What Is Mentioned vs. How It Is Related: A Comprehensive Diagnosis of Relation Blindness in Dense Text Embeddings

Anonymous ACL submission

Abstract

Dense text embeddings are central to modern NLP systems, yet they often capture what entities are mentioned more reliably than how these entities are related. We refer to this phenomenon as *relation blindness*: models preserve high representational similarity despite meaning-critical changes in relational predicates. Existing evaluations provide limited evidence on this issue. To better diagnose it, we construct a controlled benchmark that isolates relation-level sensitivity from lexical overlap and surface-form artifacts. Evaluating eight widely used embedding models, we find that relation blindness is a systematic weakness, leading to substantial performance drops on tasks that require relational understanding. Mechanistic interpretability analyses suggest that relational semantics are integrated late and are often overshadowed by entity-level lexical features. Finally, relation-aware fine-tuning improves performance on relation-intensive benchmarks while preserving competitive performance on the broader MTEB suite. Our results identify relational semantics as a critical blind spot in current dense text embeddings.

1 Introduction

Accurately representing the semantics of textual expressions is a cornerstone of robust natural language understanding, directly underpinning the performance of downstream applications such as passage retrieval (Lewis et al., 2020; Karpukhin et al., 2020), semantic clustering (Reimers and Gurevych, 2019), and cross-lingual alignment (Feng et al., 2022). In many real-world applications, semantic distinctions are determined not only by the entities mentioned, but also by the relations asserted among them. For example, "fluid in the lungs" and "fluid around the lungs" involve the same core entities, yet describe distinct clinical conditions that can require different interventions. In such cases, failing to capture relational semantics can cause models to preserve surface-level similarity while missing meaning-critical distinctions.

This requirement is strongly grounded in the Principle of Compositionality, which asserts that the semantic value of a complex expression is a deterministic function of its constituent meanings and their syntactic structure (Heim and Kratzer, 1998; Portner and Partee, 2008; Pickel and Szabó, 2025; Pelletier, 2017). Under compositional semantics, the lexical semantics of an expression's components and their combination mode (i.e., syntactic rules) jointly and uniquely determine the global meaning. Accordingly, a model adhering to the Principle of Compositionality should discriminate nuanced variations in both constituent meanings and their syntactic arrangement.

As subtle changes at the surface level could substantially altering the underlying proposition, translating this requirement into dense vector spaces remains challenging. Prior evidence (McCoy et al., 2019; Sinha et al., 2021) shows that state-of-the-art models often rely on lexical heuristics, resembling "bag-of-words" matching. Consistent with this concern, we observe relation blindness in embedding models. As shown in Figure 1, although both noun substitutions and relational predicate alterations shift the underlying semantics in human annotation, embedding models maintain anomalously higher representational proximity for relational changes than for lexical ones.

Despite growing interest in evaluating the compositional capabilities of embedding models (Grefenstette and Sadrzadeh, 2011; Magomere et al., 2025; Nagar et al., 2025; Carvalho et al., 2025), existing evaluations still lack a controlled way to isolate relation-level sensitivity from lexical overlap and surface-form artifacts. Standard benchmarks such as STS (Cer et al., 2017), SICK (Marelli et al., 2014), and MTEB (Muennighoff et al., 2023) often conflate semantic similarity with lexical overlap, allowing models with limited struc-

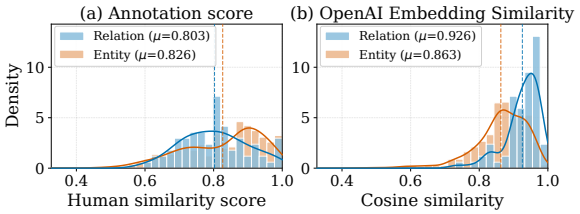

Figure 1: Comparison between human similarity labels and OpenAI embedding cosine similarities for pairs of subject–relation–object triples and their relation/entity perturbations.

tural awareness to achieve high scores by exploiting shallow surface matching. Recent perturbation-based studies (Nagar et al., 2025; Fodor et al., 2025; Chiang et al., 2023; Wang et al., 2024a) provide finer-grained tests, but mainly target lexical constituents such as nouns and verbs, leaving relational predicates underexplored.

