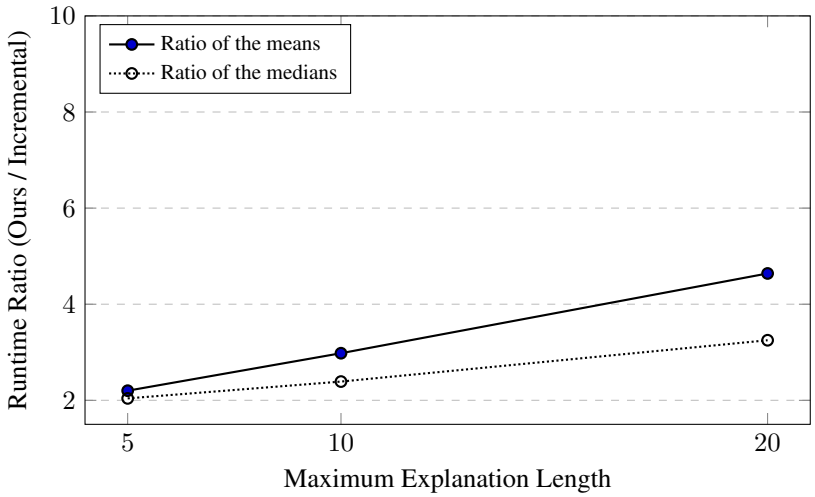

Figure 7: Runtime ratio between our approach and incremental beam search as a function of the maximum explanation length. The solid line indicates the ratio of the mean runtimes, while the dotted line indicates the ratio of the median runtimes.

even for 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 9 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 9: Average and standard deviation of 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	2	0.145 $\pm$ 0.022	0.407 $\pm$ 0.107	0.192 $\pm$ 0.043
Ours	**70**	**0.155** $\pm$ 0.022	**0.432** $\pm$ 0.094	**0.199** $\pm$ 0.036

B Metrics

This section reports the Detection Accuracy and Activation Coverage definitions, as introduced in [21, 22]. 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 [22] 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.

$$DetAcc(L, N, \mathbf{M}) = \frac{|N \cap \circ_L(\mathbf{M})|}{|\circ_L(\mathbf{M})|} \tag{13}$$

Activation Coverage [21] 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.

$$ActCov(L, N, \mathbf{M}) = \frac{|N \cap \circ_L(\mathbf{M})|}{|N|} \tag{14}$$

C User Study

To evaluate whether the additional structural complexity introduced by non-incremental explanations affects their interpretability, we conducted a human study in the NLP setting, where the two formulations can produce different logical structures.

To conduct this study, we recruited participants through Prolific. Participants were required to have English as their primary language, an approval rate greater than 90%, and more than 100 previous submissions. We additionally required either Prolific's qualified AI tasker status or a background in computer science. The study was pre-approved under an exemption by the University of California, Santa Cruz Institutional Review Board and is currently undergoing the full approval process. Informed consent was obtained from all participants. Participation was voluntary, and participants could withdraw before submitting the survey. Participants received $5 for completing the study, corresponding to an average compensation of $9.42 per hour. The final analysis included 65 qualified participants. Before the evaluation, participants completed a tutorial explaining the logical operators, parentheses, and the two evaluation criteria, followed by five qualification questions. We retained participants who answered at least four of the five qualification questions correctly.

Each qualified participant evaluated six comparisons between an incremental and a non-incremental explanation, resulting in 390 judgments per criterion. For each comparison, participants were shown eight premise–hypothesis pairs associated with a highly activating unit together with two explanations, anonymously presented as Explanation A and Explanation B. The assignment of the incremental and non-incremental explanations to A and B was randomized and balanced, and both explanations were evaluated using the same set of examples. Participants evaluated two criteria. *Descriptive power* asked: "Considering all eight sentence pairs, which explanation better describes the pattern or patterns associated with this unit's activation?" *Understandability* asked: "Which explanation was easier to understand as a description of when this unit activates?" For each criterion, participants selected one of five ordered responses: Explanation A is much better, Explanation A is better, the explanations are about equally good, Explanation B is better, or Explanation B is much better.

For the aggregate analysis, responses were grouped into three categories: preference for the incremental explanation, equivalence, and preference for the non-incremental explanation. Since each participant provided multiple judgments, statistical tests were performed at the participant level

rather than treating the 390 judgments as independent. For each criterion, we counted whether each participant favored the non-incremental formulation more often, favored the incremental formulation more often, or was tied. Directional preference was evaluated using a two-sided exact sign test, excluding ties. For understandability, we additionally performed a non-inferiority analysis. For each participant, we computed the difference between the proportion of judgments favoring the non-incremental and incremental formulations and tested non-inferiority using a 10% margin.

Results. For descriptive power, 42.8% of the judgments favored the non-incremental explanation, 24.9% considered the explanations equivalent, and 32.3% favored the incremental explanation. At the participant level, 35 participants favored the non-incremental formulation more often, 17 favored the incremental formulation more often, and 13 were tied. This corresponds to a statistically significant preference for the non-incremental formulation (two-sided exact sign test, $p = .018$).

For understandability, 30.0% of the judgments favored the non-incremental explanation, 45.4% considered the explanations equivalent, and 24.6% favored the incremental explanation. At the participant level, 29 participants favored the non-incremental formulation more often, 23 favored the incremental formulation more often, and 13 were tied. We did not observe a significant directional preference (two-sided exact sign test, $p = .488$). Using a 10% non-inferiority margin, we reject that the non-incremental formulation introduces a corresponding loss in understandability ($p < .001$).

These results suggest that the structural complexity introduced by non-incremental explanations can improve how well explanations characterize neuron behavior without making them harder to understand in the setting considered in this paper.

D Complexity Analysis

In this section, we analyze the worst-case number of candidate formulas generated by each search procedure. This quantity is the major factor influencing the running time because generated candidates require a heuristic estimate and may require an exact IoU computation, which is the most expensive operation. We characterize this quantity using the number of atomic concepts $p = |\mathcal{L}^1|$, the beam width b, the maximum explanation length n, and the number of admissible operators $o = |\mathcal{O}|$.

Incremental beam search. At each level, the beam contains at most b labels. Each label can be expanded using at most p concepts and o operators. Therefore, the incremental formulation generates at most $O(bpo)$ candidates per level. Thus, across at most n levels, incremental beam search generates $O(nbpo)$ candidate formulas.

Non-incremental beam search. Our formulation retains the incremental expansions and additionally generates candidates through internal insertion and pairwise recombination.

For internal insertion, a formula of length k contains at most $k-1$ internal operator positions at which a new concept can be inserted. Since at most b formulas are expanded using at most p concepts and o operators, the number of internal-expansion operations at level k is bounded by $O(kbpo)$. Summing over all levels results in $\sum_{k=1}^{n} O(kbpo) = O(bpo\, n^2)$ candidate formulas.

