U-CECE: A Universal Multi-Resolution Framework for Conceptual Counterfactual Explanations

Anonymous authors
Paper under double-blind review

Abstract

As AI models grow more complex, explainability is essential for building trust, yet concept-based counterfactual methods still face a trade-off between expressivity and efficiency. Representing underlying concepts as atomic sets is fast but misses relational context, whereas full graph representations are more faithful but require solving the NP-hard Graph Edit Distance (GED) problem. We propose U-CECE, a unified, model-agnostic multi-resolution framework for conceptual counterfactual explanations that adapts to data regime and compute budget. U-CECE spans three levels of expressivity: atomic concepts for broad explanations, relational sets-of-sets for simple interactions, and structural graphs for full semantic structure. At the structural level, both a precision-oriented transductive mode based on supervised Graph Neural Networks (GNNs) and a scalable inductive mode based on unsupervised graph autoencoders (GAEs) are supported. Experiments on the structurally divergent CUB and Visual Genome datasets characterize the efficiency–expressivity trade-off across levels, while human surveys and LVLM-based evaluation show that the retrieved structural counterfactuals are semantically equivalent to, and often preferred over, exact GED-based ground-truth explanations.

1 Introduction

The rapid advancements of artificial intelligence applications have widely sparked the interest of researchers and commercial users, boasting beyond human-level successes in Computer Vision (Chang et al., 2026), Multimodality (Yuan et al., 2025), Natural Language Processing (Zheng et al., 2025) and other fields (Gao et al., 2026). At the same time, the intricate and opaque nature of these black-box systems raises concerns regarding their fairness, hidden risks and rationality in decision-making (Caton & Haas, 2024). To harness these impressive capabilities while cultivating trust between AI models and humans, explainability becomes critical leaving it up to the field of Explainable Artificial Intelligence (XAI) to scrutinize complex models and impenetrable reasoning processes (Ali et al., 2023).

The ever-increasing complexity of contemporary models requires transitioning from *ante-hoc* to *post-hoc* explainability techniques: even though inherent interpretability of neural models was deemed as a necessary pre-requisite in critical domains, promoting ante-hoc explainability as the right direction (Rudin, 2019), the sheer adoption and reliance on complex systems such as Large Language Models (LLMs) underscores the need for alternatives. Post-hoc XAI techniques accept the elevated difficulty of interpreting decisions in order to maintain high performance, permitting model explanations in micro- or macro-level. More specifically, mechanistic interpretability covers an array of techniques satisfying micro-level explanations, suggesting a bottom-up approach of opening the black box (Bereska & Gavves, 2024). Ultimately, it aims to reverse engineer a primarily black-box model by offering deep and causal insights through localization and tracing information across the network, while also setting the starting point for model editing endeavors in cases where compliance is mandatory (Guo et al., 2025). Nevertheless, the need for white-box access, as required in the mechanistic case, is far from realistic in several scenarios, e.g. when utilizing proprietary models. On the other side of the spectrum lies concept-based explainability, with the majority of works accepting the black-box nature of models. An approximation of the model is performed exploiting high-level conceptual cues that lie close to human perception in a top-down fashion (Poeta et al., 2025).

In the latter case, manipulation of model inputs allows the reporting of measurable changes in the output, therefore delineating relationships between input concepts and model decisions. Throughout this process, we are able to explain black-box models without requesting *any* knowledge regarding their inner structure and workings, while providing explanations in human-understandable terms, alleviating the need for expertise when interpreting AI systems in general. Nevertheless, input-output probing should be performed in a well-defined and informative way, since the search space of input-output relationships can be large -even infinite- in the greedy case. For example, in cases of tabular input data, numerical features can be slightly altered (e.g. change a decimal digit), or receive major changes (e.g. increase or decrease the value of a feature by orders of magnitude) and everything in between. This possibility is applicable to all features and data samples, resulting in infinite combinations in practice, without any guarantee that meaningful patterns would occur from each combination. Thus, a heuristic strategy should apply some restrictions on such manipulations, instructing causal pathways between restricted input manipulations leading to observable outcomes.

The notion of minimal input changes resulting in measurable output changes has been extensively studied under the term of *counterfactual explanations* (Guidotti, 2024). Such explanations are governed by causal guarantees regarding the input-output relationships they probe (Stepin et al., 2021), staying aligned with human perception of cause-effect phenomena. In the case of concept-based explanations, a counterfactual question would explore alternative scenarios to the observed one, by answering *which concept X should be changed to X' so that a model's outcome Y changes to Y', while the transition from X to X' is minimal?*. Under this formulation, defining minimality between X and X' becomes crucial, even though multiple interpretations can be assigned to this term. Since concepts can be directly linked to semantics, imposing minimal semantic distances forms a valid strategy for crafting *counterfactual interventions*.

Despite the success of conceptual approaches in literature (Filandrianos et al., 2022; Dervakos et al., 2023; Dimitriou et al., 2024b), existing methods often force a trade-off between expressivity and computational efficiency. While simple sets of atomic concepts are lightweight, they fail to capture the relational context of a complex scene - such as the difference between a "helmet hanging on a bike and a person riding the bike" versus a "person wearing a helmet riding a bike". Conversely, explicit graph-based representations can offer high fidelity in many cases but suffer from the NP-hard complexity of requiring the solution of the Graph Edit Distance (GED) problem for accurate counterfactual retrieval.

In this paper, we propose ***U-CECE**: A **U**niversal **C**ounterfactual **E**xplanations via **C**onceptual **E**dits* framework. Unlike prior disjointed methods, U-CECE provides a unified architecture that can be adaptively scaled in terms of complexity based on the specific data regime and application requirements. Our framework integrates four distinct modules into a cohesive hierarchy:

- Level 1 (**U-CECE-Atomic**): A baseline utilizing atomic concept sets for rapid, broad explanations.
- Level 2 (**U-CECE-Relational**): A "sets-of-sets" approach that captures simple object interactions via "rolled-up" concepts.
- Level 3 (**U-CECE-Structural**): The highest fidelity tier, representing scenes as full concept graphs. This tier can adaptively fork into two operational modes: (a) **Transductive** Mode: A precision-focused path for limited quantity of data that utilizes supervised Graph Neural Networks (GNNs) to approximate ground-truth GED, and (b) **Inductive Mode**: A scalable path when training data is abundant that leverages unsupervised retrieval through Graph Autoencoders (GAEs) for real-time inference.

Our current paper marks several key contributions: **I)** The introduction of a model-agnostic framework that harmonizes diverse conceptual representations into a single unified multi-resolution pipeline. **II)** An adaptive retrieval strategy that trades off fidelity and scalability through transductive or inductive retrieval, making conceptual counterfactuals practical under different data and compute regimes. **III)** Strong empirical evidence that approximate conceptual counterfactuals preserve semantic faithfulness: across benchmarks, the framework characterizes the efficiency–expressivity trade-off, while human and Large Vision-Language Model (LVLM)-based studies show that retrieved counterfactuals are semantically aligned with, and often preferred over, exact ground truth references.

2 Background on Counterfactual Explanations

The journey toward transparent AI decision-making has addressed major challenges over the years, aiming to answer the question of how a model arrives at a particular response based on the evidence in its input. The rather static nature of a *"How does instance X influence model decision Y"* query, targeting the invocation of statistical associations in data, can be revised to allow for dynamic monitoring of input-output relationships, providing a deeper understanding of cause-effect trajectories in model behavior. A *what if* reframing indeed addresses how an intervention $X \rightarrow X'$ on the input data would stimulate a modification in Y, or reversely, how we could force an alternative outcome Y' to happen given X. By adjoining a vast amount of such interventions, we are able to form an "imaginary" reality within which retrospective queries such as *"would Y remain Y if X' had happened instead of X?"* or *"would some Y' have occurred if X' had happened instead of X?"* are compared to the existing reality, where Y appears if X happens. This comparison ultimately constitutes counterfactual explanations, perfectly aligning with black-box systems due to their observational nature (Pearl & Mackenzie, 2018). Formally, considering a model M for which $M(X) = Y$, a counterfactual explanation comprises an instance X' for which $M(X') = Y' \neq Y$, while the distance $d(X, X')$ is minimal.

Low-level Methods. Counterfactual explanations have already populated the spectrum of explainability literature. With a focus on visual classifiers, which constitute our experimental basis, there have been several endeavors operating in low-level pixel spaces: edits on specific areas of the image are proposed in order to alter a classifier's predictions. Generating counterfactual regions on images requires respecting the surrounding areas in order to contribute to plausible counterfactual instances (Chang et al., 2018), Similarly, Goyal et al. (2019) propose minimal alterations of image regions so that an image is classified as an alternative distractor class, defining minimality as the number of regions modified. Minimal additions have been also regarded as a technique for generating counterfactuals in certain datasets (Luss et al., 2019). Getting closer to the semantic edits case, minimizing the loss between an initial image and an image with manipulated attributes can instruct counterfactual generation with advanced actionability (Liu et al., 2019). With the advent of more potent generative models, such as Diffusion models (Ho et al., 2020), visual quality and feature faithfulness of generated counterfactuals reach new levels, while the guidance of synthesizing edits becomes more accurate, thus approaching conceptual desiderata (Augustin et al., 2022; Sanchez & Tsaftaris, 2022; Jeanneret et al., 2024; Deja et al., 2023; Schrodi et al., 2025; Weng et al., 2024; Jeanneret et al., 2023; Sobieski et al., 2025; Komanduri et al., 2024; Motzkus et al., 2024).

Concept-Based Methods. Moving beyond pixel-level manipulations, recent literature has shifted toward conceptual counterfactuals, where perturbations are defined in a high-level, human-understandable feature space. Unlike pixel edits, which can lead to biased or cryptic outputs, conceptual edits modify symbolic attributes while maintaining semantic coherence. While approaches such as Concept Activation Vectors (CAVs) (Kim et al., 2018; Abid et al., 2022), clustering of embeddings (Ghorbani et al., 2019), or integrating explicit intermediate concepts through concept bottleneck models (CBMs)(Koh et al., 2020) provide ways to define concepts in neural latent spaces, these methods are limited when it comes to structured reasoning about hierarchical and relational knowledge. Most concept-based methods treat concepts as isolated variables rather than interconnected entities. The Conceptual Edits as Counterfactual Explanations (CECE) framework (Filandrianos et al., 2022) addressed this by treating an image as a set of atomic concepts and defining distance via a hierarchical taxonomy. This work was later extended by the Semantic Counterfactuals (SC) approach (Dervakos et al., 2023), which introduced rolled-up concepts to capture relationships within a set-of-sets structure. The same semantic framework paired with a generation module was able to identify semantic gaps between humans and neural classifiers, establishing conceptual counterfactuals as a means of faithfully interpreting opaque models for classification (Spanos et al., 2025). However, when used as stand-alone modules, these advancements often under-represent the structural complexity of real-world scenes, as they cannot fully model the intricate dependencies between objects.

Graph Retrieval for Counterfactuals. To achieve full data expressivity while preserving conceptual desiderata, counterfactual discovery has been reframed as a structural retrieval problem using scene graphs. The Scene Graph Counterfactual Explanations (SGCE) framework (Dimitriou et al., 2024b) represents a significant departure from set-based conceptual models. In SGCE, images are represented as scene graphs

where nodes represent objects and edges represent their pairwise relationships. This allows for more descriptive, accurate, and human-aligned conceptual explanations by capturing the significance of semantic edges in the presence of intricate relationships. The utilization of scene graphs necessitates the use of Graph Edit Distance (GED), which quantifies the minimum cost of transforming one conceptual graph into another via node and edge edits. However, because the exact computation of GED is an NP-hard problem, its application in black-box conceptual interpretability requires efficient approximation techniques. The challenge lies in bypassing the exhaustive search of graph similarity for all input pairs, which is an integral but computationally expensive part of the counterfactual computation process. The acceleration of GED has been pursued in prior literature via efficient approximations, such as reductions to linear assignment problems (Jonker & Volgenant, 1988; Riesen & Bunke, 2009; Fankhauser et al., 2011) or redefinitions of the edit costs (Serratosa, 2014), as well as graph kernels (Neuhaus & Bunke, 2007). More recent works leverage Graph Neural Networks (GNNs) to obtain lower-dimensional representations, employing graph matching networks for structural similarity, and neural frameworks for learning complex graph distance functions through local or global contexts (Bai et al., 2019a; Li et al., 2019; Ranjan et al., 2022; Cheng et al., 2025). With graph embeddings at hand, counterfactual graph matching becomes equivalent to finding the closest graph embedding pairs, as showcased in SGCE, which adopted the supervised siamese GNN component from Bai et al. (2019a). Another branch in recent literature delves specifically into scene graph similarity, which is highly relevant for this study. Supervised methods, such as IRSGS (Yoon et al., 2021) or Hi-SIGIR (Wang et al., 2023), typically rely on similarity labels derived from image captions or sentence embeddings to train GNNs. However, these are often limited by the high cost of label generation and the inherent inconsistency of caption-based similarities. In contrast, unsupervised techniques based on Graph Autoencoders (GAEs) eliminate the need for labeled training data, while significantly accelerating the training burden (Dimitriou et al., 2024a), even though default GAE baselines fall behind their supervised counterparts in terms of retrieval performance. In a similar research path, the SCENIR framework (Chaidos et al., 2025) introduces a competitive GAE-driven architecture that achieves state-of-the-art concept-aware image retrieval, while advocating for GED as a deterministic and robust ground truth measure for evaluation. By integrating different parts of these diverse strategies, the U-CECE framework provides high structural expressivity while maintaining the flexibility to choose between high-precision supervised matching or scalable inductive retrieval.

3 The U-CECE Framework: Multi-Resolution Semantics

In the following sections, we introduce U-CECE, a unified framework designed to provide semantic-aware counterfactual explanations across varying levels of relational expressivity. This multi-resolution approach allows the framework to adapt to the underlying complexity of data samples and the specific requirements of the downstream application. The architectural pipeline, illustrated in Figure 1, commences with the essential step of *concept abstraction*, which converts raw inputs into human-understandable terms. We then define the formal symbolic hierarchies and taxonomies used to quantify semantic distances. Finally, we detail the three distinct levels of data expressivity - Atomic, Relational, and Structural - and the specialized U-CECE components developed to navigate the unique computational and semantic challenges of each.

3.1 Concept Abstraction

U-CECE initiates by mapping low-level inputs (e.g. pixels) to high-level symbolic spaces required for conceptual reasoning. For instance, considering the desk setup of the Input Image (Figure 1), raw visual features are represented as a distinct conceptual vocabulary. This process abstracts the visual scene components into distinct symbolic terms, e.g. screen(A), keyboard(B), and table(C). Crucially, it also captures the inherent relationships that define the scene structure, such as the spatial context of screen–on–table, keyboard–on–table, or keyboard–nextto–computer. As shown in the base of our architecture (Figure 1), this layer functions as a modality-agnostic interface that enables the subsequent multi-resolution expressivity levels.