To systematically isolate these failure modes and expose the true gap in compositional representation, we propose a controlled benchmark focusing on relation and entity changes. Specifically, we use the Visual Genome v1.4 dataset (Krishna et al., 2017), which contains over 100K images and 2.3M predominantly spatial Subject-Relation-Object (S-R-O) triples extracted from images. We generate minimally perturbed triples by replacing only one component at a time: the subject, relation, or object. Candidate perturbations are guided by the WordNet hierarchy. After filtering nonsensical perturbations and retaining only semantically valid changes, we obtain 69K relation-changed triples, 100K subject-changed triples, and 100K object-changed triples. Our benchmark is designed to isolate relational predicates from lexical constituents, enabling a direct comparison between model sensitivity to entity-level and relation-level semantic shifts under controlled surface conditions.

We conduct an extensive evaluation of eight popular embedding models and reveal a consistent relational blindness: these models are substantially less sensitive to relational predicate changes than to lexical constituent changes. At the same time, these models show a substantial performance drop on small Sick-R (Marelli et al., 2014) subsets that require relational understanding. To understand this phenomenon, we further use mechanistic interpretability analyses to trace how sentence representations evolve across layers. Our analysis shows that, models tend to emphasize subjects and objects in the attention flow, suggesting a stronger reliance on entity-level lexical features. Sentence-level meaning becomes more salient only in the final one or two layers, where the model begins to form more holistic semantic representations. This delayed integration may explain why dense embeddings often preserve high similarity even when the relation changes. Finally, we fine-tune embedding models using a mixture of our curated relation-focused dataset and public training datasets, substantially improving performance on relation-intensive benchmarks while preserving competitive performance on the MTEB suite.

2 Related Work

Compositionality is a fundamental assumption in modern semantic theories, positing that the meaning of a complex expression is determined by its syntactic structure and the meanings of its constituent parts (Heim and Kratzer, 1998; Portner and Partee, 2008; Pickel and Szabó, 2025; Pelletier, 2017). This principle provides a profound explanation for the productivity and systematicity of human linguistic competence. Guided by this principle, a substantial body of early work has sought to model compositional meaning in distributed semantic representations (Grefenstette and Sadrzadeh, 2011; Kartsaklis et al., 2013; Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010). For example, (Baroni and Zamparelli, 2010) introduced the lexical function model, in which adjectives are treated as functions mapping noun vectors to phrase vectors; Similarly, (Morzycki, 2016) systematically examines modification as a central case of compositional meaning.

In LLM era, models are typically trained in a self-supervised manner using next-token prediction objectives. Accordingly, evaluation has shifted toward examining whether model representations exhibit the desired compositional properties (Nagar et al., 2025; Carvalho et al., 2025; Fodor et al., 2025). For example, (Nagar et al., 2025) investigates how modern Transformer-based embedding models encode compositional meaning in word combinations, such as compound nouns (e.g., *bookstore*) and adjective–noun phrases (e.g., *red ball*); (Carvalho et al., 2025) propose a methodology for measuring compositional behavior in adjectival modification, focusing on adjective–noun compositions. Relational words carry limited standalone meaning and have received relatively little attention in existing research, yet they play a crucial

functional role in determining how meanings are composed. Our work is the first to evaluate whether embedding models capture this role.

Embedding models evaluation. The advent of word embeddings, ranging from static models such as Word2Vec (Mikolov et al., 2013) to contextualized models such as BERT (Devlin et al., 2019), Qwen3-Embedding Model (Zhang et al., 2025), marked a major step toward representing lexical meaning in dense vector spaces. Embedding models are commonly evaluated using general-purpose benchmarks such as MTEB (Muennighoff et al., 2023) and STS (Cer et al., 2017) for assessing broad representation quality and task-specific benchmarks (Caspari et al., 2024) such as retrieval-augmented generation (Lewis et al., 2020).

Another line of research has examined the robustness of representations to minimal input perturbations. A representative example is adversarial robustness, where small input perturbations are used to test whether model representations remain stable under attacks (Papernot et al., 2016; Ren et al., 2019). However, adversarial attacks often produce unnatural or semantically invalid inputs from a human perspective (Chiang and Lee, 2023; Morris et al., 2020). Inspired by counterfactual generation (Ribeiro et al., 2020; Wang et al., 2024d), a series of studies perturb sentences using diverse strategies, including surface-level modifications such as casing changes and typos, semantic-linguistic shifts such as negation and entity replacement, and stylistic adaptations such as LLM-based rewrites and dialectal variations (Magomere et al., 2025; Fodor et al., 2025; Chiang et al., 2023). Following this research line, our work provides the first systematic evaluation of whether embedding models are sensitive to relational words, revealing an intriguing phenomenon that we term relational myopia in embedding models.