For pairwise recombination, pairs considered at previous levels are not recomputed. When the beam B_k is generated, the algorithm considers pairs between formulas in B_k and formulas retained at preceding levels, as well as pairs within B_k. The current beam contains at most b formulas, while the search history contains at most bk formulas. Therefore, the number of new pairs considered at level k is at most $O(b^2k)$. Summing over all levels, the total number of distinct pairwise recombination attempts is $\sum_{k=1}^{n} O(b^2k) = O(b^2n^2)$.

Including incremental, internal expansions, and pairwise recombination, the total number of candidate formulas generated by our method is therefore

$$O(nbpo + bpo\, n^2 + b^2n^2) = O(bpo\, n^2 + b^2n^2)$$

Optimal search. The optimal method [20] does not constrain the search to a fixed-width beam. Therefore, in the worst case, it may need to explore the full admissible state space. Under the assumption that each concept appears at most once, the number of incremental formulas up to length n is upper-bounded by

$$\sum_{k=1}^{n} o^{k-1}(p)_k$$

where $(p)_k = \frac{p!}{(p-k)!}$ is the number of ordered selections of k distinct concepts. For fixed n, this quantity is polynomial in p, but it grows combinatorially as the maximum explanation length increases. The admissible heuristic used by the guaranteed-optimal method can substantially reduce the number of explored states in practice, but in the worst case its behavior approaches exhaustive exploration.

Our method trades the optimality guarantee for a beam search whose computational cost of candidate generation is polynomial in p, b, o, and n, while empirically recovering optimal explanations more frequently than standard incremental beam search.

Additional computational operations. The bounds above count generated candidate formulas and abstract from the elementary operations required to construct them. These costs can introduce additional dependence on formula length. For example, processing a recombination pair may require $O(k)$ operations in the worst case to inspect the concepts in formulas of length at most k, identify shared and non-shared concepts, and construct the resulting formula. Consequently, although only $O(b^2n^2)$ distinct recombination pairs are considered across the search, their structural processing can require up to $O(b^2n^3)$ operations in the worst case. Similarly, identifying compatible subformulas and constructing internally modified formulas can introduce additional formula-length-dependent operations. However, these structural operations are order-of-magnitude faster than mask-based computations used during search. For this reason, we excluded them from the complexity analysis above, and we report wall-clock runtime in the experiments to capture the remaining implementation costs.

E Heuristics

Our algorithm uses the admissible heuristic introduced in [20] for estimating the IoU of candidate formulas. Specifically, we use the sample-based label-quantity estimation. We use this heuristic since it is compatible with both vision and NLP settings and, applied recursively, does not require the incremental formulation.

The heuristic decomposes the locations in Ω according to whether they are associated with a single atomic concept (unique elements U) or multiple concepts (common elements C). We define

$$U = \left\{ \omega \in \Omega : \sum_{c \in \mathcal{L}^1} \mathbb{1}[\omega \in M_c] = 1 \right\}$$

$$C = \left\{ \omega \in \Omega : \sum_{c \in \mathcal{L}^1} \mathbb{1}[\omega \in M_c] > 1 \right\}$$

Starting from these definitions, the neuron activations are partitioned into $N^U = N \cap U$ and $N^C = N \cap C$, respectively. We also define the space available for extra locations as:

$$SE^U = U \setminus N$$

$$SE^C = C \setminus N$$

Given a candidate formula L, we define its unique and common intersection locations as:

$$I^U(L) = \circ_L(\mathcal{M}) \cap N^U$$

$$I^C(L) = \circ_L(\mathcal{M}) \cap N^C$$

and its unique and common extra locations as

$$E^U(L) = \circ_L(\mathcal{M}) \cap SE^U,$$
$$E^C(L) = \circ_L(\mathcal{M}) \cap SE^C.$$

Using these quantities, the IoU can be equivalently expressed as:

$$\mathrm{IoU}(N, \circ_L(\mathcal{M})) = \frac{|I^U(L)| + |I^C(L)|}{|N| + |E^U(L)| + |E^C(L)|}$$

Because unique locations cannot be shared by distinct concepts, their contribution can be computed exactly from the operators in the formula. Conversely, masks of different concepts may overlap in common locations. We precompute whether concept masks are disjoint. If they are disjoint, the corresponding quantities can be computed exactly. Otherwise, the heuristic estimates an upper bound on the common intersection ($I^C_{\max}$) and a lower bound on the common extras ($E^C_{\min}$).

Following [20], the bounds are defined for a formula $L = L_{\leftarrow} \oplus L_{\rightarrow}$ in terms of the quantities of its two operands. In [20], $L_{\leftarrow}$ is a label retained by beam search and $L_{\rightarrow}$ an atomic concept. Since non-incremental candidates do not necessarily have this form, we apply the same rules recursively over the syntax tree of the candidate. The recursion stops at operands that are either atomic concepts or labels that entered the beam at a previous level. In this case, their quantities are known exactly, and the heuristic only needs to estimate how these quantities combine under the operator $\oplus$, and in particular the unknown overlap within the common locations.

For a formula $L = L_{\leftarrow} \oplus L_{\rightarrow}$, these bounds are computed per sample. For OR:

$$|I^C_{\max}(L)_x| = \min\left(|I^C_{\max}(L_{\leftarrow})_x| + |I^C_{\max}(L_{\rightarrow})_x|, |N^C_x|\right)$$
$$|E^C_{\min}(L)_x| = \max\left(|E^C_{\min}(L_{\leftarrow})_x|, |E^C_{\min}(L_{\rightarrow})_x|\right)$$

For AND:

$$|I^C_{\max}(L)_x| = \min\left(|I^C_{\max}(L_{\leftarrow})_x|, |I^C_{\max}(L_{\rightarrow})_x|\right)$$
$$|E^C_{\min}(L)_x| = \max\left(|E^C_{\min}(L_{\leftarrow})_x| + |E^C_{\min}(L_{\rightarrow})_x| - |SE^C_x|, 0\right)$$

Finally, for AND NOT:

$$|I^C_{\max}(L)_x| = \min\left(|I^C_{\max}(L_{\leftarrow})_x|, |N^C_x| - |I^C(L_{\rightarrow})_x|\right)$$
$$|E^C_{\min}(L)_x| = \max\left(|E^C_{\min}(L_{\leftarrow})_x| - |E^C(L_{\rightarrow})_x|, 0\right)$$

These estimates correspond to the most favorable overlap compatible with the information available for the two operands. The heuristic IoU is then computed as:

$$h(L) = \frac{\sum_{x \in \mathcal{D}} \left(|I^U(L)_x| + |I^C_{\max}(L)_x|\right)}{|N| + \sum_{x \in \mathcal{D}} \left(|E^U(L)_x| + |E^C_{\min}(L)_x|\right)}$$

Additionally, candidates that cannot improve over one of their operands (e.g., disjunctions whose added operand cannot increase the intersection with N) or that contain unverified negations are discarded without computing their bound.