To populate the conceptual layer of U-CECE, semantic metadata can be sourced through either existing annotated datasets or automated extraction pipelines. In many research settings, concepts are already available through human-curated datasets where annotations, captions, or metadata explicitly link pixels with semantics. Notable examples include Visual Genome (Krishna et al., 2017), which provides dense scene

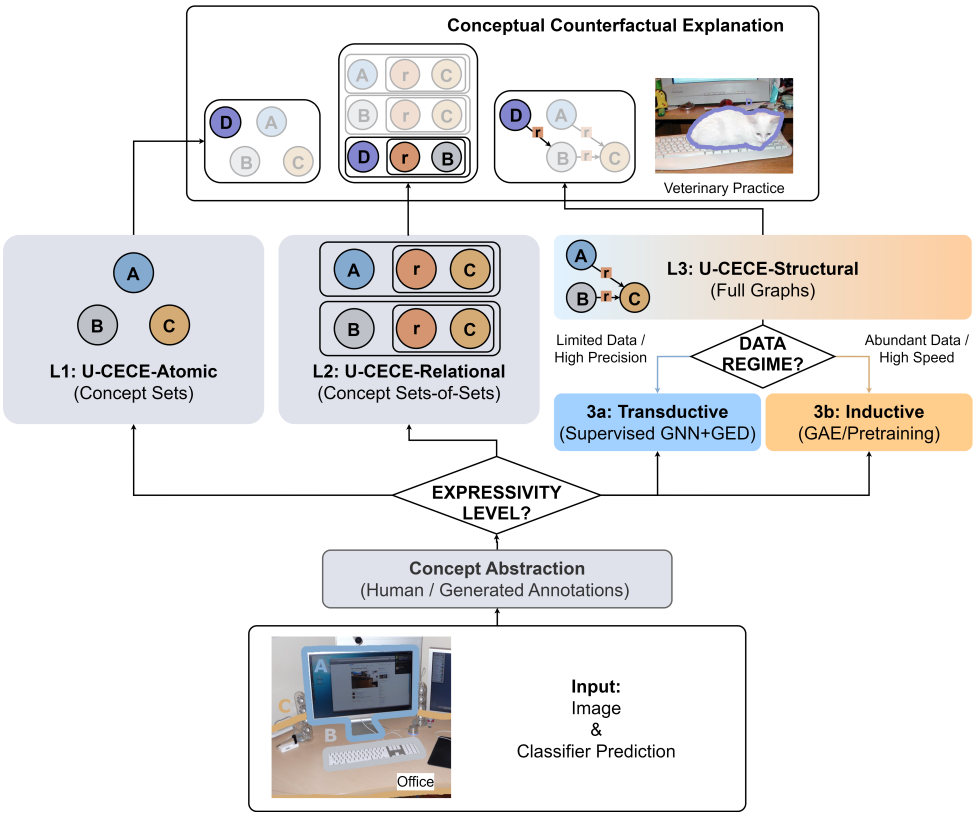


Figure 1: **The U-CECE Framework** for Multi-Resolution Conceptual Counterfactuals. The pipeline begins with Concept Abstraction, mapping raw inputs to a symbolic vocabulary (Computer(A), Keyboard(B), Table(C), Cat(D)) and their relations (on(r)). Segmentation visualizations are for clarity, not to imply a specific abstraction pipeline. The central Expressivity Level decision directs queries through three increasingly complex representations: Atomic Sets (L1), Sets-of-Sets (L2), or Full Scene Graphs (L3). In the high-fidelity L3 path, U-CECE forks into different engines based on the Data Regime: Supervised Transductive GNNs (3a) for high-precision matching on limited data, or Unsupervised Inductive GNNs (3b) for scalable retrieval on abundant data. All paths converge to provide a Conceptual Counterfactual Explanation - such as flipping an "Office" prediction to "Veterinary Practice" by introducing a cat on a keyboard.

graph annotations encompassing objects, attributes, and relationships, or COCO (Lin et al., 2014), which offers captions and annotated bounding boxes. Alternatively, in the absence of manual ground-truth (GT) labels, a common approach in the visual domain is Scene Graph Generation (SGG), where an image is parsed automatically into objects, attributes, and pairwise relations. In this setting, detectors identify objects (e.g., dog, person, bike), while relational models capture interactions (e.g., person–rides–bike, dog–next to–person). Recent SGG research has advanced along several directions: improving object and relation detection with transformer-based architectures (Zhao et al., 2021; Kong et al., 2022), mitigating dataset biases through debiasing and causal inference techniques (Tang et al., 2020; Yang et al., 2021; Chen et al., 2021), and enhancing generalization with open-vocabulary or weakly supervised methods (Lu et al., 2016; Zareian et al., 2020; Dong et al., 2022).

Careful consideration of granularity, domain specificity, and the intended audience ensures that these concepts serve as a faithful and interpretable basis for counterfactual reasoning. This is the essence of “choosing your data wisely” (Dervakos et al., 2023): the selected concept vocabulary directly shapes both the fidelity and the usefulness of explanations. Building on this, once these concepts are curated and extracted, organizing them into expressive structures ranging from atomic sets to full graphs is essential for enabling the effective multi-resolution explanations provided by the U-CECE framework.

3.2 Representation Hierarchy: From Atoms to Graphs

The U-CECE framework, being a conceptual explainability approach, aims to shift the focus from low-level, often unintuitive, features to high-level, human-understandable semantic concepts. While deep learning architectures typically rely on pixel-level correlational patterns (e.g., pixel brightness, color), we follow the foundational principle of Conceptual Edits as Counterfactual Explanations (CECE) (Filandrianos et al., 2022): Instead of identifying pixels to be perturbed, we identify the minimal set of semantic edits (e.g., "adding a helmet" or "removing a laptop" instead of "decreasing the pixel intensity in this region") required to flip a model's decision. This way, U-CECE enables structured reasoning across multiple dimensions, harnessing hierarchical taxonomies—and extending them to relational structures for higher expressivity—to provide an effective pipeline for navigating the trade-offs between symbolic abstraction and computational efficiency. Ultimately, U-CECE ensures that the minimal counterfactual edit is found not in a raw feature space or a disjointed set, but within a semantically coherent and organizationally sound knowledge representation.

To enable multi-resolution counterfactuals, we must first formalize how concepts are structured and related within our framework. U-CECE treats the representation of a data sample not as a static label, but as a symbolic entity that can scale in complexity from a simple category to a dense relational graph.

3.2.1 Symbolic Atoms and Taxonomies

At the most fundamental level, we define the building blocks of our semantic space. Formally, a *concept* C within our framework is defined as any of the following:

- An **atomic concept** A (e.g., cat, laptop), representing a basic category.
- The **top concept** $\top$, representing the universal domain.
- The **bottom concept** $\bot$, representing the empty or impossible category.

These concepts are organized into *taxonomies*, i.e. formal hierarchies termed **TBoxes**. For example, the axiom $\mathsf{cat} \sqsubseteq \mathsf{animal}$ indicates that "every cat is an animal." Visualized as a directed acyclic graph, these hierarchies allow for automated semantic reasoning: recognizing a cat inherently implies the presence of more generic concepts like LivingThing through transitivity. This hierarchical grounding is essential for Level 1 (Atomic) counterfactuals, where the distance of an edit is determined by the path through this taxonomy.

3.2.2 Relational Enriched Concepts (Roles)

Beyond simple inclusion, concepts participate in directed relationships, formally termed *roles* (e.g., person–wears–helmet). To maintain computational tractability while increasing expressivity, we adopt the "rolling up" strategy of the SC framework (Dervakos et al., 2023). By incorporating outgoing relationships directly into node labels, we produce enriched concepts of the form $\exists r.C$ (e.g. $\exists$wears.helmet), interpreted as "something that wears a helmet." This allows a **set-of-sets** representation, effectively preserving the relational structure within a symbolic format, bridging the gap between atomic sets and full graphs.

3.2.3 Structural Scene Graphs

The final step in our formalization incorporates full interactions through **graph-structured representations**. While sets and "rolled-up" concepts are computationally efficient, they can obscure critical distinctions in complex scenes. Consider the visual domain, where a scene might contain concepts such as man, on.bike, and wearing.helmet. Representing these only as an unordered label set obscures important distinctions - for example, whether there are multiple riders, and if the helmets belong to the men on bikes or to other men standing nearby. Such ambiguities are not trivial; they define the difference between safe and unsafe scenarios in high-stakes domains. To resolve this, we leverage the graph-based formalism established by the SGCE framework (Dimitriou et al., 2024b). Each data sample (image I) is represented as a concept graph $G = (V, E)$, where nodes V correspond to objects (with associated feature vectors X) and edges E capture their explicit pairwise relationships, typically encoded in an adjacency matrix A. In this structural view, the

underlying hierarchical semantics of the TBox remain intact, but they are now embedded within a topology that is both richer and more faithful to the underlying data. This structural fidelity forms the basis for our Level 3 expressivity tier, where counterfactuals are discovered through graph matching.

3.3 Expressivity Levels

The U-CECE framework is built on the principle that not all data samples or application domains require the same level of relational detail. To this end, we formalize three tiers of expressivity, each defining "minimality" and "distance" according to the complexity of the underlying symbolic representation.

3.3.1 Level 1 (Atomic)

At the most fundamental resolution, U-CECE-Atomic leverages the principles of the CECE framework, representing an image as a set of atomic concepts. The "minimality" of a counterfactual is governed by the *conceptual distance* (d_c) within a structured semantic space. Rather than treating concepts as isolated labels, they are embedded in a hierarchical structure, denoted $TBox(V_T, E_T)$. Nodes v_T correspond to concepts (e.g., cat, mammal), while edges E_T encode semantic inclusion relationships (e.g., every cat is also a mammal). The distance $d_c(A, B)$ between two concepts A, B is defined as the length of the shortest path between them:

$$d_c(A, B) = \min_{p \in P(A,B)} \sum_{(u,v) \in p} w(u, v) \tag{1}$$

where $P(A, B)$ is the set of all paths connecting A and B in $TBox$ and $w(u, v)$ is the weight or cost of the transition between concepts. Typically, this cost defaults to 1 but can be weighted to reflect domain-specific actionability. From a computational perspective, $P(A, B)$ are computed using Breadth First Search (BFS), unless weighted edges are introduced, in which cases the Dijkstra algorithm is executed instead. If no path exists (disconnected components), $d_c(A, B) = \infty$, indicating that the corresponding edit is infeasible.

A graph-based notion of distance captures semantic relatedness (concepts lying closer in the hierarchy are cheaper to substitute, while also being semantically closer), and provides a principled way to assign costs to *edit operations* for counterfactuals. In our paradigm, the following three edit operations can be implemented:

- **Replacement ($\boldsymbol{R}$):** Substituting concept A with B has a cost equal to their distance $d_c(A, B)$.
- **Deletion ($\boldsymbol{D}$):** Removing concept A is treated as replacing it with a generic placeholder, with cost proportional to its distance from the top-level category $\top$ ($d_c(A, \top)$).
- **Insertion ($\boldsymbol{I}$):** Introducing a new concept B is modeled symmetrically, as the distance from the generic placeholder $\perp$ to B ($d_c(\perp, B)$).

As illustrated in Figure 1, this deterministic approach allows U-CECE-Atomic to retrieve counterfactuals by treating the abstracted concepts as an unordered set. For our office scene example (initially containing A, B, and C), the L1 path identifies the minimal edit necessary to flip the classification to "Veterinary Practice". While for illustrative simplicity the optimal counterfactual image remains the same across all expressivity levels in this instance, from an explanatory standpoint, the mere addition of concept D (cat) to the existing set suffices at this resolution. Overall, the L1 tier provides a high-speed, transparent explanations, identifying D as the key semantic component required to change the model's prediction in the illustrated case.

3.3.2 Level 2 (Relational)

Moving beyond isolated atoms, Level 2 (U-CECE-Relational) adopts the SC approach to capture simple object interactions. In this tier, images are treated as *exemplars* $\mathcal{E}$ - formulated as collections of "roled-up" concepts (i.e., source objects paired with their relational targets). Because this algorithm builds its understanding of the scene exclusively through valid edge connections, any isolated atomic concepts (objects without structural relationships) are excluded from the exemplar. Since each exemplar is essentially a set of concepts, distances are computed hierarchically generalizing the atomic metric to a *set-edit distance*.

Set-level edit distance. Given two finite sets $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, the set-edit distance is computed via a minimum-cost bipartite matching:

$$d_{\text{set}}(A, B) \;=\; \min_{M \in \mathcal{M}(A,B)} \sum_{(a,b) \in M} d_r(a, b) \tag{2}$$

where: $\mathcal{M}(A, B)$ is the set of all full matchings between A and B (allowing dummy matches with a fixed penalty δ if $|A| \neq |B|$). The distance $d_r(\cdot, \cdot)$ evaluates the transformation of the entire relational unit. While for atomic concepts C, D, their distance is given by the metric $d_c(C, D)$ defined in Section 3.3.1, for roled-up concepts of the form $\exists r.C$ and $\exists s.D$, we re-define it to $d_r(\exists r.C, \exists s.D)$ and treat them as two-element sets $\{r, C\}$ and $\{s, D\}$. Their distance is then computed as a set-edit distance between these trivial sets:

$$d_r(\exists r.C, \exists s.D) \;=\; d_{\text{set}}(\{r, C\}, \{s, D\}) \tag{3}$$

In the paradigm discussed within this paper, exemplars are essentially images within a dataset modeled as a "bag of relational facts". The distance between two images $\mathcal{I}_1, \mathcal{I}_2$ is then:

$$d_{\mathcal{E}}(\mathcal{I}_1, \mathcal{I}_2) = d_{\text{set}}(\mathcal{I}_1, \mathcal{I}_2) \tag{4}$$

where unmatched elements incur a penalty δ. Level 2 provides a crucial middle ground: it preserves the semantics of interactions (the "on-ness" or "near-ness") without the computational complexity of full graphs.

As shown in the middle path of Figure 1, U-CECE-Relational moves beyond the mere presence of concepts to explain their structural context. For our example, the L2 path explicitly identifies the relational transition. The resulting explanation highlights the addition of the roled-up concept $\exists$on.keyboard (r.B) which refers to the added cat concept. While this level provides a significantly more nuanced explanation than a simple label, it serves as a semantic hint, establishing that the relationship exists within the scene's set of features. However, this strict edge-centric focus means the metric inherently penalizes structural shifts while remaining blind to unconnected background objects. For example, if an isolated but highly relevant object (e.g. doctor) were present in the scene without any explicit edges linking it to other entities, the relational metric would entirely ignore its presence. Consequently, while L2 successfully captures relational state changes, it leaves the explicit, point-to-point topological binding to the higher-resolution structural tier.

3.3.3 Level 3 (Structural)

At the highest resolution, U-CECE-Structural represents scenes as full conceptually-rich graphs $G = (V, E)$. This tier captures the complete topology of the data, ensuring that counterfactuals are sensitive to the specific relationships between multiple instances of the same concept, effectively addressing L2 restrictions.

Finding the minimal counterfactual here equates to solving the **Graph Edit Distance (GED)** problem (Sanfeliu & Fu, 1983). Graph matching stands as a natural extension of placing concepts on a bipartite graph and then calculating the perfect matching; instead of finding the *sets of concepts* to be added, deleted or replaced, we need to adapt and extend this formulation by discovering *matchings between graphs*. In that case, edits correspond to finding the closest subgraphs between a query graph and its candidates and then defining what needs to be changed in the form of neighbor-informed *concept-role-concept* triplets.

Specifically, by assuming a query image I_q and a set of candidates $I_c^i, i = 1, 2, ..., N$, their corresponding scene graphs are denoted as G_q and $G_c^i, i = 1, 2, ..., N$ respectively. If I_q belongs to a class A, a counterfactual image should be any $I_c^i \neq I_q$ that belongs to some class $B^i \neq A$. This translates in finding the closest scene graph $G_c^i \in B^i$ to the query $G_q \in A$, which culminates in the GED problem. By considering a set of graph edits $\{e_1^i, e_2^i, ...e_m^i\} \in \{R, D, I\}$ of non-negative costs $\{c_{e_1^i}, c_{e_2^i}, ...c_{e_m^i}\}$ needed to transit from G_q to some other counterfactual scene graph G_c^i, GED is mathematically formulated as:

$$GED(G_q, G_c^i) = \min_{e_1, e_2, \ldots e_m} \sum_{j=1}^{m} c_{e_j^i} \quad \text{such that } G_q \in A, G_c^i \in B^i, A \neq B^i \tag{5}$$

GED is an NP-hard problem which given $N-1$ graph candidates, needs to be computed $N-1$ times between G_q and all G_c^i in order to conclude to the counterfactual graph $G_c^k \in B^k$, posing an even greater overhead.