3 Data and Methods

3.1 Problem Formulation

Grounded in formal semantics and first-order logic (Heim and Kratzer, 1998; Portner and Partee, 2008), the underlying meaning of a factual proposition derived from its syntactic constituents and their combinatorial rules. We formally define the semantic meaning $\mathcal{M}$ of an utterance $X = \{r, s, o\}$ as a functional application:

$$\mathcal{M}(X) := r(s, o) \tag{1}$$

where s and o are the entity arguments (subject and object), and $r \in \mathcal{R}$ is the relational predicate acting as a semantic operator that binds them (e.g., $in(\text{cat}, \text{room})$ or $after(\text{shower}, \text{workout})$).

Let $f : \mathcal{X} \to \mathbb{R}^d$ denote an embedding function that maps an utterance X to a dense vector. Under a compositional account of meaning, dense embeddings should be sensitive not only to which constituents are present, but also to how they are structurally related. To evaluate the model's compositional sensitivity, we introduce a structural perturbation operator $\mathcal{P}$. This operator selectively substitutes either the relational predicate ($r \to r'$) with its semantic or topological antonym, or one entity argument (s or o) with a different one, allowing us to isolate the specific locus of the model's representation failure. This yields a perturbed utterance X', introducing a semantic shift such that the new meaning is mutually exclusive to the original (i.e., $\mathcal{M}(X') \perp \mathcal{M}(X)$).

Under the assumption of compositionality, the embedding function f should intrinsically capture the divergence induced by the perturbation. It is expected to reflect the underlying logical contradiction as a large spatial separation in its vector space:

$$\mathcal{D}\Big(f(X_{\mathcal{T}}), f(X_{\mathcal{T}'})\Big) \geq \tau \tag{2}$$

where $\mathcal{D}$ is a distance metric and τ establishes the minimum margin required to represent semantic mutual exclusivity.

3.2 Data Collection

A benchmark for systematically measuring compositionality in embedding models should satisfy these requirements. First, it should support controlled perturbations over both entities and relations, enabling direct comparison between lexical-constituent changes and relational changes. Second, these perturbations should induce genuine semantic changes, rather than relying on bag-of-words overlap as a proxy for semantic similarity. Third, the benchmark should be sufficiently large to support comprehensive and statistically reliable evaluation.

Our data collection pipeline is shown in Figure 2. We use the Visual Genome v1.4 dataset (Krishna et al., 2017) as the seed dataset, which contains over 100K images and 2.3 million Subject-Relation-Object (S-R-O) triples extracted from images. These triples cover a broad range of everyday visual scenes and provide a natural source of

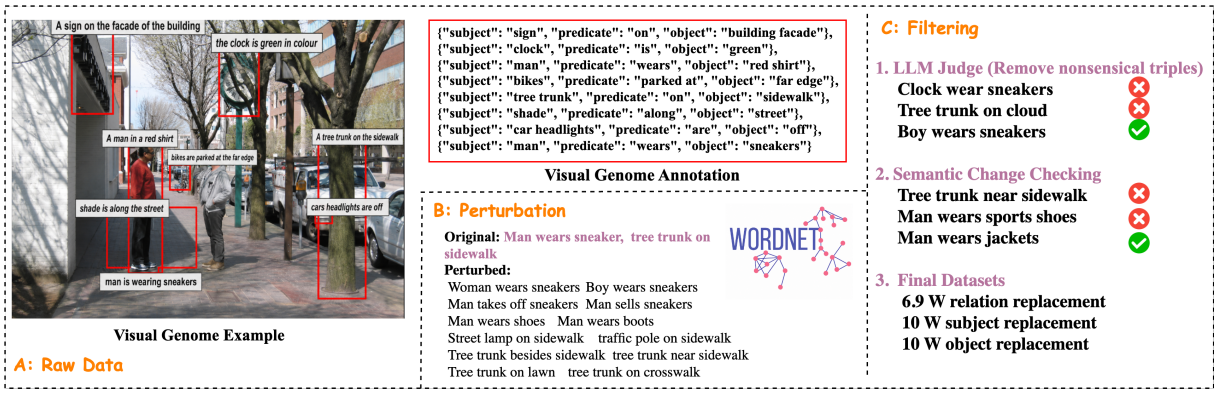

Figure 2: Overview of the data collection pipeline. Starting from Visual Genome annotations, we extract raw SRO triples, generate controlled perturbations over subjects, relations, and objects. Finally, we use an LLM judge removes nonsensical triples, and semantic checking filters variants that do not induce the intended semantic change.