The heuristic is admissible because the unique quantities are exact and the estimated common quantities satisfy $|I^C_{\max}(L)_x| \geq |I^C(L)_x|$ and $|E^C_{\min}(L)_x| \leq |E^C(L)_x|$. Thus, the heuristic cannot underestimate the IoU numerator or overestimate its denominator, and therefore $h(L) \geq \mathrm{IoU}(N, \circ_L(\mathcal{M}))$. A complete derivation and proof of admissibility for the considered operators are provided in [20]. In the case of non-incremental formulas, the recursive application preserves admissibility. Indeed, the only difference is the replacement of the exact quantities of $L_{\rightarrow}$ with their upper and lower bounds, as computed by the heuristic itself, and using an overestimated intersection and underestimated extras in the rules can only overestimate the intersection and underestimate the extras of L.

This admissibility property justifies the stopping criterion used in our beam selection. Since candidates are processed in decreasing order of $h(L)$, once the next candidate has a heuristic value lower than the exact IoU of the lowest-scoring formula in a full beam, neither that candidate nor any subsequent candidate can enter the beam.

F Algorithm Design Choices

This section motivates and further expands on some of the design choices we took in proposing the non-incremental algorithm. Overall, our choices and rules should be viewed as a conservative first design rather than as an optimal trade-off between expressivity and computation.

Expansion Rules Motivations. The proposed expansion rules represent a first attempt to support non-incremental search strategies. We designed the proposed rules to enlarge the reachable search space while keeping candidate generation manageable and preserving the assumption that each concept appears at most once in an explanation. The incremental expansion preserves the standard search paths considered in prior work. The internal insertion addresses a failure mode observed when a concept that improves the explanation may need to be inserted within an existing subformula rather than appended to the complete formula [20]. Pairwise recombination instead provides a limited form of backtracking, allowing the algorithm to combine structures retained at different stages of the search rather than extending only the current beam.

Recombinations When combining formulas from the beam history, formulas that do not share concepts provide the simplest case because no duplicated concepts need to be resolved. When the formulas share concepts, we restrict recombination to the structures considered in Rules 2 and 3, where duplicated concepts can be removed without requiring general logical simplification. More general recombinations between arbitrary overlapping formulas could introduce repeated concepts or require transformations to preserve the semantics of the original formulas. We leave these broader recombination strategies to future work.

Recombinations beyond Beam Candidates. In preliminary experiments, we also considered the configuration in which recombination used the top-q candidates at each level, with $q > b$, rather than only formulas retained in the beam. These variants produced some additional improvements in alignment, but increased both runtime and the amount of search history that needed to be maintained. They would also combine the effect of non-incremental recombination with that of retaining a larger candidate pool than standard beam search. We therefore restrict recombination to beam candidates in the formulation evaluated in this work.

G 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 does not improve or degrade the broader impact of explainability methods in this area of research.

H Reproducibility

This section serves as a brief summary and documentation of the experimental setup used by our framework, along with the resources required.

H.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
- BERT
  - Accessible at: https://huggingface.co/ccenli/dense_weights/blob/main/bert_dense.pth
  - License: unknown
- LLaMA
  - Accessible at: https://huggingface.co/ccenli/dense_weights/blob/main/llama_dense.pth
  - License: unknown

Explanations Our framework generates explanations through a heuristic search guided by the heuristic proposed in [20]. We reimplemented the heuristic and applied it to the NLP domain following the pipeline available at https://github.com/jayelm/compexp. Regarding the hyperparameters, we follow the setup proposed in [26], 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 concept set defined in [26].

H.2 Resources

The proposed algorithm introduces additional computational and memory overhead with respect to incremental compositional explanations, as discussed in Section D. 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 ($\sim$ 10 GB). In the NLP setting analyzed in this paper, they are vectors, leading to a significantly lighter setup ($\sim$ 1 GB). The number of concepts is also related to the data domain: a larger concept set increases memory usage. As a reference, the $\sim$ 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 $\sim$ 20 GB for LLaMA.

I 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. Therefore, the images shown do not represent the only and specific samples explained by the labels, but they represent a limited representative subset of the state space over which the explanations are computed.

I.1 Vision

For vision, we follow the setup used in [21]. 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 8 to 13) are not cherry-picked and correspond to the first six differing explanations for the penultimate layer of each explained model (ResNet18 [13], AlexNet [18], and DenseNet161 [15]). 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.

I.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 14 to 31) are not cherry-picked and correspond to the first six units where the explanations differ for the MLP layer of each analyzed model (BiLSTM [26], BERT [7], and LLaMA [32, 35]).

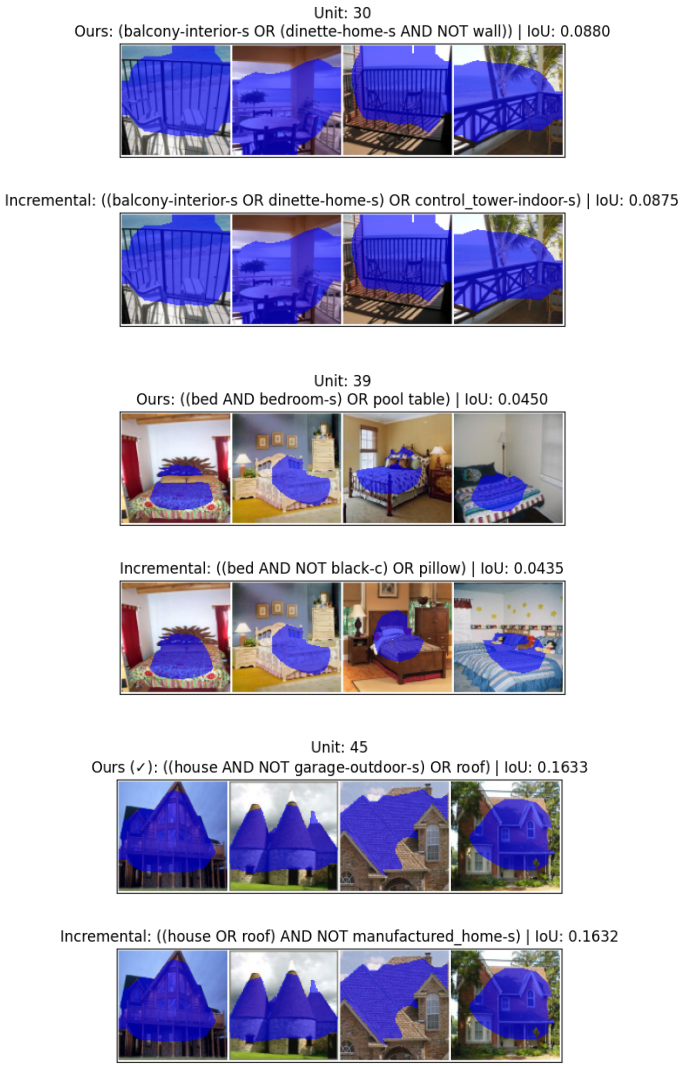

Figure 8: Alignment detected incremental and non-incremental beam-search methods on ResNet18. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.