The calculation of G_c^k is mathematically termed in the following equation:

$$G_c^k = \arg \min_{G_c^i \in B^i, i=1,2,\ldots,N-1} GED(G_q, G_c^i) \tag{6}$$

Due to its deterministic characteristics, GED proves well-suited for counterfactual graph retrieval, though computational optimizations are necessary to improve its efficiency. U-CECE employs two distinct retrieval modes: a high-precision, computationally expensive path and a less precise path that operates inductively. Both engines utilize an identical retrieval pipeline, differing only in the methodology for computing graph embeddings and the underlying training regime. The matching procedure of the embeddings between G_q and all G_c^i during inference yields the closest graph $G_c^k \in B^k$ to the query G_q. Then GED will have to be calculated only *once*. After that, G_c^k comprises the counterfactual graph of G_q and $GED(G_q, G_c^k)$ provides a measure of conceptual distance. Consequently, the *R, D, I* graph edits $\{e_1, e_2, ...e_m\}$ required to transform $G_q \rightarrow G_c^k$ comprise a *local* counterfactual explanation for these instances, describing *what needs to be altered* in G_q to become indistinguishable from G_c^k, which is equivalent to outlining the conceptual cues - such as a specific relationship between a cat and a keyboard - that discriminate the closest instances from the two classes. This process should be repeated for all the N scene graphs in our test subset, so that we obtain N counterfactual scene graph pairs, by accepting N *GED* operations in total.

As shown in the final path of Figure 1, U-CECE-Structural provides the most complete explanation. While L1 identifies the object (cat) and L2 identifies the interaction (on − keyboard), only the L3 tier explicitly binds these elements together into a cohesive graph, resolving any ambiguity found in the lower tiers. Furthermore, by formulating the explanation as a subgraph, this approach captures unconnected objects, successfully incorporating any isolated nodes that were invisible to the relational metric. This ensures that the explanation reflects not just a general state of the room, but a precise structural configuration: it is the cat itself that is on the keyboard. From a human perspective, this yields a more intuitive counterfactual. While a cat simply sitting in the corner of a room might not be enough to override the "Office" classification, a cat specifically on the keyboard serves as a strong semantic cue that the space is not a functional workplace, shifting the context to one that prioritizes the presence of animals over professional tasks.

4 U-CECE-Structural Retrieval Modes

While GED provides a mathematically rigorous foundation for counterfactual matching, its computational complexity necessitates neural approximations to remain viable for large-scale applications (Dimitriou et al., 2024b). To this end, we introduce the two approximate retrieval scene graph retrieval engines that comprise our L3 mode. Depending on the availability of ground truth (GT) labels and the required inference speed, U-CECE selects between a supervised precision-focused engine or an unsupervised efficiency-focused one. This modularity ensures that U-CECE remains applicable across diverse data regimes -whether utilizing abstracted concepts that stem from human-curated annotations or are automatically generated.

4.1 Transductive Mode: The Precision Path

To implement this structural tier focused on precision, we utilize a transductive, supervised engine (U-CECE-Transductive) that leverages the deterministic strengths of GED. Here, transductive refers to a retrieval setting in which the model is trained with respect to a fixed candidate graph set from the target dataset, rather than being optimized to generalize to unseen graph collections. This mode is designed for scenarios where the "minimality" of an explanation is essential, ensuring that the retrieved counterfactual is the closest possible semantic neighbor within the available data. The core philosophy of the transductive engine is to "learn" the GED metric directly. By training a model specifically on the dataset of interest, we can accelerate the counterfactual search between each G_q to all other G_c^i during inference.

Supervised GED Approximation. The training pipeline considers a GNN (f_θ) tasked to map given semantic graphs G into an embedding space $h_G = f_\theta(G) \in \mathbb{R}^d$, such that geometric proximity in the embedding space reflects semantic similarity as quantified via GED. In technical terms, upon constructing the GT supervision pairs comprising randomly sampled graphs (G^{i_1}, G^{i_2}), a Siamese GNN (Krivosheev

et al., 2020), comprising two identical node embedding units, is trained to produce a fixed-dimensional vector $h_G \in \mathbb{R}^d$ after performing L rounds of message passing. Initial node features $\mathbf{h}_v^{(0)} = \mathbf{x}_v$ are extracted using pre-trained Glove (Pennington et al., 2014) embeddings. Considering $\mathcal{N}(v)$ as the neighborhood of node v, $\mathbf{W}^{(\ell)}$ as learnable weight matrices, and σ as a nonlinear activation function (e.g., ReLU), for each GNN layer $l = 1, ..., L$, each node embedding $\mathbf{h}_v^l$ is updated as:

$$\mathbf{h}_v^\ell = \sigma\left(\mathbf{W}^\ell \cdot \text{AGGREGATE}\left(\left\{\mathbf{h}_u^{(\ell-1)} \mid u \in \mathcal{N}(v) \cup \{v\}\right\}\right)\right), \tag{7}$$

After L layers, global graph embeddings are obtained via pooling operations:

$$h_G = \text{POOL}\left(\left\{\mathbf{h}_v^L\right\}_{v \in V}\right) \tag{8}$$

To enhance structural consistency in the embedding space, we define a GED-driven regression loss function $\mathcal{L}$ for the graph pair (G^{i_1}, G^{i_2}), which enforces the graph embeddings $h_{G^{i_1}}, h_{G^{i_2}}$ to approximate their true semantic distance, by penalizing discrepancies between embedding distances and the GT GED values:

$$\mathcal{L} = (||h_{G^{i_1}} - h_{G^{i_2}}|| - GED(G^{i_1}, G^{i_2}))^2 \tag{9}$$

By incorporating geometric learning principles like Multi-Dimensional Scaling (MDS) (Williams, 2000; Bai et al., 2019b), we ensure that the latent space preserves the global distance topology of the TBox-driven graph edits.

Transductive Training. The training procedure is transductive, i.e. the model s optimized directly on a subset of the graphs it will be tasked to explain. The training procedure is enforced on a predefined random graph subset of cardinality $p < N$ (empirically we set $p \approx N/2$). To maintain efficiency for the transductive engine, the supervision signal (GED) is computed for only $N/2$ graph combinations within the dataset. As demonstrated by Dimitriou et al. (2024b), this significantly accelerates the training process in the supervised setting without sacrificing substantial retrieval accuracy.

Counterfactual Retrieval. After training, counterfactual retrieval is transformed from an NP-hard search into a high-speed vector comparison. Ultimately, by considering a graph $G_q \in A$ as the query, having a graph embedding $h_{G_q} \in \mathbb{R}^d$, its counterfactual graph $G_c^k \in B^k \neq A$, corresponding to the graph embedding $h_{G_c^k} \in \mathbb{R}^d$, is retrieved by maximizing the cosine similarity between their embeddings:

$$G_c^k = \arg\max_{G_c^i \in B^i \neq A} \frac{h_{G_q}^\top h_{G_c^k}}{\|h_{G_q}\| \cdot \|h_{G_c^k}\|}, i = 1, ..., N. \tag{10}$$

4.2 Inductive Mode: The Efficiency Path

The computational overhead of generating pairwise GED supervision signals in the transductive case (even with a reduced number of pairs) poses challenges for large-scale or real-time applications. To address this, we introduce U-CECE-Inductive, a lightweight alternative that circumvents GED-related operations during training. Inductive refers to a model trained to learn reusable graph representations that can generalize beyond the specific graph pairs or candidate set seen during training. This path utilizes unsupervised Graph Autoencoders (GAEs), which learn to encode a given scene graph into a low-dimensional latent representation and then reconstruct it from that representation. By training on individual graphs rather than pairs, these frameworks maintain a training complexity that scales linearly $\mathcal{O}(N)$ with dataset size, as opposed to the quadratic $\mathcal{O}(N^2)$ scaling of supervised metric learning, resulting in significant training time reduction.

In our experiments, we explore a spectrum of unsupervised architectures, where each method introduces specific mechanisms to enhance the quality of the learned latent embeddings.

Graph Autoencoders at their foundational formulation comprise an encoder $f_\theta(G) = \mathbf{Z}$ that maps a graph G with feature matrix $\mathbf{X}$ and adjacency matrix $\mathbf{A}$ to a latent embedding matrix $\mathbf{Z}$, and a decoder tasked to reconstruct the adjacency matrix $\hat{\mathbf{A}}$ via an embedding inner-product (Kipf & Welling, 2016). The

Variational Graph Autoencoder (VGAE) extends this by learning a probabilistic distribution (of mean μ and variance σ) rather than deterministic embeddings. The VGAE objective function balances reconstruction accuracy with a regularization term ensuring the latent distribution approximates a Gaussian prior:

$$\mathcal{L}_{\text{VGAE}} = \mathcal{L}_{\text{edge_recon}} + \mathcal{L}_{\text{KL}} \tag{11}$$

where $\mathcal{L}_{\text{edge_recon}} = \mathbb{E}[\log p(\mathbf{A}|\mathbf{Z})]$ maximizes the likelihood of the reconstructed graph structure (implemented as sampled Binary Cross-Entropy loss between edge predictions and ground truth adjacency matrix), and $\mathcal{L}_{\text{KL}}$ represents the Kullback-Leibler divergence between the learned distribution $q(\mathbf{Z}|\mathbf{X}, \mathbf{A})$ and the prior $p(\mathbf{Z})$.

Adversarially Regularized Variational Graph Autoencoders (ARVGA) (Pan et al., 2018) integrate principles from Generative Adversarial Networks (GANs) to further enforce a robust latent space distribution. They employ a discriminator network $\mathcal{D}$ alongside the VGAE generator $\mathcal{G}$ (encoder) to distinguish between the learned latent codes and samples from a prior distribution (typically Gaussian). This adversarial training scheme involves a min-max game where the generator tries to fool the discriminator, adding a regularization term that encourages the embeddings to match the prior more effectively than KL divergence alone. The objective combines the standard VGAE loss with this adversarial mechanism:

$$\mathcal{L}_{\text{ARVGA}} = \mathcal{L}_{\text{VGAE}} + \mathcal{L}_{\text{adv}} \tag{12}$$

where $\mathcal{L}_{\text{adv}}$ denotes the standard GAN minimax loss, in which the generator tries to fool the discriminator into classifying the embeddings as samples drawn from the prior p_z.

Graph Feature Autoencoders (GFA) (Hasibi & Michoel, 2021) further boost node information learning by incorporating a dedicated feature decoder, contrary to the topological reconstruction focus (predicting edges) of standard GAEs. Instead of solely reconstructing $\mathbf{A}$, GFA explicitly reconstructs the node feature matrix $\mathbf{X}$, ensuring that semantic attributes are preserved in the latent space. The loss function is modified to include feature reconstruction error:

$$\mathcal{L}_{\text{GFA}} = \mathcal{L}_{\text{VGAE}} + \mathcal{L}_{\text{feat_recon}} \tag{13}$$

where $\mathcal{L}_{\text{feat_recon}}$ is implemented as a Mean Squared Error loss between ground truth $\mathbf{X}$ and predicted $\hat{\mathbf{X}}$.

SCENIR is a state-of-the-art scene graph retrieval framework that synthesizes the aforementioned components into a unified architecture (Chaidos et al., 2025). It employs a split-encoder design (independent GNN branches for μ and σ parameters) to better capture distinct structural and uncertainty features. Furthermore, it replaces the standard inner-product decoder with parallel MLP-based decoders for both edges and features to prevent over-smoothing and capture complex relations. Its training objective combines all previous aspects: reconstruction of features and edges, adversarial regularization, and variational inference:

$$\mathcal{L}_{\text{SCENIR}} = \lambda_1(\mathcal{L}_{\text{feat_recon}} + \mathcal{L}_{\text{edge_recon}}) + \lambda_2\mathcal{L}_{\text{adv}} + \lambda_3\mathcal{L}_{\text{KL}} \tag{14}$$

The defining advantage of SCENIR is its natural alignment with inductive settings, though it remains versatile enough to also support transductive scenarios. While transductive models are "dataset-locked"(trained and evaluated on a fixed set of instances), our inductive engine learns a general transformation function that can generalize to unseen queries at inference time without re-training. Regardless of the specific architecture used, global graph embeddings are obtained via pooling operations on the latent node embeddings, allowing for efficient nearest-neighbor retrieval in the latent space, approximating the retrieval results of GED.

4.3 Performance and Complexity Analysis

The selection between the **Transductive** and **Inductive** engines involves a fundamental trade-off between topological precision and computational scalability. While both seek to approximate GED via latent space retrieval instead of direct GED learning, their operational overhead is governed by distinct complexity classes.

Training Complexity and Scaling. The primary distinction between the two modes lies in the construction of the latent manifold. The **Transductive Mode** relies on supervised metric learning, where the model is optimized to minimize the regression loss between embedding distances and ground-truth GED. For a dataset of N scene graphs, the absolute computational cost encompasses both the GNN forward passes and the generation of the pairwise supervision signal: $\mathcal{O}(N \cdot \text{cost}(\text{GNN}) + N^2 \cdot \text{cost}(\text{GED}))$. Because there are $\binom{N}{2} \in \mathcal{O}(N^2)$ potential pairwise combinations, the quadratic supervision bottleneck strictly dominates the linear message-passing operations. Even with empirical $N/2$ optimization, the effective training complexity reduces to:

$$T_{\text{train(trans)}} \propto \mathcal{O}(N^2 \cdot \text{cost}(\text{GED})) \tag{15}$$

where cost(GED) represents the computational overhead of a single graph matching operation. While exact solvers based on A^* search scale exponentially ($\mathcal{O}(2^N)$), practical implementations of the Transductive engine often utilize the Volgenant-Jonker (VJ) or Hungarian algorithms (Jonker & Volgenant, 1988; Kuhn, 1955) to reduce the matching to a Linear Assignment Problem (LAP). This substitution brings the cost down to $\mathcal{O}(|V|^3)$, where $|V|$ is the number of vertices in the larger graph. However, even with this polynomial reduction, the $\mathcal{O}(N^2)$ scaling of the pairwise training regime remains a significant bottleneck, effectively bounding the Transductive engine to high-fidelity, sparse data regimes.

In contrast, the **Inductive Mode** leverages self-supervision via reconstruction. Because the loss function is computed per graph instance rather than per pair, the $\mathcal{O}(N^2)$ supervision bottleneck is entirely eliminated. The training complexity scales strictly *linearly* with the dataset size, bounded only by the efficiency of the underlying GNN backbone:

$$T_{\text{train(induc)}} \propto \mathcal{O}(N \cdot \text{cost}(\text{GNN})) \tag{16}$$

where L is the number of message-passing rounds, $|E|$ the number of edges, and d the embedding dimensionality. Here, cost(GNN) varies depending on the specific message-passing architecture utilized (e.g., GCN, GIN, GAT). For a network with L layers and an embedding dimensionality of d, the cost generally scales as $\mathcal{O}(L \cdot (|V|d^2 + |E|d))$, accounting for both the node-wise feature transformations and the edge-wise neighborhood aggregations (including attention coefficient computations in GAT). In large-scale datasets, the total number of graphs N vastly exceeds the structural parameters of any individual scene graph ($N \gg |V|, |E|$), making cost(GNN) a dataset-independent constant factor. By confining the complexity strictly to this linear $O(N)$ scaling, the inductive mode enables the processing of vast data distributions that remain computationally inaccessible to transductive supervised methods.

Inference and Retrieval Efficiency. At inference time, both engines utilize an identical retrieval pipeline, transforming the NP-hard structural matching bottleneck into a vector similarity task. Given a query graph G_q, the complexity of retrieving the counterfactual G_c^k involves a forward pass through the GNN f_θ followed by a nearest-neighbor search across the $N-1$ candidates:

$$T_{\text{inf}} \propto \mathcal{O}(\text{cost}(\text{GNN}q) + N \cdot d) \tag{17}$$

where cost(GNNq) represents the linear message-passing cost for the single query graph, and d is the latent dimensionality. However, the **Inductive Mode** provides superior operational flexibility. While the transductive model is effectively "dataset-locked" ($\mathcal{G}_{\text{train}} \approx \mathcal{G}_{\text{test}}$), the inductive engine learns a generalizable mapping $f : \mathcal{G} \to \mathbb{R}^d$. This allows for the immediate generation of counterfactuals for novel, out-of-distribution scene graphs without the need for the $\mathcal{O}(N^2)$ retraining cycle required by transductive supervised signals.

Precision vs. Topological Generalization. The *Precision Path* is optimized for the minimality of the semantic edit, making it the preferable choice for applications where preserving fidelity to the original structure is critical and arbitrary deviations are undesirable. Conversely, the *Efficiency Path* captures the broader manifold of the scene graph distribution. Rather than blindly relying on granular, potentially noisy input edges, this inductive approach leverages the learned global topology to infer structural relationships probabilistically, effectively attaining generalizability.