Figure 9: Alignment detected incremental and non-incremental beam-search methods on ResNet18. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

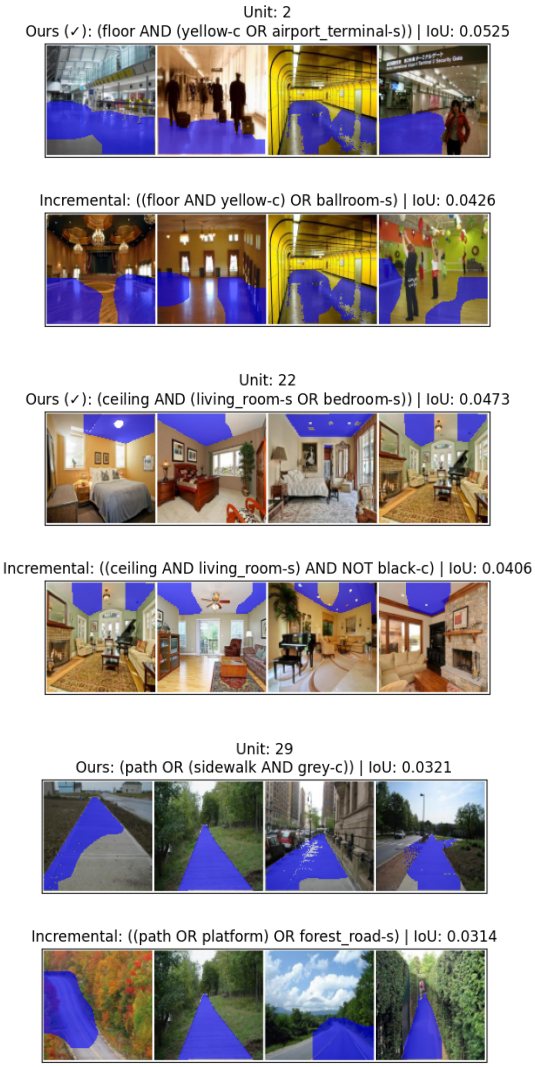

Figure 10: Alignment detected incremental and non-incremental beam-search methods on AlexNet. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.



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

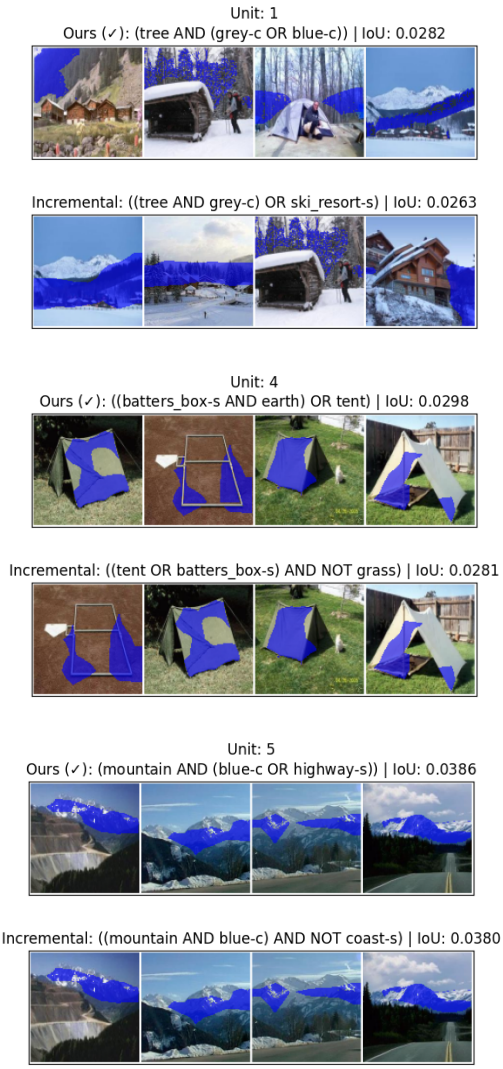

Figure 12: Alignment detected incremental and non-incremental beam-search methods on DenseNet161. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

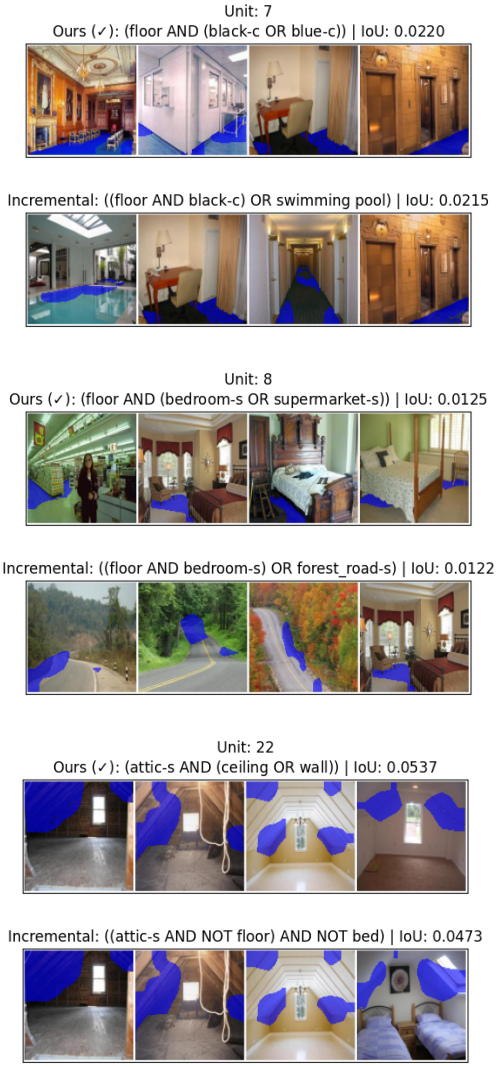

Figure 13: Alignment detected incremental and non-incremental beam-search methods on DenseNet161. The blue regions indicate areas of neuron activation within the considered range. The checkmark indicates whether one of two explanations is optimal.

Unit 0 – Incremental Beam Search

IoU: 0.136

Explanation

((((hyp:tok:alone ∨ pre:tok:sleeves) ∧ ¬hyp:tag:rb) ∨ hyp:tok:sleeps) ∧ 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

(((pre:tok:sleeves ∨ hyp:tok:alone) ∨ hyp:tok:sleeps) ∧ (pre:tag:vbn ∨ 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 14: Comparison between incremental and non-incremental compositional explanations for Unit 0 in the BiLSTM model.

Unit 15 – Incremental Beam Search

IoU: 0.246

Explanation

((((hyp:tok:sitting ∧ neighbor:pre:tok:around) ∧ ¬pre:tok:sitting) ∧ ¬pre:tag:rb) ∧ ¬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

((hyp:tok:sitting ∧ ((pre:tok:walking ∨ neighbor:pre:tok:between) ∨ hyp:tok:bench)) ∧ ¬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 15: Comparison between incremental and non-incremental compositional explanations for Unit 15 in the BiLSTM Model.

Unit 16 – Incremental Beam Search

IoU: 0.129

Explanation

((((neighbor:hyp:tok:trick ∧ neighbor:hyp:tok:him) ∨ hyp:tok:eye) ∧ ¬hyp:tok:man) ∨ 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

(((neighbor:hyp:tok:trick ∨ neighbor:hyp:tok:drummer) ∧ (neighbor:hyp:tok:him ∨ neighbor:hyp:tok:after)) ∨ 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 16: Comparison between incremental and non-incremental compositional explanations for Unit 16 in the BiLSTM model.

Unit 17 – Incremental Beam Search

IoU: 0.133

Explanation

((((hyp:tok:girls ∧ ¬neighbor:pre:tok:cheerleaders) ∧ neighbor:pre:tok:pair) ∨ pre:tok:teen) ∧ ¬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

((((hyp:tok:girls ∧ ¬neighbor:pre:tok:cheerleaders) ∨ hyp:tok:tv) ∧ neighbor:pre:tok:pair) ∨ 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 17: Comparison between incremental and non-incremental compositional explanations for Unit 17 in the BiLSTM model.

Unit 19 – Incremental Beam Search

IoU: 0.113

Explanation

((((hyp:tok:boys ∧ ¬pre:tok:boys) ∨ neighbor:hyp:tok:atv) ∧ neighbor:pre:tok:row) ∨ hyp:tok:dryer)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
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: 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 and ATV.	🟢
P: A woman hanging out her laundry on the line.	
H: A man fixes a dryer.	🔴

Unit 19 – Our Non-Incremental Search

IoU: 0.118

Explanation

((((pre:tok:blows ∨ neighbor:hyp:tok:atv) ∨ neighbor:hyp:tok:inflatable) ∨ neighbor:hyp:tok:cheer) ∧ 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 18: Comparison between incremental and non-incremental compositional explanations for Unit 19 in the BiLSTM Model.

Unit 22 – Incremental Beam Search

IoU: 0.156

Explanation

((((pre:tok:dress ∧ ¬neighbor:hyp:tok:female) ∧ neighbor:hyp:tok:men) ∨ pre:tok:routine) ∧ ¬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

(((pre:tok:dress ∨ neighbor:pre:tok:nun) ∧ (hyp:tok:man ∨ hyp:tok:boy)) ∧ ¬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 19: Comparison between incremental and non-incremental compositional explanations for Unit 22 in the BiLSTM model.

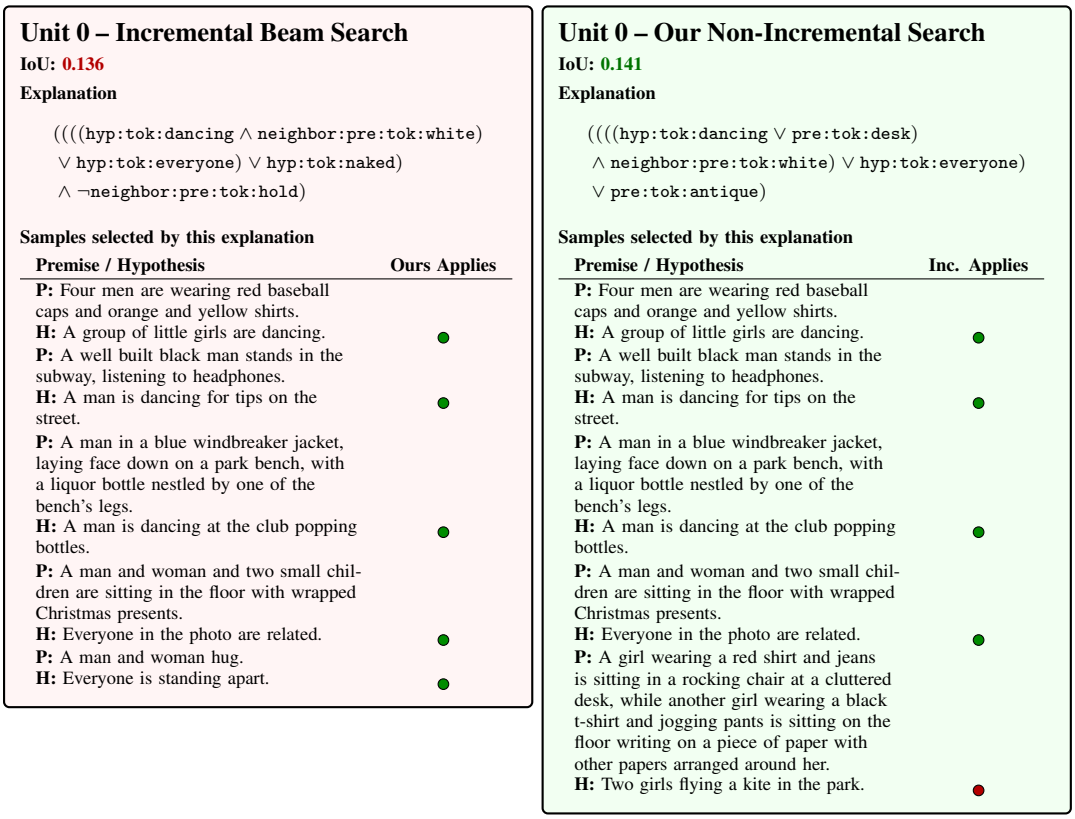

Figure 20: Comparison between incremental and non-incremental compositional explanations for Unit 0 in the BERT model.

Unit 1 – Incremental Beam Search

IoU: 0.197

Explanation

((((pre:tok:surgery ∨ hyp:tok:belly) ∨ hyp:tok:dances) ∨ hyp:tok:appears) ∨ hyp:tok:captured)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: The tiny girl dressed in a butterfly ballet costume, waves her wand near a nutcracker statue. **H:** A small girl dances in the nutcracker.	●
P: Three women wearing belly dancing costumes are pictured in color while the background of the shot is in black and white. **H:** Three women dance in belly dancing costumes	●
P: A surgeon is performing an open surgery. **H:** A doctor is doing surgery.	●
P: Two doctors perform surgery on patient. **H:** Doctors are performing surgery.	●
P: A large man in white dances in the middle of a crowd holding a beverage in his hand. **H:** A man dances at a wedding.	●

Unit 1 – Our Non-Incremental Search

IoU: 0.205

Explanation

((((neighbor:hyp:tok:hula ∧ neighbor:hyp:tok:women) ∨ pre:tok:surgery) ∨ hyp:tok:dances) ∨ hyp:tok:belly)

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: The tiny girl dressed in a butterfly ballet costume, waves her wand near a nutcracker statue. **H:** A small girl dances in the nutcracker.	●
P: Three women wearing belly dancing costumes are pictured in color while the background of the shot is in black and white. **H:** Three women dance in belly dancing costumes	●
P: A surgeon is performing an open surgery. **H:** A doctor is doing surgery.	●
P: Two women on the dance floor with other people dancing. **H:** Two women are dancing on the dance floor.	●
P: Two doctors perform surgery on patient. **H:** Doctors are performing surgery.	●

Figure 21: Comparison between incremental and non-incremental compositional explanations for Unit 1 in the BERT model.