5 Experimental Analysis

The evaluation of U-CECE is structured across the following primary axes: i) *Performance Across Expressivity Modes:* a traditional benchmarking of the different expressivity levels using standard local counterfactual metrics to establish a baseline of efficacy; ii) *Structural Engine Optimization:* an exhaustive evaluation of various GNN architectures for the Transductive and Inductive engines at the highest expressivity tier, including an assessment of inductive capabilities; iii) *Human-Centric Validation:* an evaluation of explanation quality through manual inspection of qualitative examples and a systematic human survey to validate the semantic faithfulness of our most accurate structural engine against human visual intuition; and iv) *Neural Interpretability:* an exploration of counterfactual interpretability from the perspective of Large Vision-Language Models (LVLMs) as a Judge by recreating our human survey with neural-based annotators to measure the alignment between machine and human perception.

Notably, direct comparisons with external baselines are omitted, as the fundamental retrieval engines utilized within U-CECE have been previously validated against state-of-the-art methods in prior literature (Dimitriou et al., 2024a;b; Chaidos et al., 2025); consequently, our evaluation prioritizes the framework's internal consistency across its various expressivity tiers and its alignment with human understanding.

5.1 Datasets

Conceptual counterfactuals are inherently versatile and applicable across a broad spectrum of data domains. Previous research has established their efficacy on standard XAI benchmarks or automatically generated datasets (Dervakos et al., 2023; Dimitriou et al., 2024b). In this study, our primary objective is to rigorously evaluate the distinct expressivity levels and retrieval modes of U-CECE. To this end, we focus on real-world *image* datasets characterized by high-quality annotations at both the concept and graph levels. These controlled datasets provide the necessary structural richness to be employed horizontally throughout our experimental analysis, ensuring a consistent and robust evaluation of our transductive and inductive engines.

All tiers of expressivity are evaluated on the exact same subset of Visual Genome (VG) (Krishna et al., 2017) images, varying only in the scope of the structural information extracted. Specifically, to serve L1 expressivity, concept abstraction is restricted to isolated object annotations. For L2 expressivity and higher, the representations incorporate full relational data. Furthermore, both L1 and L2 leverage the WordNet synsets provided by VG to enrich their respective elements (utilizing synsets exclusively for objects in L1, and for both objects and relations in L2). For these experiments, we adopt the dataset split introduced by SGCE (Dimitriou et al., 2024b), which treats edge density as a decisive factor for the selected explanation algorithm. Specifically, the VG-DENSE split contains 500 highly interconnected graphs with fewer isolated nodes, while the VG-RANDOM split contains 500 arbitrarily selected graphs. We further use the Caltech-UCSD Birds (CUB) dataset (Wah et al., 2011), which offers fine-grained concept-level annotations of bird parts and their visual attributes. While graph-level annotations are not directly provided, they can be straightforwardly constructed and connected to WordNet (Dimitriou et al., 2024b), resulting in 422 tree-shaped graphs of fairly small depth whose size depends on the number of visible parts in each image.

The roles of these datasets are strategically partitioned across our evaluation axes. CUB, a widely adopted benchmark in visual XAI, serves as the primary benchmark here as well, being used horizontally throughout the entire experimental section; its fine-grained attributes and manageable graph structures make it uniquely suited for evaluating the complete pipeline, ranging from standard quantitative metrics and GNN optimization to the human perception surveys and neural interpretability studies. In contrast, VG is utilized specifically to investigate the impact of structural expressivity. By leveraging VG-DENSE and VG-RANDOM, where edge density serves as a proxy for relational importance, we report on performance across our retrieval engines and provide qualitative insights into how U-CECE handles complex (edge-dense) scenes.

5.2 Model Details and Training Protocols

To evaluate our retrieval engines, we utilize several GNN architectures across two distinct training protocols. For U-CECE-Transductive, we leverage GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2017), and GIN (Xu et al., 2019) within the Siamese similarity framework of SGCE. For the unsupervised U-CECE-Inductive

engine, we employ GAE-based variations (Section 4.2 –GFA, ARVGA). We conduct thorough ablation studies across these variants for both engines to identify the optimal configurations for counterfactual retrieval. Unsupervised GNNs are evaluated in both transductive and inductive settings. While the transductive models are trained and evaluated directly on specific benchmark subsets (consisting of 500 VG graphs and 422 CUB graphs), the inductive capability is assessed by pretraining on larger sets of both datasets (containing 28k graphs for VG / 11k graphs for CUB), strictly excluding the evaluation samples, followed by an optional fine-tuning phase. By anchoring all experiments to the same 500 VG and 422 CUB test graphs, we ensure a consistent and fair comparison between the specialized representations of the transductive models and the generalized latent spaces of the inductive engine. Details are provided in Appendix A.

As classifiers under explanation, we utilize the setup layed out by Dervakos et al. (2023); Dimitriou et al. (2024b), i.e. a ResNet-50 for CUB to ensure experimental alignment and a Places365 model Zhou et al. (2017) to better capture the scene-centric nature of VG, which object-focused ImageNet classifiers are less equipped to represent. For datasets with GT classification labels (CUB) the target class for the counterfactual explanation is selected as the category most frequently confused with the source class (Vandenhende et al., 2022). In scenarios where GT labels are unavailable (VG), the target class is instead defined as the category of the single highest-ranked instance that belongs to any class other than the source one.

To ensure a high-fidelity and internally consistent evaluation across all expressivity tiers of U-CECE, we implement several strategic refinements to the underlying experimental components. To provide a more robust representation of scene semantics, the classifiers are optimized to assign improved labels. The mathematical protocol for calculating structural edits - particularly the averaging of edge-level operations - is standardized to provide a more precise and stable measure of topological change across datasets. For the L1 and L2 expressivity modes, we utilize a more accurate WordNet synset initialization to provide a more rigorous semantic grounding. By ensuring that all submodules operate on a consistent, high-quality baseline, the resulting evaluation more accurately captures the inherent performance of the unified framework. Consequently, while the reported metrics may deviate from those in previous publications due to these optimizations, this standardized approach ultimately maintains empirical integrity.

5.3 Evaluation Metrics

To evaluate the U-CECE-Structural retrieval engines, we utilize ranking metrics to measure the alignment between GNN-based approximate rankings and the exact GED GT. We report Precision@k ($P@k$) and Normalized Discounted Cumulative Gain ($nDCG@k$) for $k \in \{1, 2, 4\}$. While counterfactuals primarily prioritize the top-ranked result ($k = 1$) to ensure structural minimality, the inclusion of $k = 2$ and $k = 4$ during ablation studies provides a more comprehensive overview of the quality and stability of different engine variants. In the default case, all valid retrieved items are considered equally relevant. However, since the top-ranked instance returned by the exact GED search represents the optimal counterfactual, we also employ *binary* metric counterparts. This stricter evaluation measures the engine's precision in recovering the definitive GT for structural minimality. All aforementioned metrics are bounded in the range $[0, 1]$, where higher values indicate better retrieval accuracy and closer alignment with the GT ranking.

More tailored to counterfactuals, we analyze edit-level metrics based on the number of node and edge operations (R, D, I). These counts are averaged across the dataset for the top-ranked retrieval ($k = 1$) and denoted as $\bar{e}_{node}$, $\bar{e}_{edge}$, and $\bar{e}_{total}$. These results are paired with the mean GED ($\overline{\text{GED}}$) of the top-ranked retrievals to offer a view into the cost of the operations as well. These structural metrics have a theoretical lower bound of 0 (no edits needed), with lower values being preferred as they signify greater structural minimality and a more concise transition from the source to the counterfactual state. By quantifying these transitions, we gain insight into the logical complexity of the explanations and the effectiveness of the U-CECE engines in approximating the foundational graph-based reasoning.

When comparing the retrieval power across the U-CECE expressivity tiers (L1-Atomic, L2-Relational, and L3-Structural), we strictly evaluate the first-ranked proposed counterfactual ($k = 1$). In this context, the focus shifts from the performance of the retriever in isolation to a deterministic measure of how closely the final generated explanation approximates the gold standard GED-based counterfactual. To this end, in these comparative experiments, we report the average edit counts and costs, and only $P@1$.

5.4 Results

5.4.1 Expressivity Level Mode Performance

In Table 1, we evaluate the performance of the three U-CECE expressivity tiers in a strict top-1 counterfactual retrieval task. Across all datasets, $P@1$ scores appear relatively low (ranging from 0.078 to 0.248). This is indicative of the challenging nature of the task: retrieving a single, mathematically optimal gold instance from a candidate pool of hundreds of potentially conceptually similar images. This is particularly evident in the CUB dataset, where inter-class variation is minimal. Bird species often share a restricted subset of attributes and differ primarily in the visibility or orientation of specific parts. Consequently, while $P@1$ is conservative, the retrieval quality improves significantly as the rank threshold k increases (Appendix B).

For the CUB dataset, we observe that L1 and L2 tiers actually outperform the L3 engine in $P@1$. This finding is consistent with the nature of CUB graphs, which are essentially tree-shaped structures of low depth representing bird parts without explicit spatial orientation. In such a domain, a coarser explanation often approximates the GT more effectively. Interestingly, while the L2 engine achieves the lowest edge edit count ($\bar{e}_{edge} = 2.9$), its associated cost ($\overline{GED} = 219.638$) is the highest in the cohort. This discrepancy suggests a form of local structural disruption: since the L2 tier groups relations into "triplet sets" and ignores the broader neighboring topology, it may suggest fewer edits that are semantically "expensive" or nonsensical, such as swapping a functional part (e.g., a beak) for a purely visual attribute (e.g., a solid pattern).

The merits of L3 expressivity are most prominent in the VG-DENSE split. Here, the Structural engine achieves a $P@1$ of 0.248, representing a 3.6% and 8.2% improvement over the L1 and L2 tiers, respectively. In scenarios where graphs are highly interconnected, higher structural expressivity is required to navigate the complex relational dynamics and retrieve an accurate counterfactual. Furthermore, the L3 engine achieves the lowest total edit count $\bar{e}_{total}$ and the lowest $\overline{GED}$ cost, proving its efficiency in dense environments. In contrast, VG-RANDOM exhibits trends more similar to CUB. When scene graphs are arbitrarily selected and potentially sparser, the performance gap between L1 (0.216) and L3 (0.210) diminishes. This reinforces our intuition that the necessity for structural detail is a function of relational density; when object interactions are sparse, a set-based or atomic view of concepts is often sufficient for retrieval.

Table 1: Comparison of the three expressivity U-CECE tiers across CUB, VG-DENSE, and VG-RANDOM datasets. For U-CECE-Structural, the best-performing variant and architecture are selected. **Bold** for best result.

	Level	$P@1 \uparrow$	$\bar{e}_{node} \downarrow$	$\bar{e}_{edge} \downarrow$	$\bar{e}_{total} \downarrow$	$\overline{GED} \downarrow$
CUB	Atomic	**0.175**	**5.750**	3.475	9.225	**210.594**
	Relational	0.137	6.175	**2.900**	**9.075**	219.638
	Structural	0.138	7.506	6.038	13.544	217.131
VG DENSE	Atomic	0.212	**4.564**	11.236	15.800	109.170
	Relational	0.166	5.018	10.702	15.720	113.532
	Structural	**0.248**	4.942	**10.312**	**15.254**	**105.194**
VG RANDOM	Atomic	**0.216**	**11.176**	12.216	**23.392**	160.982
	Relational	0.078	13.150	11.868	25.018	179.868
	Structural	0.210	12.392	**11.840**	24.232	**159.356**

The L3 engine emerges as a robust and versatile retriever; even in domains where it does not achieve the highest precision, it remains highly competitive while providing the most verbose and topologically faithful explanations. The L1 engine is remarkably consistent and cheap, making it an excellent candidate for datasets with low relational importance. Finally, the L2 engine shows its greatest potential in dense settings but can suffer from high semantic costs. Ultimately, these quantitative metrics confirm that *no single expressivity tier is universally superior*. Instead, the choice of engine should be informed by the underlying data complexity. To further comprehend the subjective utility of these levels, we provide a qualitative inspection (Section 5.4.3), exploring how such numeric differences translate into human-understandable explanations.

5.4.2 Structural Retrieval Engine Trade-offs

Having established the efficacy of the structural L3 tier, we now provide a fine-grained analysis of the architectural backbones and training paradigms within the U-CECE-Structural engine. This section eval-

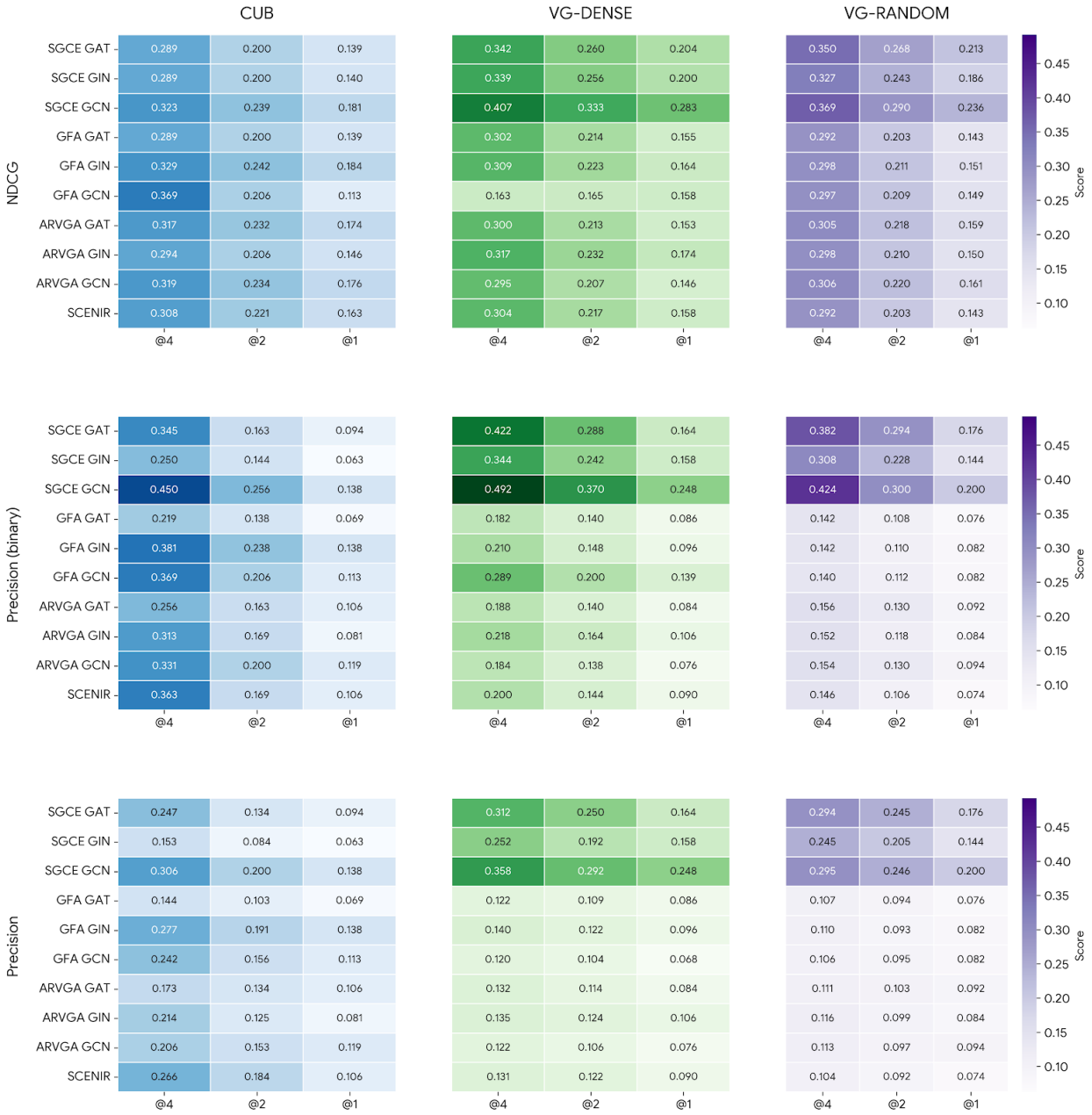


Figure 2: Retrieval performance comparison of underlying GNN components of the U-CECE-Structural (Transductive) mode, including SGCE, GFA, ARVGA, and SCENIR architectures. Ablations on the GNN variant are presented where applicable. Color intensity in heatmaps is normalized globally across all metrics and datasets to facilitate direct comparisons. Distinct color hues represent the different datasets.

uates the trade-offs between supervised and unsupervised regimes, specifically comparing various GNN backbones across the transductive and inductive setups. To establish a fair comparison, we first utilize a standardized transductive protocol for all GNN-based variations. We perform systematic ablations on the message-passing backbones—interchanging GCN, GAT, and GIN —within the supervised SGCE framework (U-CECE-Transductive) and the unsupervised GFA, ARVGA, SCENIR autoencoders (U-CECE-Inductive). SCENIR is implemented with GIN to ensure best retrieval performance (Chaidos et al., 2025).

Table 2: Comparison of edit metrics and $\overline{GED}$ across CUB, VG-DENSE, and VG-RANDOM datasets for all U-CECE-Structural models in the transductive setting.

		CUB				**VG-DENSE**				**VG-RANDOM**			
	Model	$\bar{e}_{node}\downarrow$	$\bar{e}_{edge}\downarrow$	$\bar{e}_{total}\downarrow$	$\overline{GED}\downarrow$	$\bar{e}_{node}\downarrow$	$\bar{e}_{edge}\downarrow$	$\bar{e}_{total}\downarrow$	$\overline{GED}\downarrow$	$\bar{e}_{node}\downarrow$	$\bar{e}_{edge}\downarrow$	$\bar{e}_{total}\downarrow$	$\overline{GED}\downarrow$
SGCE	GAT	7.281	5.506	12.788	224.71	5.278	10.892	16.170	108.64	12.810	11.978	24.788	159.36
	GIN	9.732	10.181	19.913	237.63	5.106	10.766	15.872	116.17	12.792	11.394	24.186	173.72
	GCN	7.506	6.038	13.544	**217.13**	4.942	**10.312**	**15.254**	**105.19**	12.392	11.840	24.232	**159.32**
GFA	GAT	8.775	5.194	13.969	243.44	5.070	11.564	16.634	131.58	12.988	12.412	25.400	198.78
	GIN	6.488	4.650	11.138	217.37	**4.856**	10.786	15.642	129.86	12.916	12.218	25.134	200.97
	GCN	6.456	**3.738**	**10.194**	220.84	5.030	11.432	16.462	132.05	12.696	12.338	25.034	201.04
ARVGA	GAT	8.050	4.900	12.950	238.03	5.062	11.482	16.544	133.48	12.508	12.192	24.700	196.95
	GIN	**6.444**	3.919	10.363	228.74	4.894	10.912	15.806	128.94	**12.234**	**11.794**	**24.028**	198.65
	GCN	6.625	4.106	10.731	223.14	5.048	11.316	16.364	132.29	12.464	12.088	24.552	198.46
SCENIR	GIN	6.369	4.7	11.069	223.91	5.028	11.546	16.574	131.56	12.646	12.352	24.998	197.68

Figure 2 illustrates the retrieval performance for $k \in \{1, 2, 4\}$. Across all heatmaps, we observe a monotonic improvement in retrieval fidelity as k increases. The results confirm the clear superiority of the supervised SGCE models in approximating the GT GED rankings: specifically, SGCE-GCN achieves the most robust performance on the VG subsets, where its ability to model complex relational densities is most pronounced. Interestingly, the performance deltas between architectures are more significant in the P@k and Binary P@k metrics, whereas nDCG@k exhibits a smoother distribution across variants. This suggests that while the exact recovery of top-ranked instances varies by backbone, the overall rank-order stability remains relatively consistent across different supervised GNNs. In contrast, unsupervised models generally underperform in the transductive setting, as they lack the explicit similarity-driven signal used by SGCE. While there is no definitive winner between GFA, ARVGA and SCENIR globally, the models that utilize a feature decoder (GFA, SCENIR) exhibit competitive performance on CUB, particularly for higher k values, suggesting that the concentrated distribution of node and edge attribute semantics in CUB is more readily captured by feature-reconstruction backbones than the more relational, structure-heavy scene graphs of VG.

While the retrieval metrics establish ranking efficacy, the edit metrics reveal a divergence between topological minimality and semantic relevance (Table 2). In VG-DENSE, the supervised SGCE-GCN variant remains the dominant performer, achieving the lowest total edit count ($\bar{e}_{total} = 15.254$) and the lowest semantic cost ($\overline{GED} = 105.19$). For SGCE-GIN, similarly to the retrieval metrics, there is a significant performance drop specifically on CUB ($\bar{e}_{total} = 19.913$, $\overline{GED} = 237.63$); despite extensive optimization, we hypothesize this stems from GIN's rigid focus on structural isomorphism, which causes it to overfit to the uniform anatomical topologies of the bird graphs rather than exploiting the critical semantic variance found within the continuous attributes. Furthermore, in the CUB and VG-RANDOM subsets, we observe a distinct behavior: unsupervised models such as GFA-GCN and ARVGA-GIN achieve lower raw edit counts—frequently nearly 25% lower than their supervised counterparts—yet maintain significantly higher $\overline{GED}$ costs, confirming that while unsupervised autoencoders often find "shorter" topological paths, these transitions frequently result in semantically noisier shortcuts that fail to respect the learned semantic manifold. Consequently, these results reinforce our argument that raw edit numbers are an insufficient standalone metric for conceptual counterfactuals and must be evaluated alongside $\overline{GED}$ to accurately measure the logical complexity and structural fidelity of the retrieved explanations.

Finally, we evaluate the performance of U-CECE-Inductive, utilizing all unsupervised GNN models that can intuitively be adapted to the inductive setup. As detailed in our methodology, the inductive regime offers a significant computational advantage: by pretraining on a massive unlabeled corpus, counterfactual identification is reduced to a single inference pass, eliminating the need for labels or task-specific training. To this end, a detailed evaluation of the U-CECE-Inductive framework across all unsupervised autoencoder architectures

Table 3: U-CECE-Inductive performance across various GNN backbones (SCENIR, ARVGA, GFA). In underlined instances, the inductive model strictly outperforms its transductive counterpart.

	Model	NDCG (↑)			Precision (binary) (↑)			Precision (↑)			Edits & GED (↓)			
		@4	@2	@1	@4	@2	@1	@4	@2	@1	$\bar{e}_{node}$	$\bar{e}_{edge}$	$\bar{e}_{total}$	$\overline{GED}$
CUB	SCENIR	0.328	0.244	0.187	0.288	0.194	0.125	0.220	0.150	0.125	**6.425**	4.750	11.175	225.68
	ARVGA-GAT	0.319	0.233	0.176	**0.388**	0.213	0.125	0.266	0.181	0.125	6.563	4.775	11.338	221.681
	ARVGA-GIN	0.289	0.200	0.139	0.300	0.175	0.088	0.256	0.172	0.088	6.556	5.144	11.700	224.294
	ARVGA-GCN	**0.332**	**0.248**	**0.191**	0.350	**0.225**	0.144	**0.278**	**0.200**	**0.144**	6.894	5.281	12.175	222.15
	GFA-GAT	0.292	0.203	0.143	0.231	0.138	0.081	0.172	0.116	0.081	7.694	5.250	12.944	235.55
	GFA-GIN	0.296	0.208	0.149	0.275	0.163	0.081	0.188	0.141	0.081	7.506	4.644	12.150	234.869
	GFA-GCN	0.325	0.188	0.100	0.295	0.207	**0.148**	0.250	0.163	0.100	6.719	**4.163**	**10.881**	221.563
VG-DENSE	SCENIR	**0.304**	**0.217**	**0.158**	0.188	0.130	0.086	0.125	0.098	0.086	5.032	11.198	16.230	132.38
	ARVGA-GAT	0.298	0.210	0.150	0.186	0.136	0.082	0.138	0.118	0.082	5.208	11.798	17.006	131.964
	ARVGA-GIN	**0.304**	**0.217**	**0.158**	0.210	0.156	0.092	**0.150**	**0.124**	**0.092**	5.210	11.832	17.042	132.688
	ARVGA-GCN	0.296	0.207	0.148	0.174	0.120	0.076	0.122	0.112	0.076	5.228	12.086	17.314	134.63
	GFA-GAT	0.300	0.212	0.153	0.176	0.130	0.084	0.128	0.116	0.084	5.180	11.922	17.102	133.386
	GFA-GIN	0.299	0.212	0.152	**0.212**	0.150	0.086	0.144	0.121	0.086	**4.874**	**11.166**	**16.040**	**128.406**
	GFA-GCN	0.293	0.205	0.145	0.172	0.116	0.072	0.113	0.103	0.072	5.338	12.190	17.528	133.284
VG-RANDOM	SCENIR	0.293	0.204	0.144	0.144	0.114	0.078	0.116	0.101	0.078	12.802	**11.878**	**24.680**	**195.12**
	ARVGA-GAT	**0.305**	**0.218**	**0.159**	0.152	0.124	**0.096**	0.118	**0.102**	**0.096**	13.074	12.682	25.756	195.508
	ARVGA-GIN	0.285	0.196	0.135	0.140	0.102	0.070	**0.119**	0.097	0.070	13.190	12.922	26.112	200.248
	ARVGA-GCN	0.290	0.201	0.141	0.128	0.096	0.074	0.113	0.099	0.074	13.134	12.860	25.994	197.842
	GFA-GAT	0.298	0.210	0.151	0.146	0.116	0.086	0.118	**0.102**	0.086	13.114	13.086	26.200	197.98
	GFA-GIN	0.290	0.202	0.141	0.148	0.108	0.076	0.118	0.100	0.076	**12.620**	12.278	24.898	196.546
	GFA-GCN	0.296	0.208	0.149	0.146	0.108	0.082	0.108	0.094	0.082	13.068	12.870	25.938	200.368

and their respective GNN backbones is presented in Table 3. Comparing these zero-shot inductive results with the previously detailed transductive baselines reveals a striking trend: the inductive models consistently match or exceed their transductive counterparts in global ranking fidelity. For example, on CUB, the inductive SCENIR and ARVGA-GCN achieve nDCG@4 scores of 0.328 and 0.332, respectively—distinctly outperforming transductive SCENIR (0.308). This suggests that large-scale pretraining allows the engines to learn a highly robust global ranking structure, regardless of the specific autoencoder formulation.

However, a subtle trade-off exists regarding exact retrieval. In highly dense environments like VG-DENSE, domain-specific transductive supervision retains a slight edge in Top-1 precision (e.g., transductive SCENIR achieves P@1 = 0.090 compared to its inductive counterpart's 0.086), better isolating exact targets within complex local neighborhoods. Yet, despite this zero-shot evaluation, the inductive suite maintains highly competitive semantic costs. Notably, the inductive GFA-GIN model on VG-DENSE actually improves upon the semantic cost of its transductive version (dropping $\overline{GED}$ from 129.86 to 128.406). This indicates that the generalized inductive space naturally avoids the topologically short but semantically noisy shortcuts more typical of transductive autoencoders, ultimately yielding counterfactuals that are logically sound.

We observe strong relative performance across the inductive models on sparser datasets, most notably in VG-RANDOM, where architectures like ARVGA-GAT achieve highly competitive retrieval metrics (P@1 = 0.096, nDCG@4 = 0.305). This performance profile can be directly attributed to the nature of the pretraining data; since the vast majority of the full VG dataset consists of sparse scene graphs, the inductive latent space is naturally biased toward such topological motifs, allowing it to easily map the VG-RANDOM distribution.

Crucially, the robust generalization of these models represents a significant success for U-CECE. While the L1 and L2 expressivity tiers do not require training and are inherently "inductive," they lack an optimized continuous latent space, making their exact graph-matching inference significantly slower. Thus, U-CECE-Inductive provides a unique and highly desirable combination: the high-expressivity and semantic richness of the L3 tier, coupled with rapid inference and robust generalization to entirely unseen graph topologies.

5.4.3 Qualitative Analysis

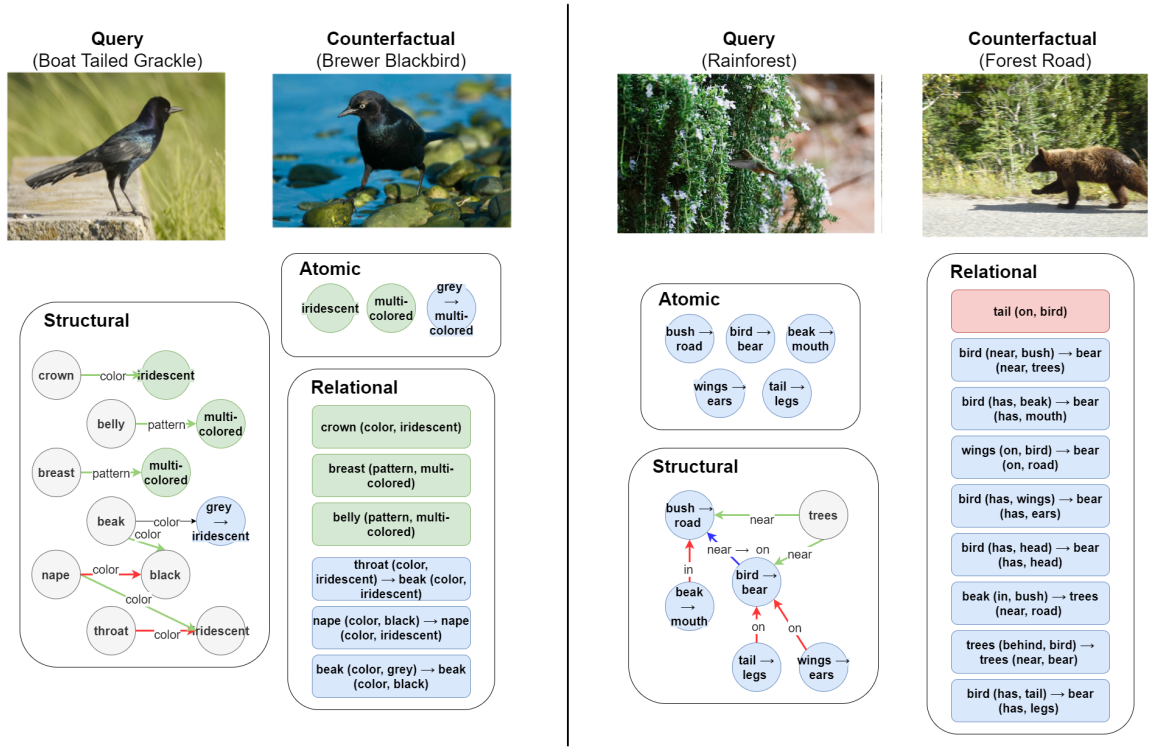


Figure 3: U-CECE expressivity tiers yielding the same counterfactual for different datasets (**left:** CUB example, **right:** VG-DENSE example).

While the previous sections establish a quantitative standard for retrieval, the primary goal of U-CECE is to provide human-interpretable explanations. In this section, we transition from numerical metrics to a visual inspection of the counterfactuals generated at each expressivity tier. For the Structural level, we utilize the SGCE-GCN variant, as it consistently demonstrated the highest retrieval fidelity. By visualizing these transitions, we aim to evaluate how the mathematical minimality captured by different types of GED approximations aligns with logical, intuitive shifts in conceptual reasoning. In all figures, green indicates insertion, red deletion and blue replacement. Grey nodes are context nodes, that do not get edited.