Unit 2 – Incremental Beam Search

IoU: 0.141

Explanation

((((pre:tok:gloves ∨ pre:tok:backs) ∧ neighbor:pre:tok:shirt) ∨ pre:tok:lemons) ∨ 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

((((pre:tok:you ∨ pre:tok:denim) ∧ ¬pre:tag:vbz) ∨ pre:tok:backs) ∨ 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 22: Comparison between incremental and non-incremental compositional explanations for Unit 2 in the BERT model.

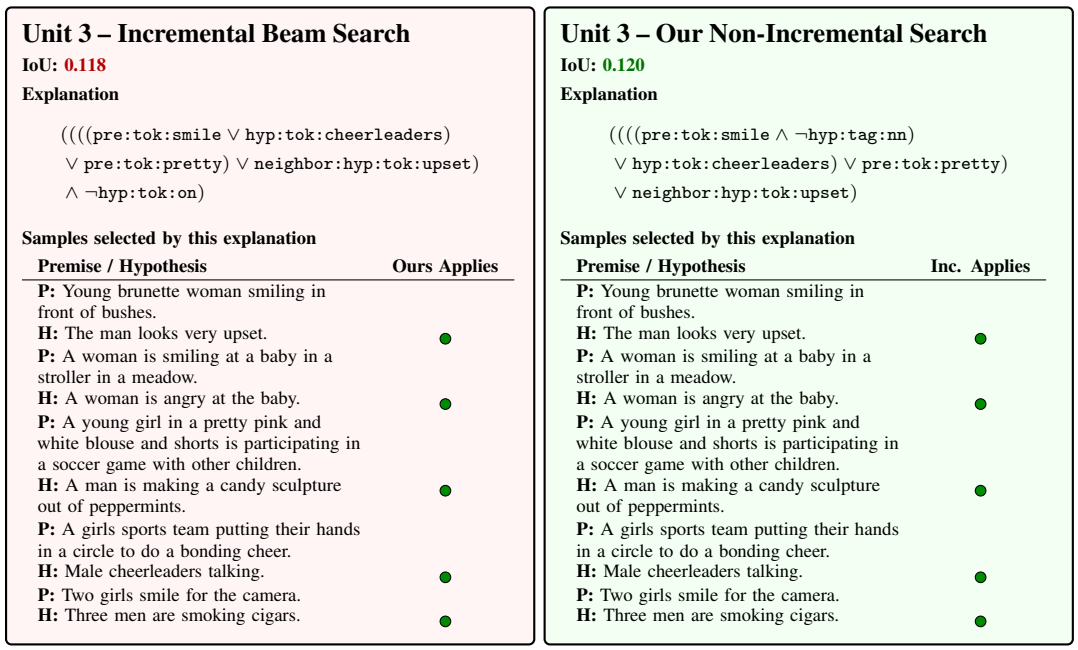

Figure 23: Comparison between incremental and non-incremental compositional explanations for Unit 3 in the BERT model.

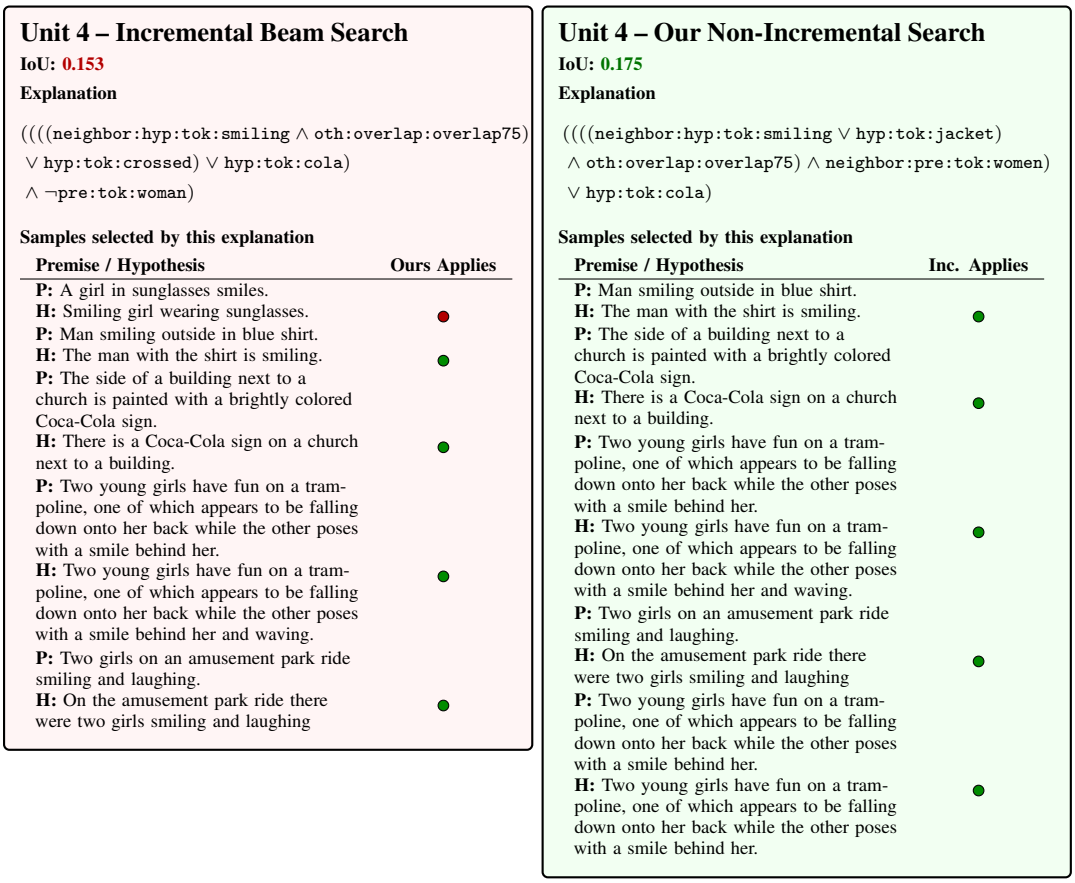

Figure 24: Comparison between incremental and non-incremental compositional explanations for Unit 4 in the BERT model.