Figure 3 provides a qualitative visual bridge to our earlier quantitative findings, illustrating how the three expressivity tiers interpret the same counterfactual transition. In the CUB example (Left), the L1 tier offers a general sense of the necessary attribute shifts (e.g., color changes) but lacks grounding, as it cannot specify which anatomical parts these attributes refer to. The L2 level serves as a "middle man," adding this missing context; however, it occasionally introduces confusion during replacements. For instance, the transition of "throat (color,iridescent)" to "beak (color,iridescent)" suggests a peculiar "transfer" of properties between parts rather than a logical state change. In contrast, the L3 tier provides the full conceptual picture: it clarifies that attributes are being added or modified on specific, existing parts without misrepresenting the parts themselves—accurately reflecting that black color is added to the beak while iridescent tones are mapped to the nape. This aligns with our observation that while the L3 tier may have higher raw edit counts ($\bar{e}_{total}$), its semantic fidelity ($\overline{GED}$) is superior because it avoids the logical shortcuts taken by the lower tiers. For the VG-DENSE sample (Right), the limitations of the L2 tier become more pronounced. Its triplet-based replacements, such as "beak (in,bush)" transitioning to "tree (near,road)", can be difficult to decode when relational density is high. The Structural subgraph, however, explicitly maps the logic required to transition from "Rainforest"→"Forest Road": it swaps the bird for a bear (an entity more intrinsic to the target environment) and physically introduces a road where greenery previously existed. The inclusion of extra edges in the L3 tier refines the conceptual hierarchy (e.g., increasing tree density), which explains why

Figure 4: Qualitative comparison of diverging counterfactuals retrieved by U-CECE expressivity tiers for CUB example (same CommonTern target class).

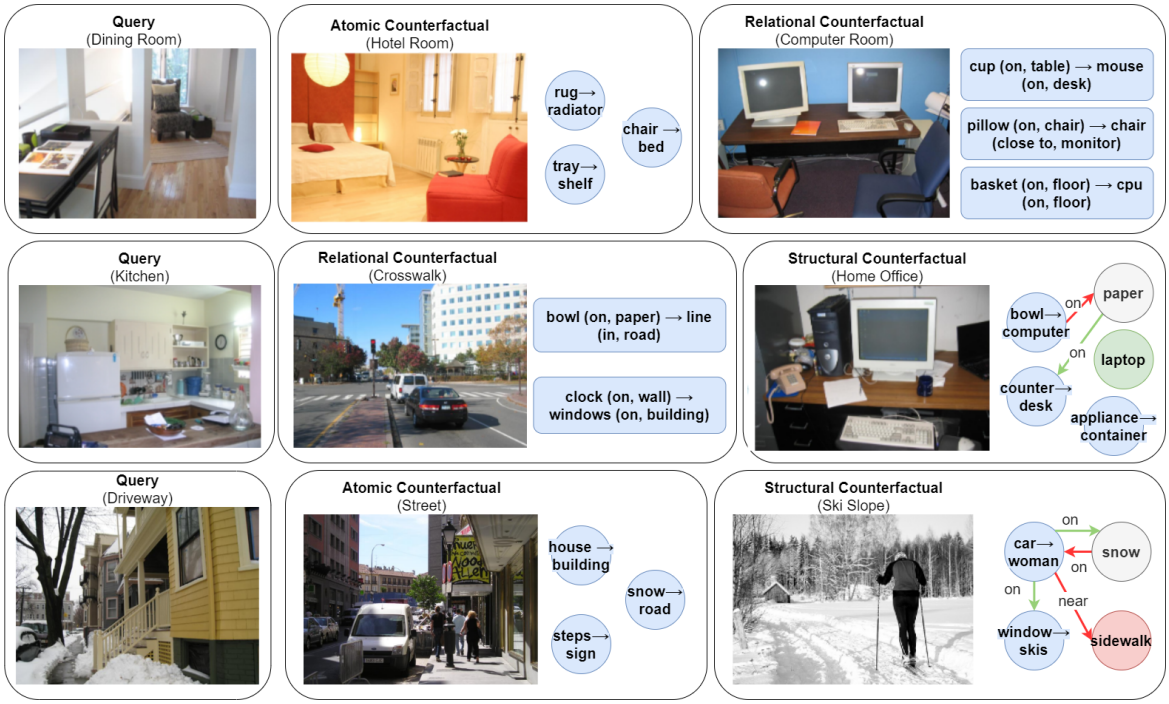


Figure 5: Qualitative comparison of diverging counterfactuals retrieved by U-CECE expressivity tiers for VG datasets (pairwise comparisons). A subset of edits is depicted for clarity purposes.

this tier achieved the highest $P@1$ in dense settings. Interestingly, the L1 tier remains surprisingly useful for single-concept transitions, reinforcing our quantitative finding that often simple concept-based retrieval is a low-cost alternative for broad conceptual shifts.

In cases where the different expressivity tiers retrieve distinct counterfactual images for the same target class (Figure 4), we gain further insight into the specific "reasoning" of each engine. While all three retrieved counterfactuals are semantically valid representatives of the CommonTern target class, the L3 engine identifies a candidate that is arguably dominant in terms of pose and visibility alignment with the original query. The structural explanation provides a highly straightforward narrative for the transition: to move from a ForstersTern to a CommonTern, one must eliminate the "solid" pattern and specifically reduce the prominence of "red" in the beak. While this core logic is partially reflected in the lower tiers—most notably in the L1 tier's identification of a "red-to-black" shift—the L3 tier provides the most refined and anatomically consistent roadmap. This reinforces our earlier quantitative findings: while the L1 tier is consistent at finding a correct bird through general attribute matching, the L3 tier excels at selecting the right bird that preserves the query's underlying graph-based context, even if it was not the optimal GT sample.

In Figure 5, we present pairwise comparisons on VG to highlight the specific logic of each expressivity mode. For clarity, we focus on a subset of representative replacements rather than displaying the complete edit graphs as in previous sections. Row 1 compares L1 vs. L3 mode. In sparse VG-RANDOM scenes like the Dining Room, where annotations often include isolated nodes (e.g. rug), the L1 tier suggests broad

conceptual swaps like chair → bed to retrieve a Hotel Room. These edits provide a general category shift but lack topological grounding. In contrast, the L2 tier utilizes triplet context to identify objects with similar scene interactions—such as swapping a basket on floor for a CPU on floor, thereby retrieving a Computer Room that feels more situationally grounded. Row 2 highlights the merit of structural expressivity in sparse scenarios. In the cluttered Kitchen query, the limitations of triplet-based retrieval become evident. Since the L2 tier over-indexes on specific triplets while ignoring the numerous isolated nodes typical of cluttered scenes, it suffers from semantic drift—matching a bowl on paper to a line in road and retrieving an irrelevant crosswalk. The L3 engine, however, accounts for both interconnected and isolated nodes. It successfully identifies a Home Office as a semantically closer counterfactual, logically swapping the counter for a desk to preserve the "indoor tabletop" nature of the scene. Finally, row 3 compares L1 vs. L3 mode in relationally denser scenes (like the depicted snow scene from VG-DENSE). While the L1 tier simply swaps house for building, it fails to capture how the depicted objects interact with the pervasive snow. The L3 engine re-purposes this interaction: it replaces the snow-covered car with a woman and her skis who is also interacting with the snow. This transition proves that the L3 tier does not just list presence; it refines the conceptual hierarchy to reflect the environmental truth, consistent with its superlative performance in our dense-relational quantitative benchmarks.

6 Testing Human Perception on U-CECE counterfactuals

While previous experiments and analysis address explanation minimality, retrieval quality, and engine efficiency, a systematic evaluation of the explanations themselves to the final recipients is required. This is particularly vital for the L3 variant which, as the most expressive tier by design, prioritizes feature interdependence. To justify the selection of the L3 variant for our human evaluation on the CUB dataset—despite lower-expressivity tiers occasionally showing higher $P@1$ scores—we anchor our study in the proven pedagogical efficacy of this specific engine: as demonstrated in Dimitriou et al. (2024b), the structural representations produced by SGCE are exceptionally robust for machine teaching tasks, with human participants successfully learning to categorize complex species based exclusively on the structural edit subgraphs, even in the complete absence of visual imagery during the teaching phase. By extending these findings, our current human perception experiments evaluate whether the neural-based retrieval engine serves as a semantically sensible alternative to the deterministic GED GT. This is a vital distinction: while GED provides a strict baseline, the GNN is trained on thousands of pairs to internalize a latent manifold that represents the conceptual essence of a "Bird". Consequently, these human trials can verify that even when the engine's top-ranked retrieval deviates from the exact GED, it remains within the same semantic neighborhood, preserving the logical integrity and explanatory power required for effective human-AI collaboration.

6.1 Design of Human Survey

We design three distinct survey formats to isolate different aspects of human judgment. Consistent with established visual XAI research (Vandenhende et al., 2022; Dervakos et al., 2023; Dimitriou et al., 2024b), we select the CUB dataset for bird species classification, to have humans evaluate the generated explanations. This selection serves as a control for prior knowledge: by employing non-expert annotators, we ensure that similarity assessments are based on the provided visual counterfactuals rather than pre-existing taxonomical expertise. To ensure participants evaluate high-level semantic concepts rather than low-level image artifacts, they are provided with the following standardization criteria:

> "*Assess whether the images are semantically equivalent by concentrating exclusively on the species and characteristics of depicted birds. Ignore any other elements, such as the surroundings, pose, or position of the birds.*"

The survey experiments are: i) **Select**: Participants are shown the source image alongside both the GT (GED) and the retrieved (GNN) counterfactuals. They are asked to identify which of the two images is more appropriate as a counterfactual, or if they are conceptually equal given the source; ii) **Sem-Eq-Triplets**: Given the same source-GED-GNN setup, participants are asked if the two counterfactuals are semantically equivalent in the context of the source; iii) **Sem-Eq-Pairs:** To assess baseline semantic consistency without

Auklet classes (Parakeet → Least) Blackbird classes (Rusty → Brewer)

Figure 6: Examples of visual form layouts showing two triplets of images used in the experiments. Each shows (Source, Image A (GT-GED), Image B (U-CECE-Structural w/ GCN)). As noted, in the **Sem-Eq-Pairs** experiment, the source image was not provided, nor was it revealed that images A, B were alternative counterfactuals for a specific query.

the influence of the "counterfactual task," participants evaluate the equivalence of the two counterfactual images in isolation. For brevity, we refer to the latter two together as **Sem-Eq-x** in our analysis.

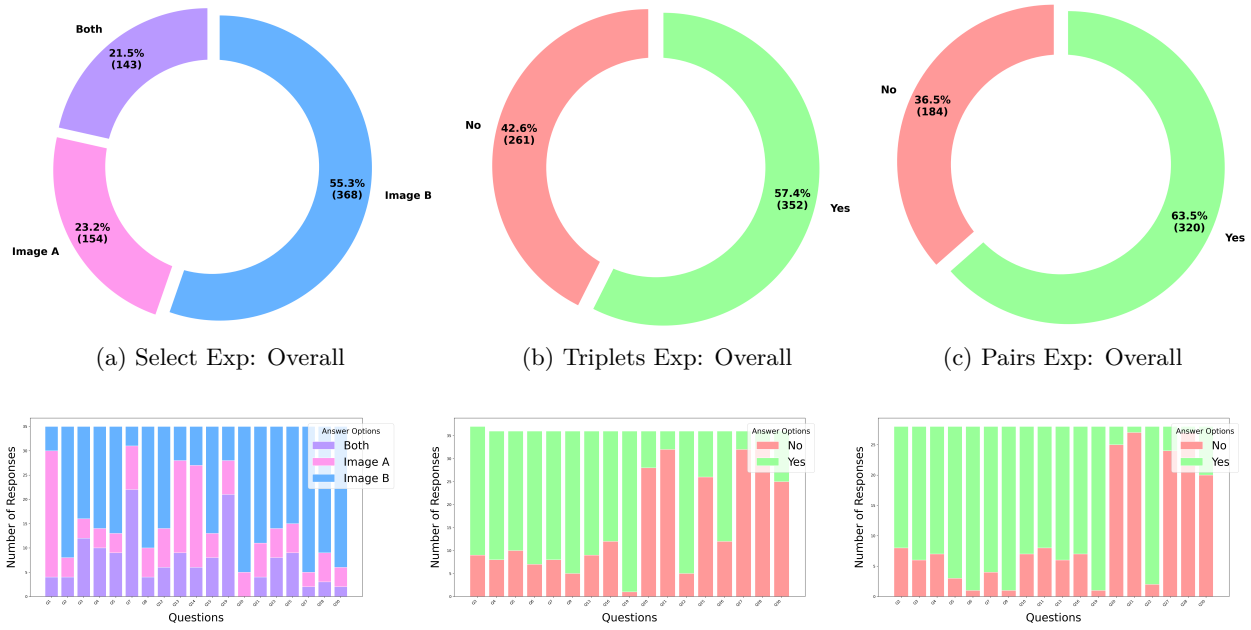
Each survey comprises 30 randomly selected instances from two specific counterfactual class transitions: RustyBlackbird → BrewerBlackbird, and Parakeet → LeastAuklet. These pairs are selected based on prior claims that they are widely confused by image classifiers during explainability tasks (Vandenhende et al., 2022). For each instance, we gather multiple independent human annotations (35 for **Select**, 36 for **Sem-Eq-Triplets**, and 28 for **Sem-Eq-Pairs**). The visual layout of these forms is illustrated in Figure 6, while comprehensive details regarding participant information and additional examples are available in App. C.

To transform subjective survey responses into mathematically grounded data, we move beyond simple majority voting. Specifically, we apply a **Chi-Square Goodness-of-Fit Test** (Liski, 2007) to every question. This allows us to segment our results into two distinct "buckets": i) Signal, i.e. questions where the null hypothesis (random guessing) is rejected ($p < 0.05$) representing cases where human intent is clear and consensus is statistically significant, and ii) Noise, i.e. questions where responses are statistically indistinguishable from random noise ($p \geq 0.05$), which identify edge cases where semantic differences are too subtle or subjective for universal agreement. The results of these surveys are presented below.

Assessing the semantic similarity of bird instances proved challenging, with approximately 1/3 of the instances exhibiting high inter-rater variance (noise). Notably, our Sem-Eq-x experiments demonstrate high consistency across annotators, with 25 out of 30 instances categorized the same way. This overlap hints that the inclusion of the source image does **not** fundamentally alter the perceived semantic relationship between counterfactuals. In contrast, agreement diminishes in the **Select** experiment, likely because the forced-choice format introduces subjective variance absent in binary tasks. Ultimately, 11 instances remain consistent across all setups, while only 4 are entirely excluded from the primary analysis due to noise.

The quantitative results of the human evaluation study are illustrated in Figure 7. To focus on mathematically validated human consensus, these visualizations present the results from the *signal dataset* ($p < 0.05$). Corresponding analysis detailing the high-entropy 'Noise' dataset are provided in Appendix C.3.

The results on signal data indicate that while the GNN(GCN)-retrieved counterfactuals are mathematically suboptimal in exact GED, human evaluators predominantly perceive them as semantically preferable over the GT GED counterfactuals. First of all, the **Select** experiment demonstrates that the GCN-based retrieval mechanism does not merely approximate the exact GED GT; it frequently surpasses it in human preference. The GCN-generated counterfactual are preferred in 55.3% of responses, compared to only 23.2% for the exact GED baseline (Figure 7a). Interestingly, the cognitive complexity introduced by comparing candidates against a third source image actually appears to favor GCN-retrieved instances. While this does not invalidate exact GED as a rigorous, mathematically sound evaluation metric for structural minimality, it highlights a critical distinction in human-centric explainability: as the GCN projects structural topology and rich GloVe features into a continuous latent space, it performs a form of "semantic smoothing." This enables the neural network to construct holistic instance representations that blend semantics and structure, capturing



(a) Select Exp: Overall (b) Triplets Exp: Overall (c) Pairs Exp: Overall

(d) Select Exp: Per Question (e) Triplets Exp: Per Question (f) Pairs Exp: Per Question

Figure 7: **Human evaluation results (signal dataset).** The top row displays the aggregate consensus for each experiment, while the bottom row details the distribution for each individual question.

nuanced visual patterns that exceed the capacity of rigid, discrete graph edits. Ultimately, this proves that for explanations requiring perceived semantic similarity, continuous neural representations are inherently better aligned with human visual intuition. Furthermore, in the **Sem-Eq-Triplets** setup, where the source image is present as a reference, 57.4% of respondents affirm GCN equivalence (Figure 7b). Notably, in the **Sem-Eq-Pairs** experiment, which removes the source image, consensus on GCN equivalence increases to 63.5% (Figure 7c). This suggests that the cognitive load of triangulating visual similarities across three distinct images reinforces perceptual ambiguity. Instead, when evaluated in isolation, the retrieved and GT images are more readily recognized as sharing the same conceptual essence, demonstrating that our retrieval results are semantically robust independent of the source context. Crucially, these findings validate the L3 (Structural) expressivity of U-CECE. By confirming visual equivalence, human subjects implicitly verify that the GCN-based approximation successfully preserves the complex relational topology required for high-expressivity explanations, effectively mirroring the NP-hard exact calculation. Finally, an analysis of the per-question distributions (Figures 7e, 7f) reveals a concentrated cluster of negative responses in the latter trials, which predominantly featured Blackbird classes. We hypothesize that the uniform, prominent black feathering of these instances made semantic assessment uniquely challenging, forcing non-expert participants to rely on irrelevant characteristics, such as absolute size or pose, to artificially differentiate the images.

To complement our behavioral analysis, we contextualize the human survey (Sem-Eq-x) results by comparing categorical responses to the distribution of pre-computed GEDs for both GT and GCN-retrieved counterfactuals (Figure 8). We aggregate responses from the noise dataset, where no statistical consensus is reached, as a baseline for semantic ambiguity. This fine-grained analysis of the human experiments maps subjective perception to a rigid structural metric, providing a clear picture of what semantic preference entails for non-expert observers. As expected, the GCN-retrieved results are mathematically suboptimal, yielding larger mean GED compared to the GT across all categories. However, a clear pattern emerges within the GCN-retrieved group: instances where humans reached a "Yes" consensus consistently exhibit lower average GEDs than those labeled "No". This suggests that even when the GCN fails to find the absolute mathematical minimum, its successful approximations (as perceived by humans) still reside in a tighter structural neighborhood than its failures.

The relationship between structural distance and human judgment shifts slightly depending on the experimental context. In the **Sem-Eq-Triplets** experiment, the noise category (represented by the grey marker with the highest inter-rater variance) sits squarely between the "Yes" and "No" signals, highlighting a zone where the structural distance from the source seems ambiguous. In contrast, in the **Sem-Eq-Pairs** experiment, the noise category is pushed further to the right, implying that without the source image as a reference, humans become more sensitive to larger structural discrepancies, requiring a higher degree of mathematical similarity to affirm semantic equivalence or not.

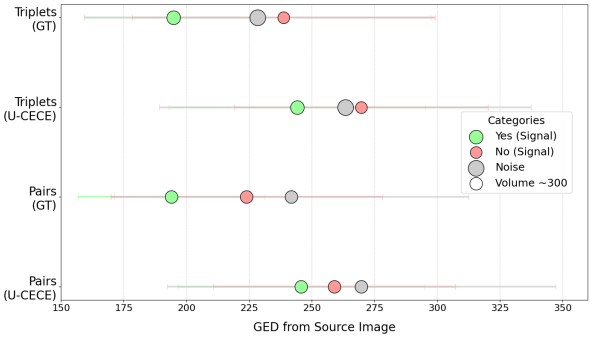


Figure 8: Distribution of GED across the Sem-Eq-x human perception surveys. Point markers represent the mean GED from the source image for GT and GCN-retrieved counterfactuals, categorized by participant consensus: *Yes* (Signal), *No* (Signal), and *Noise*. Error bars denote standard deviation. The area of each marker is proportional to the total volume of user annotations in that specific category.

It is crucial to note that while the GCN acts as a "bird expert" capturing fine-grained structural cues, the participants (non-experts) likely utilized a coarser semantic filter. The relatively tight standard deviation (horizontal bars) for "Yes" responses in the GT categories suggests that when structural minimality is high, even non-experts provide highly consistent semantic labels, whereas the GNN's increased variance in the "No" and "Noise" regions reflects the difficulty of maintaining semantic coherence as structural distance grows.

6.1.1 Human-LVLM Perception Alignment

Foundational Models such as Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in visual reasoning tasks, even acting as judges prompted for evaluating vision-language settings (Chen et al., 2024). In our case, we consider LVLMs to be an alternative to human evaluators, exploring their potential to obtain additional insights regarding our retrieved counterfactuals. To this end, we repeat the questions given to humans as prompts, explicitly formulated to ensure proper LVLM responses. We experiment with two cases: first, we perform a single-step prompting process, directly equivalent to the evaluation that humans performed. In the second case, we perform the reasoning-enhanced analyze-then-judge procedure proposed by Chen et al. (2024), which first assists the LVLM into understanding the images under consideration, and then perform any querying regarding them as the next step. Since LVLMs may hallucinate, we repeat the exact same evaluation experiment 10 times, simulating each LVLM session as a separate annotator, and keep the most frequent value for each answer choice. However, we observe that the used LVLM perfectly replicates its answers across runs, indicating confidence about its answers.

Experimental setup. The model used to perform the questionnaire judgment is ChatGPT 5.4 thinking, using the following default zero-shot prompt:

Annotation Prompt (single-step)

You are going to evaluate a questionnaire regarding counterfactual explanations on images. Provide your annotations without relying on any prior knowledge about bird species, and simulate a reasoning process similar to that of an engineering student. Focus on salient semantic characteristics of the bird instances provided, which are likely to represent discriminative features of the bird class, and ignore superficial factors such as lighting, viewing angle, or bird positioning. Return your responses in the form of a dictionary: `{question number: label}`.

For the analyze-then-judge experiment, the prompt is the following:

Annotation Prompt (two-step)

Now, I want you to perform re-annotation on the questionnaire following a different reasoning process. You have first to analyze the image visually and as a second step, you have to decide the label upon your previous analysis. Once again, you are going to provide your annotations without reliance to any prior knowledge about bird species, while you have to simulate your way of thinking similar to an engineering student. You should rely on salient semantic characteristics of the bird instances provided (which most likely would represent discriminative features of the bird class) and ignore other superficial features such as lighting, angle, bird positioning etc. Return your responses in a dictionary: `{question number: label}`.

Results of LVLM evaluation are presented in Figure 9, as an analogy to human evaluation. In all "Yes"/"No" experiments, the significantly higher "Yes" percentage indicates that the retrieved counterfactual is in fact semantically equivalent to the GT one in most cases. By also observing the leftmost column, the most frequent answer "B" reveals that selecting the GCN-retrieved instance is more preferable over the GT, reaching a consensus with human preferences (Figure 7a) and validating human-LVLM conceptual alignment.

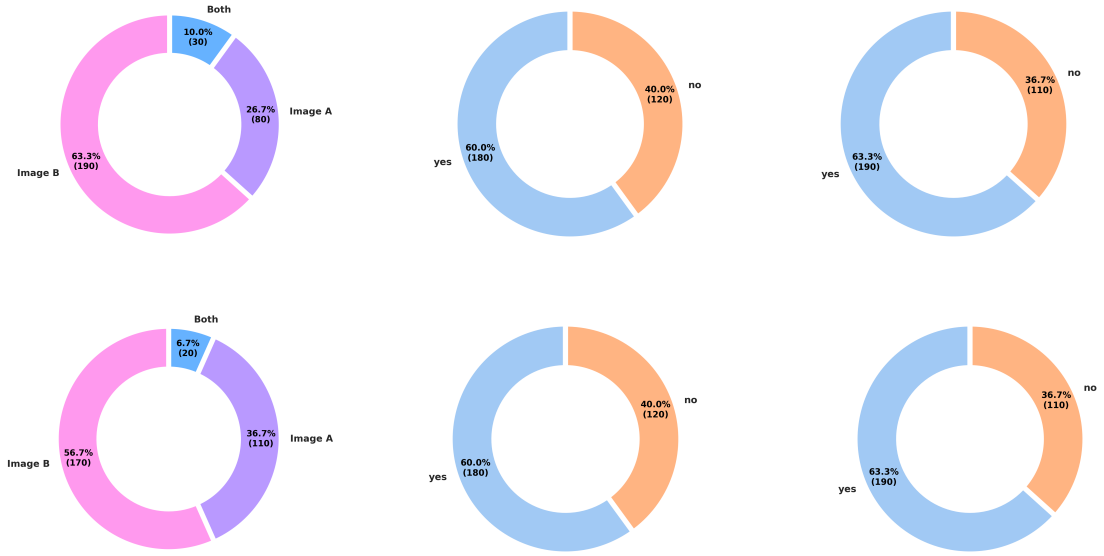
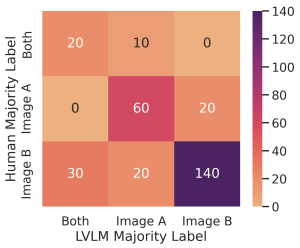


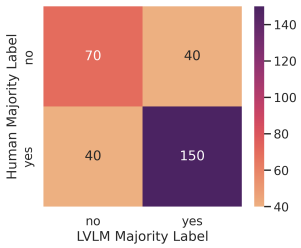
Figure 9: Results of LVLM-as-Judge. Top row for single step evaluation, second row for the analyze-then-judge procedure. The leftmost column corresponds to the **Select** experiment, the middle column represents the **Sem-Eq-Triplets** experiment, and the rightmost column corresponds to the **Sem-Eq-Pairs** experiment, all replicated for LVLM annotators.

Human-LVLM agreement is presented in Figure 10. Across settings, human–LVLM agreement is consistently above chance yet moderate, with performance varying by format. Agreement is highest in the **Select** experiment, particularly with reasoning (76.67% exact agreement, κ=0.5954), moderate in **Sem-Eq-Pairs** (73.33%, κ=0.4258), and lowest in **Sem-Eq-Triplets** (66.67%, κ=0.3056). Analyze-then-judge reasoning improves agreement only in the **Select** condition, suggesting that its benefit depends on task structure.

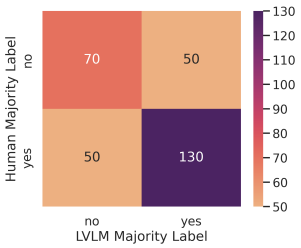
We attribute the observed discrepancies in LVLM responses to the varying cognitive demands and ambiguity of each task. The **Sem-Eq-Pairs** setting restricts the model to a binary contrast; while this simplicity facilitates higher raw agreement, the moderate κ score reveals this alignment is largely a byproduct of the limited label distribution. By contrast, the **Sem-Eq-Triplets** setting expands the search space, requiring fine-grained discrimination that increases ambiguity and consequently lowers human–LVLM agreement. The **Select** setting yields the highest overall agreement and is the only scenario that materially benefits from explicit reasoning. Rather than forcing symmetric comparisons, it tasks the model with isolating the most



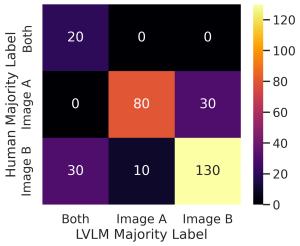
Exact agreement: 0.7333 (73.33%)
Cohen's kappa: 0.5285



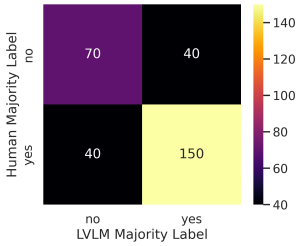
Exact agreement: 0.7333 (73.33%)
Cohen's kappa: 0.4258



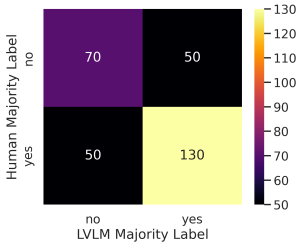
Exact agreement: 0.6667 (66.67%)
Cohen's kappa: 0.3056



Exact agreement: 0.7667 (76.67%)
Cohen's kappa: 0.5954



Exact agreement: 0.7333 (73.33%)
Cohen's kappa: 0.4258



Exact agreement: 0.6667 (66.67%)
Cohen's kappa: 0.3056

Figure 10: Human vs LVLM responses agreement confusion matrices directly corresponding to Fig. 9
.

prominent cue. This indicates that reasoning only helps when prioritizing distinct features, offering no advantage in environments that are overly constrained (Triplets) or highly ambiguous (Pairs). Ultimately, LVLMs excel at feature prioritization but struggle with relational integration—aligning closely with human judgment when isolating key discriminators, but faltering when coordinating multiple competing alternatives.

7 Limitations

U-CECE is limited by the quality of the conceptual abstraction stage: errors in concept extraction, relation parsing, or graph construction can propagate to retrieval and explanation quality. Our experiments mitigate this by focusing on richly annotated datasets such as CUB and VG, but this likely overestimates robustness relative to noisier real-world settings. The human study is restricted to a controlled bird domain and exhibits a noisy subset, while the LVLM-as-a-judge analysis shows only moderate human–LVLM agreement and is sensitive to task format. Finally, although the U-CECE framework establishes conceptuality in counterfactual explanations and improves semantic faithfulness and structural minimality, broader issues such as robustness to abstraction noise, fairness, and actionability are left for future work.

8 Future directions

The human-machine explanatory gap Conceptual edits have been proven crucial for unraveling current explainability issues in different experimental setups, such as counterfactual image generation (Spanos et al., 2025). In this case, conceptualization acts as an indispensable tool for diagnosing the *explanatory gap* between humans and visual classifiers: when explanations lack conceptual grounding, attempts on explaining a black-box classifier in human-explainable terms becomes futile, since the semantic level perceived by humans widely differs from a neural semantic level. Thus, the inability of a conceptual-driven image generation framework to accurately design salient, discriminative concepts between image classes stems not from the generator's capacity, but from the inherently misaligned signal that the classifier provides to the generator in the first place. In essence, the more a classifier's semantics abstract away from human conceptualization, the more misaligned the classifier's signal guiding the generating process will be, leading to limited interpretable

substance in the finally derived explanations. The work of Spanos et al. (2025) confirms the claims presented in the current paper, setting new core desiderata in counterfactual image generation literature.

Robustness assessments of conceptual counterfactuals An important future avenue is the stress-test on the fragility of conceptual counterfactuals, attributing and quantifying sources of brittleness in the image itself, as well as during concept extraction, relational parsing, graph construction, and retrieval stages. Specifically, future experimentation can focus on concept 'survival' across perturbation intensity and type and the discovery of bifurcation points where small input or concept changes lead to qualitatively different counterfactual explanations. Related findings will establish such analysis as a measure of ambiguity margins in derived explanations, while also exposing concept importance in human terms. At the same time, accompanying curation strategies will enhance the meaningfulness, trust and causal validity of conceptual counterfactual explanations.

Conceptual Uncertainty Estimation and Mitigation Ambiguity analysis inspires the exploration of a wider research direction: the uncertainty estimation of conceptual counterfactuals, particularly when semantic evidence is noisy, weakly specified, or inherently ambiguous. Targeted stress tests are able to reveal whether uncertainty stems from annotation noise, class overlap, or genuine aleatoric ambiguity, offering discrete, semantic-aware measures on the data themselves. Since such conceptual uncertainty can be well-measured and calibrated, any ambiguity of consequent explanation systems can be better isolated and gauged, while also mitigation endeavors can be framed more accurately, allowing for ambiguity-aware retrieval, conceptual enrichment and data-specific curation. Ultimately, this would position conceptual counterfactuals not only as explanatory artifacts, but also as tools for uncertainty estimation and reduction.