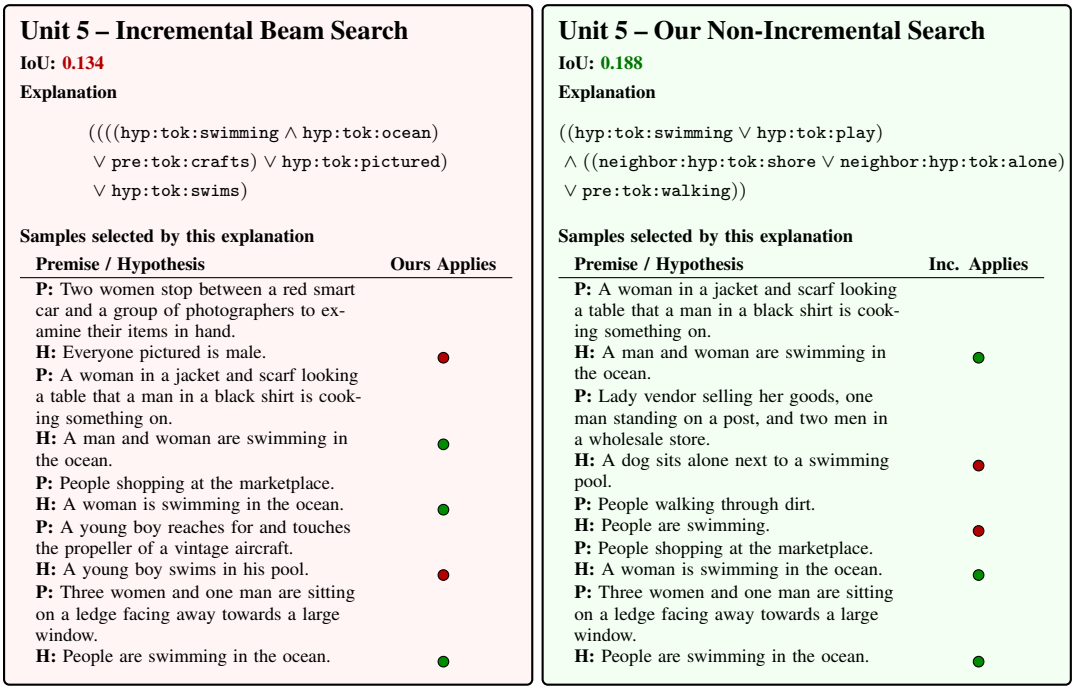

Figure 25: Comparison between incremental and non-incremental compositional explanations for Unit 5 in the BERT model.

Unit 0 – Incremental Beam Search

IoU: 0.145

Explanation

((((hyp:tok:bridge ∨ hyp:tok:plaid) ∨ pre:tok:lighting) ∧ oth:overlap:overlap50) ∨ pre:tok:artist)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: A very young child in a red plaid coat and pink winter hat makes a snowball in a large pile of snow. **H:** A child in a red plaid coat and pink winter hat makes a snowball in a large pile of snow.	●
P: A very young child in a red plaid coat and pink winter hat makes a snowball in a large pile of snow. **H:** A very young child in a red plaid coat and pink winter hat makes a snowball.	●
P: A woman walking across an old rope and driftrood bridge between mountains. **H:** A woman walking on an old bridge near a mountain.	●
P: An artist dressed in a gray sweatshirt is working on a brightly colored painting of a town. **H:** A person is holding a baseball bat and standing at the plate awaiting an oncoming pitch.	●
P: There are people walking down a busy nighttime street with signs lighting the way. **H:** People are walking down a street holding homemade signs.	●

Unit 0 – Our Non-Incremental Search

IoU: 0.159

Explanation

((((pre:tok:winter ∨ hyp:tok:bridge) ∨ hyp:tok:tent) ∨ pre:tok:lighting) ∧ oth:overlap:overlap50)

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: A very young child in a red plaid coat and pink winter hat makes a snowball in a large pile of snow. **H:** A child in a red plaid coat and pink winter hat makes a snowball in a large pile of snow.	●
P: A very young child in a red plaid coat and pink winter hat makes a snowball in a large pile of snow. **H:** A very young child in a red plaid coat and pink winter hat makes a snowball.	●
P: Two people are sitting under a tent while another stands. **H:** Two people sitting under a tent while another bakes a cake.	●
P: A woman walking across an old rope and driftrood bridge between mountains. **H:** A woman walking on an old bridge near a mountain.	●
P: There are people walking down a busy nighttime street with signs lighting the way. **H:** People are walking down a street holding homemade signs.	●

Figure 26: Comparison between incremental and non-incremental compositional explanations for Unit 0 in the LLaMA 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: 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: 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: 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: 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: 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: 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 27: Comparison between incremental and non-incremental compositional explanations for Unit 1 in the LLaMA model.

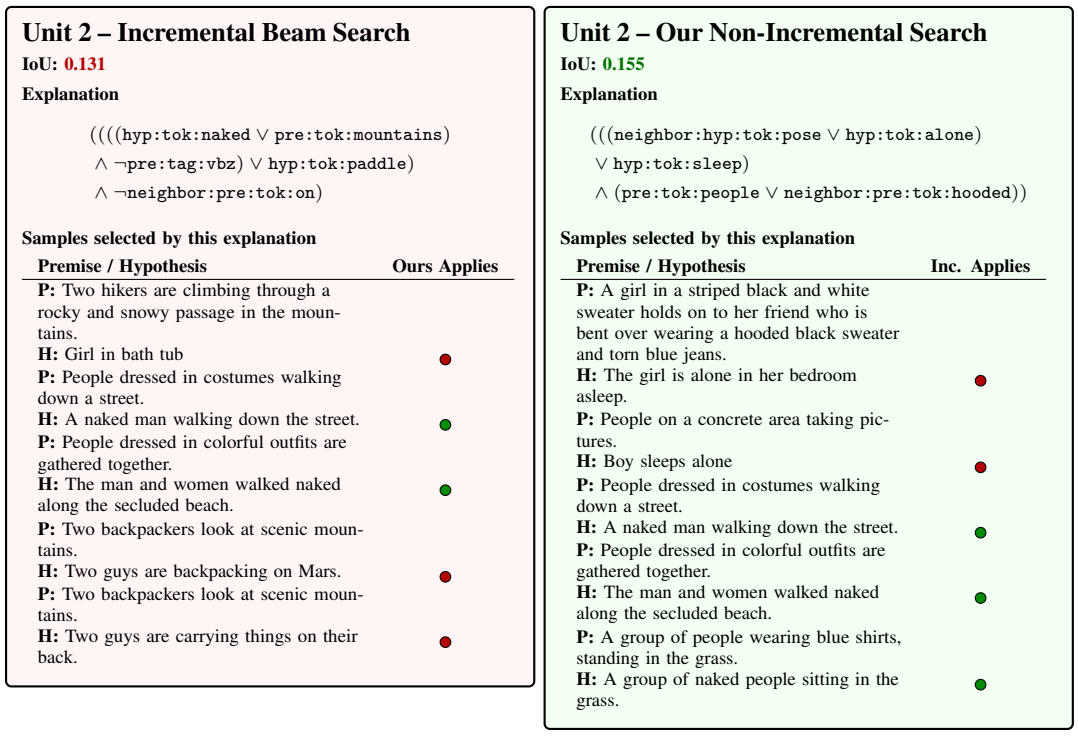

Figure 28: Comparison between incremental and non-incremental compositional explanations for Unit 2 in the LLaMA model.