9 Conclusion

In this paper, we introduced U-CECE, a unified multi-resolution framework for conceptual counterfactual explanations that bridges atomic, relational, and structural concept representations within a single pipeline. More broadly, our findings reinforce the value of conceptual counterfactuals as a human-centered alternative to low-level counterfactual explanations: by operating on semantically meaningful concepts rather than pixel-level edits, they produce explanations that are better aligned with how humans perceive, compare, and reason about visual changes. Within this perspective, the three expressivity levels of U-CECE make explicit a central trade-off in conceptual explainability. The atomic level provides fast and lightweight explanations, the relational level captures simple interactions at moderate cost, and the structural level delivers the most faithful representation of complex scenes by preserving full relational topology, albeit with higher computational demands. Our experiments on CUB and Visual Genome show that no single level is universally optimal; instead, explanatory structure should be selected according to the relational complexity of the data and the practical constraints of the application. At the highest expressivity tier, this trade-off is further refined through the complementary transductive and inductive modes: the former offers high-precision retrieval when sparse data and strict structural minimality are paramount, while the latter supports scalable and flexible deployment when data are abundant and computational constraints favor efficient generalization. Human evaluation further demonstrates that the retrieved structural counterfactuals are not only semantically robust, but often preferred over exact GED-based references, highlighting an important distinction between mathematical optimality and perceived explanatory quality. While GED is rightfully established as a foundational metric for the evaluation and optimization of conceptual counterfactuals, this divergence reveals its limitations. It underscores that human-in-the-loop validation remains indispensable, as the true suitability and cognitive resonance of an explanation is notoriously difficult to capture through purely quantitative metrics. The LVLM-as-a-judge results complement this picture by largely reproducing the same preference and equivalence trends, while also showing that human–LVLM alignment depends on task structure and is strongest when the task emphasizes salient semantic cue prioritization. Taken together, these findings position U-CECE not simply as a retrieval framework, but as a principled foundation for studying how semantic abstraction, relational structure, and computational efficiency jointly shape the quality of counterfactual explanations. Overall, U-CECE offers a practical, extensible, and human-aligned foundation for conceptual counterfactual explanation—one that supports flexible deployment today while opening a broader path toward richer, more faithful, and more cognitively grounded explainability in the future.

References

Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 66–88, 2022.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Yunsheng Bai, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. Unsupervised inductive graph-level representation learning via graph-graph proximity. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 1988–1994, 2019a. ISBN 9780999241141.

Yunsheng Bai, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. Unsupervised inductive graph-level representation learning via graph-graph proximity. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 1988–1994, 7 2019b. doi: 10.24963/ijcai.2019/275. URL https://doi.org/10.24963/ijcai.2019/275.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety - a review. *Transactions on Machine Learning Research*, Aug 2024. URL https://openreview.net/forum?id=ePUVetPKu6.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Nikolaos Chaidos, Angeliki Dimitriou, Maria Lymperaiou, and Giorgos Stamou. Scenir: visual semantic clarity through unsupervised scene graph retrieval. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25, 2025.

Chun-Hao Kingsley Chang, Elliot Creager, Anna Goldenberg, and David Kristjanson Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:52962991.

Ziyi Chang, George A. Koulieris, Hyung Jin Chang, and Hubert P.H. Shum. On the design fundamentals of diffusion models: A survey. *Pattern Recognition*, 169:111934, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111934. URL https://www.sciencedirect.com/science/article/pii/S0031320325005941.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Long Chen, Wenbo Ma, Shizhe Xiao, Shih-Fu Chang, and Qi Jin. Counterfactual scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15255–15264, 2021.

Qihao Cheng, Da Yan, Tianhao Wu, Zhongyi Huang, and Qin Zhang. Computing approximate graph edit distance via optimal transport. *Proceedings of the ACM on Management of Data*, 3(1):1–26, 2025.

Kamil Deja, Tomasz Trzciński, and Jakub M. Tomczak. Learning data representations with joint diffusion models. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pp. 543–559, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43415-0.

Edmund Dervakos, Konstantinos Thomas, Giorgos Filandrianos, and Giorgos Stamou. Choose your data wisely: A framework for semantic counterfactuals. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pp. 382–390, 2023. doi: 10.24963/ijcai.2023/43. URL https://doi.org/10.24963/ijcai.2023/43.

Angeliki Dimitriou, Nikolaos Chaidos, Maria Lymperaiou, and Giorgos Stamou. Graph edits for counterfactual explanations: A comparative study. In *Explainable Artificial Intelligence*, pp. 100–112. Springer Nature Switzerland, 2024a. ISBN 978-3-031-63797-1.

Angeliki Dimitriou, Maria Lymperaiou, Georgios Filandrianos, Konstantinos Thomas, and Giorgos Stamou. Structure your data: Towards semantic graph counterfactuals. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10897–10926. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/dimitriou24a.html.

Hao Dong, Kaihua Tang, Jiebo Luo, Chuan-Sheng Xu, and Zhongfei Li. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 569–577, 2022.

Stefan Fankhauser, Kaspar Riesen, and Horst Bunke. Speeding up graph edit distance computation through fast bipartite matching. In *International Workshop on Graph-Based Representations in Pattern Recognition*, pp. 102–111. Springer, 2011.

Giorgos Filandrianos, Konstantinos Thomas, Edmund Dervakos, and Giorgos Stamou. Conceptual edits as counterfactual explanations. In *AAAI Spring Symposium: MAKE*, 2022.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. Dreamdojo: A generalist robot world model from large-scale human videos, 2026. URL https://arxiv.org/abs/2602.06949.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/goyal19a.html.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, September 2024. ISSN 1573-756X. doi: 10.1007/s10618-022-00831-6. URL https://doi.org/10.1007/s10618-022-00831-6.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=92oBV5HAGl.

Ramin Hasibi and Tom Michoel. A graph feature auto-encoder for the prediction of unobserved node features on biological networks. *BMC Bioinformatics*, 22, 2021. URL https://api.semanticscholar.org/CorpusID:240006620.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. *Computer Vision and Image Understanding*, 249:104207, 2024. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2024.104207. URL https://www.sciencedirect.com/science/article/pii/S1077314224002881.

Roy Jonker and Ton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. In *DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*, pp. 622–622. Springer, 1988.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2673–2682, 2018.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders, 2016. URL https://arxiv.org/abs/1611.07308.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *Proceedings of the 27th European Conference on Artificial Intelligence*, 2024.

Xuwang Kong, Tianshui Chen, Xu Xu, Wenyu Chen, Liang Peng, and Xiangnan He. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19494–19503, 2022.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision (IJCV)*, volume 123, pp. 32–73, 2017.

Evgeny Krivosheev, Mattia Atzeni, Katsiaryna Mirylenka, Paolo Scotton, and Fabio Casati. Siamese graph neural networks for data integration, 2020. URL https://arxiv.org/abs/2001.06543.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2 (1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning*, pp. 3835–3845. PMLR, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Erkki P. Liski. An introduction to categorical data analysis, 2nd edition by alan agresti. *International Statistical Review*, 75(3):414–414, 2007. doi: https://doi.org/10.1111/j.1751-5823.2007.00030_6.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2007.00030_6.x.

Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, 2019. URL https://api.semanticscholar.org/CorpusID:195833361.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision (ECCV)*, pp. 852–869, 2016.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthikeyan Shanmugam, and Chun-Chen Tu. Generating contrastive explanations with monotonic attribute functions. *ArXiv*, abs/1905.12698, 2019. URL https://api.semanticscholar.org/CorpusID:170078674.

Franz Motzkus, Christian Hellert, and Ute Schmid. Cola-dce – concept-guided latent diffusion counterfactual explanations, 2024. URL https://arxiv.org/abs/2406.01649.

Michel Neuhaus and Horst Bunke. *Bridging the Gap Between Graph Edit Distance and Kernel Machines*. World Scientific Publishing Co., Inc., USA, 2007. ISBN 9789812708175.

Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Adversarially regularized graph autoencoder for graph embedding, 2018.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *ACM Comput. Surv.*, November 2025. ISSN 0360-0300. doi: 10.1145/3774643. URL https://doi.org/10.1145/3774643.

Rishabh Ranjan, Siddharth Grover, Sourav Medya, Venkatesan Chakaravarthy, Yogish Sabharwal, and Sayan Ranu. Greed: A neural framework for learning graph distance functions. In *Advances in Neural Information Processing Systems*, 2022.

Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision computing*, 27(7):950–959, 2009.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 647–668. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/sanchez22a.html.

Alberto Sanfeliu and King-Sun Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (3):353–362, 1983.

Simon Schrodi, Karim Farid, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations. In *Pattern Recognition*, pp. 295–311. Springer Nature Switzerland, 2025. ISBN 978-3-031-85181-0.

Francesc Serratosa. Fast computation of bipartite graph matching. *Pattern Recognition Letters*, 45:244–250, 2014.

Bartlomiej Sobieski, Jakub Grzywaczewski, Bartłomiej Sadlej, Matthew Tivnan, and Przemyslaw Biecek. Rethinking visual counterfactual explanations through region constraint. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nikolaos Spanos, Maria Lymperaiou, Giorgos Filandrianos, Konstantinos Thomas, Athanasios Voulodimos, and Giorgos Stamou. V-CECE: Visual counterfactual explanations via conceptual edits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NKPud1Jnyt.

Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9: 11974–12001, 2021.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3716–3725, 2020.

Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. Making heads or tails: Towards semantically consistent visual counterfactuals. In *European Conference on Computer Vision*, pp. 261–279. Springer, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *6th International Conference on Learning Representations*, 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Yulu Wang, Pengwen Dai, Xiaojun Jia, Zhitao Zeng, Rui Li, and Xiaochun Cao. Hi-sigir: Hierachical semantic-guided image-to-image retrieval via scene graph. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 6400–6409, 2023.

Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVI*, pp. 338–357. Springer-Verlag, 2024. ISBN 978-3-031-73015-3. doi: 10.1007/978-3-031-73016-0_20. URL https://doi.org/10.1007/978-3-031-73016-0_20.

Christopher Williams. On a connection between kernel pca and metric multidimensional scaling. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/d4b2aeb2453bdadaa45cbe9882ffefcf-Paper.pdf.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Jiaxin Yang, Zhiwu Lu, Jinhui Chen, and Tao Zhang. Causal attention for unbiased scene graph generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10718–10726, 2021.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57:1 – 34, 2025. URL https://api.semanticscholar.org/CorpusID:275631365.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14393–14402, 2020.

Hanwang Zhao, Can Cui, Yandong Liu, and Mingkui Tan. Tsp: Transformer-based scene graph parsing with instance-level constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5622–5631, 2021.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17733–17750, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.895. URL `https://aclanthology.org/2025.emnlp-main.895/`.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A Experimental Details

For the supervised GNN models, we utilize a single-layer architecture with a dimension of 2048, optimized using Adam (no weight decay or dropout) with a batch size of 32 over 50 epochs. The learning rate is set to 0.04 for GCN and 0.02 for GAT and GIN; GAT specifically employs 8 attention heads, while the ϵ parameter for GIN is non-learnable. In the unsupervised setting, training is divided into a 30-epoch pretraining phase and a 20-epoch finetuning phase. For the VG dataset, a learning rate of 10^{-4} is maintained for both stages. For the CUB dataset, we utilize a learning rate of 10^{-3} during pretraining and 10^{-4} during finetuning.The number of training pairs, denoted as p, is set following the heuristic $p \approx N/2$ from the SGCE paper, resulting in 70K and 50K pairs for datasets with 500 and 422 graphs, respectively. Finally, all experiments were executed on a single NVIDIA T4 GPU, ensuring a consistent and accessible computational baseline.

B Complementary Experimental Findings on Expressivity Modes

Table 4: Retrieval Metrics ($k = 2, 4$) across Expressivity Tiers. P_{bin} denotes binary precision.

Dataset	**Level**	**P@2** ↑	**P@4** ↑	**nDCG@2** ↑	**nDCG@4** ↑	P_{bin}**@2** ↑	P_{bin}**@4** ↑
CUB	Atomic	**0.241**	**0.330**	**0.276**	**0.357**	**0.325**	**0.506**
	Relational	0.197	0.294	0.248	0.332	0.238	0.388
	Structural	0.200	0.306	0.239	0.323	0.256	0.450
VG DENSE	Atomic	0.278	0.306	0.297	0.375	0.338	0.418
	Relational	0.208	0.261	0.262	0.344	0.242	0.318
	Structural	**0.292**	**0.358**	**0.333**	**0.407**	**0.370**	**0.492**
VG RANDOM	Atomic	**0.256**	**0.302**	**0.293**	**0.372**	0.296	0.388
	Relational	0.099	0.143	0.201	0.290	0.110	0.182
	Structural	0.246	0.295	0.290	0.369	**0.300**	**0.424**

As shown in Table 4, retrieval metrics increase significantly at higher k values. Notably, the binary precision for locating the exact GT counterfactual within the top 2 or 4 samples approaches 50%, doubling the maximum $P@1$ ($\sim 25\%$). Trends remain consistent across all datasets and methods.

C Human Survey

C.1 Details on Participants

Prior to participation, informed consent was obtained from individuals to utilize their anonymized responses for research purposes. To preserve privacy, fine-grained demographic data was not recorded. However, the survey was distributed to a group of undergraduate and graduate students - age range 23 - 35, of both sexes. All participants share an educational background in software engineering. Crucially, none of the participants

possessed prior domain expertise in ornithology. This lack of specialized knowledge was intentional; it ensures that their ability to evaluate the fine-grained bird counterfactuals relied entirely on the structural and visual explanations provided by the framework, rather than any pre-existing subject familiarity.

C.2 Form Design and Instructions

The human perception study was conducted using the Google Forms platform. This choice of medium allowed participants to leverage standard browser utilities, such as zooming in on high-resolution images, to carefully inspect the fine-grained visual characteristics of the birds before making their assessments. The study was divided into three distinct experimental setups, aligning with the analysis on the main paper: **Select**, **Sem-Eq-Triplets**, and **Sem-Eq-Pairs**. Each form began with a specific set of instructions to align the participants' focus purely on semantic equivalence (species and anatomical traits) rather than environmental noise.

"Triplets-Select" Initial Instruction: *"The first image is the source, while the second and the third are proposed counterfactual images of the source. Focus on the species and the characteristics of the counterfactuals (Image A and Image B) and ignore other image details, bird pose or position. Which of the two counterfactuals, Image A or Image B, is the most appropriate to the source? If both are equally good counterfactuals, answer 'Both'."*

For each instance, participants answered the question: **"Which of the two counterfactuals, Image A or Image B, is the most appropriate to the source?"** using a multiple-choice format (Image A, Image B, Both)

"Sem-Ed-Triplets" Initial Instruction: *"The first image is the source, while the second and the third are proposed counterfactual images of the source. Are the proposed counterfactuals (second and third image) semantically equivalent? Focus on whether the counterfactuals are of the same species and have similar characteristics."*

Participants were asked to answer **"Are the proposed counterfactuals (Image A and Image B) semantically equivalent?"** with a simple *Yes* or *No*

"Sem-Eq-Pairs" Initial Instruction: *"Given two images, assess whether they are semantically equivalent by concentrating exclusively on the species and characteristics of the birds depicted. Ignore any other elements, such as the surroundings, pose, or position of the birds."*

Participants answered the question: **"Are the two images semantically equivalent?"** (*Yes* / *No*).

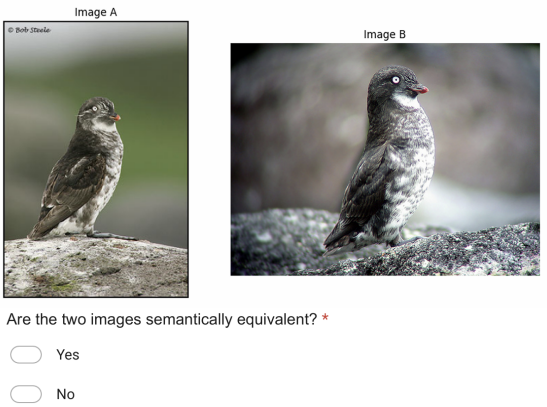


Figure 11: An example of the **Pairs** experimental layout. The Source Image is removed to test unanchored semantic equivalence.

C.3 Noise Results

Analysis of the human evaluation results for the Noise dataset across all experimental tiers reveals a lack of consensus, with response distributions approximating a random choice baseline. In the Select experiment, participants displayed near-equal preference between Image A (34.0%), Image B (32.7%), and the "Both" option (33.2%), mirroring the 1/3 probability of a random selector. This trend persists in the binary tiers, where "Yes" votes reached only 50.6% for Pairs and 52.6% for Triplets, effectively hovering at the 50% chance threshold. These data validate our categorization of these instances as semantically ambiguous "noise," confirming that for these specific samples, the neural-based U-CECE-Structural retrievals and the exact GED ground truth are perceived as equally plausible (or equally suboptimal) from a human perspective.