Unit 3 – Incremental Beam Search

IoU: 0.323

Explanation

((((hyp:tok:person ∧ ¬hyp:tag:.)
∧ ¬hyp:tag:jj) ∧ neighbor:pre:tok:blue)
∨ hyp:tok:least)

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 male child with blond-hair, green eyes, and a sticky looking mouth wearing a multicolored t-shirt has his hand wrapped around a popsicle stick and is chewing and has it inserted into his mouth where you can see 4 bright white teeth on the bottom row.	
H: A person in a shirt	●
P: This young man is wearing blue jeans and a cowboy hat, and is walking away from a calf that is currently roped within an arena.	
H: A person in jeans	●
P: Man and a woman walking on the street	
H: There are at least two people in the picture.	●
P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.	
H: A person in a suit	●

Unit 3 – Our Non-Incremental Search

IoU: 0.328

Explanation

((((hyp:tok:person ∧ ¬hyp:tag:.)
∧ neighbor:pre:tok:blue) ∨ hyp:tok:least)
∧ ¬hyp:tag:jj)

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 male child with blond-hair, green eyes, and a sticky looking mouth wearing a multicolored t-shirt has his hand wrapped around a popsicle stick and is chewing and has it inserted into his mouth where you can see 4 bright white teeth on the bottom row.	
H: A person in a shirt	●
P: This young man is wearing blue jeans and a cowboy hat, and is walking away from a calf that is currently roped within an arena.	
H: A person in jeans	●
P: Man and a woman walking on the street	
H: There are at least two people in the picture.	●
P: Man in a black suit, white shirt and black bowtie playing an instrument with the rest of his symphony surrounding him.	
H: A person in a suit	●

Figure 29: 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:woman)
∧ ¬neighbor:hyp:tok:runner) ∨ hyp:tok:cook)
∨ pre:tok:skydiving)

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 little girl in pink boots runs down the street.	
H: A little girl in pink boots is sitting in a chair.	●
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: 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.	●
P: Girl with helmet riding a brown horse	
H: The girl is wearing tall boots.	●

Figure 30: Comparison between incremental and non-incremental compositional explanations for Unit 4 in the LLaMA model.

Unit 5 – Incremental Beam Search

IoU: 0.187

Explanation

((((neighbor:hyp:tok:post ∧ neighbor:hyp:tok:here) ∨ pre:tok:catcher) ∨ hyp:tok:least) ∧ ¬neighbor:pre:tok:female)

Samples selected by this explanation

Premise / Hypothesis	Ours Applies
P: The person and the dog are sitting on the yellow canoe. **H:** There are at least two figures in the picture.	●
P: Seagulls are flying above the crashing waves. **H:** There are flying animals in this picture.	●
P: A racing motorcycle rider has fallen off his bike and is sliding on the ground while another rider is driving behind. **H:** There are at least two riders.	●
P: A left-handed pitcher wearing the number 40 is throwing a ball to a catcher. **H:** A dog sleeps	●
P: There is not an image here. **H:** There is an image here.	●

Unit 5 – Our Non-Incremental Search

IoU: 0.208

Explanation

(((neighbor:hyp:tok:background ∨ neighbor:pre:tok:racers) ∨ neighbor:pre:tok:empty) ∧ (hyp:tag:ex ∨ pre:tok:sits))

Samples selected by this explanation

Premise / Hypothesis	Inc. Applies
P: Trucks racing **H:** There are vehicles.	●
P: The person and the dog are sitting on the yellow canoe. **H:** There are at least two figures in the picture.	●
P: Seagulls are flying above the crashing waves. **H:** There are flying animals in this picture.	●
P: A racing motorcycle rider has fallen off his bike and is sliding on the ground while another rider is driving behind. **H:** There are at least two riders.	●
P: A race cars muffler catches on fire. **H:** There is a fire	●

Figure 31: Comparison between incremental and non-incremental compositional explanations for Unit 5 in the LLaMA model.

Algorithm 1 Non-Incremental Beam Expansion

Input: Beam B_k, beam history $B_{<k}$, concept set $\mathfrak{L}^1$, operator set $\mathcal{O}$, target length $k+1$
Return: Candidate set $\mathcal{BC}_{k+1}$
$\mathcal{B}_{k+1} \leftarrow \emptyset$
for each $L \in B_k$ **do**
  for each $c \in \mathfrak{L}^1$ such that $c \notin \mathcal{C}(L)$ **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 $|\mathcal{C}(L_{\text{ni}})| = k+1$ **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 $L_1 \in B_k$ **do**
  for each $L_2 \in B_{<k}$ **do**
    if $|\mathcal{C}(L_1) \cup \mathcal{C}(L_2)| = k+1$ **then**
      $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{k+1} \cup \mathcal{R}_{k+1}(L_1, L_2)$
    end if
  end for
end for
for each (L_1, L_2) in B_k **do**
  if $|\mathcal{C}(L_1) \cup \mathcal{C}(L_2)| = k+1$ **then**
    $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_{k+1} \cup \mathcal{R}_{k+1}(L_1, L_2)$
  end if
end for
Return: $\mathcal{B}_{k+1}$

Algorithm 2 Functional-Equivalence-Aware Beam Selection at Level $k+1$

Input: Ordered candidate list $\mathcal{BC}_{k+1}$, activation mask N, concept masks $\mathbf{M}$, beam width b
Return: Beam B_{k+1}, recombination candidates $\mathcal{B}_{k+1}^{FE}$
$B_{k+1} \leftarrow \emptyset$
$\mathcal{B}_{k+1}^{FE} \leftarrow \emptyset$
for each $L_j \in \mathcal{BC}_k$ **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**
    if $IoU(N, \circ_{L_j}(\mathbf{M})) < IoU(N, \circ_{L_{\min}}(\mathbf{M}))$ **then**
      continue
    end if
  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**
    if $|B_{k+1}| < b$ **then**
      $B_{k+1} \leftarrow B_{k+1} \cup \{L_j\}$
    else
      $B_{k+1} \leftarrow (B_{k+1} \setminus \{L_{\min}\}) \cup \{L_j\}$
    end if
  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**
    if $|B_{k+1}| < b$ **then**
      $B_{k+1} \leftarrow B_{k+1} \cup \{L_j\}$
    else
      $B_{k+1} \leftarrow (B_{k+1} \setminus \{L_{\min}\}) \cup \{L_j\}$
    end if
    $\mathcal{B}_{k+1}^{FE} \leftarrow \mathcal{B}_{k+1}^{FE} \cup \mathcal{R}_t(L_j, L_t)$
  end if
end for
Return: $B_{k+1}, \mathcal{B}_{k+1}^{FE}$

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 include a link to the official code 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 H.

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 release and provide open access to the full code and scripts to faithfully reproduce all the experimental results.

   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 H 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. For the human evaluation, we report participant-level exact sign tests and a non-inferiority analysis in Section C.

   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 H.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 G.

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 H 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: [Yes]

    Justification: We release the implementation of the proposed method, together with the scripts and documentation required to reproduce the experiments, in the public repository linked in the paper. Section H additionally documents the datasets, models, experimental settings, licenses, and computational resources used in the experiments.

    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: [Yes]

    Justification: Section C reports the recruitment procedure, qualification criteria, task design, and compensation for the human study. The full instructions provided to participants and screenshots of the study interface are included in the supplementary material.

    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: [Yes]

Justification: The study was pre-approved under an exemption by the University of California, Santa Cruz Institutional Review Board and is currently undergoing the full approval process. Potential risks were disclosed to the 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.