relational expressions, with approximately 36K relation expressions. In addition, its S-R-O structure allows us to isolate changes to the subject, relation, or object while keeping the remaining components fixed. The entities are explicitly linked to WordNet synsets (Miller, 1995), facilitating controlled semantic replacement too. During preprocessing, we first remove duplicate triples and entries with missing fields. We also exclude triples whose subject or object cannot be mapped to a WordNet synset. This process yields 500K high-quality raw triples for subsequent perturbation and validation.

In the perturbation stage, we apply controlled replacement rules to the subject, object, and relation, producing three types of minimal replacements. For subject and object perturbations, we first map each subject/object to one of the 26 WordNet supersense categories. Candidate substitutions are restricted to the top 25% most frequent entities within the same WordNet supersense category to preserve semantic coherence. We also discard candidate substitutions with Wu-Palmer (WUP) similarity above 0.9 to avoid near-synonymous replacements. For relation perturbation, we construct a two-level hierarchical taxonomy with 8 top-level categories, including temporal, spatial, causal, and possession relations, and a set of fine-grained subcategories. Candidate relation substitutions are restricted to the same fine-grained subcategory, ensuring that the perturbed triples remain coherent. For each original triple, we generate up to 20 perturbed variants for each perturbation type. The complete and detailed processing rules are provided in the Appendix A, will be released online.

Our rule-based perturbations inevitably introduce some nonsensical triples, such as "tree trunk on cloud." We therefore adopt an LLM-as-a-Judge approach (Zheng et al., 2023) to filter these invalid cases. In particular, we prompt GPT-4.1-mini to judge the commonsense validity of each generated triple and provide a confidence score. The detailed prompt is provided in Figure 7 in the Appendix. We retain only triples that are judged to be valid with high confidence (> 0.8). To verify this approach, we conducted a rigorous manual annotation of 1,000 samples. Evaluated against these human annotations, the LLM judgments achieved a precision of 0.969, demonstrating that our filtering strategy reliably yields a high-quality and natural probing dataset.

Although the generated perturbations thus far are plausible, they do not guarantee substantive semantic shifts. To enforce strict semantic divergence, we restrict the relation perturbations exclusively to antonymous relations, yielding 69K relation pairs. For subject and object perturbations, we select the top 100K entity pairs that exhibit the lowest Wu-Palmer similarity.

3.3 Model Finetuning

To mitigate relation blindness, we fine-tune embedding models using our collected relation-sensitive data. Since training only on antonymous relation pairs will easily distort the representation space and lead to overfitting to relation-specific patterns, and this finally compromises the models' general semantic textual similarity capabilities. Therefore, we incorporate the SimCSE training set (Gao et al., 2021) and a modified, asymmetric version of MS MARCO (Bajaj et al., 2016), which help maintain general semantic matching ability and reduce degradation on non-STS downstream tasks.

Following previous work (Zhang et al., 2025; Reimers and Gurevych, 2019; Wang et al., 2024b), we adopt a contrastive objective to fine-tune the sentence encoder, pulling semantically equivalent pairs closer while pushing relation-based hard negatives and standard negatives apart. For our antonymous relation pairs, we introduce the following objective:

$$\mathcal{L}_{\text{rel}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(\frac{\text{sim}(a_i, n_i)}{\tau}\right)\right), \tag{3}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, τ is the temperature parameter, and N is the batch size. This objective penalizes high similarity scores between relation-based negative pairs, encouraging the encoder to separate them in the embedding space.

For the SimCSE and MS MARCO datasets, which are organized as anchor–positive–negative triples, we apply the standard InfoNCE loss (Oord et al., 2018), denoted as $\mathcal{L}_{\text{nli}}$ and $\mathcal{L}_{\text{ms}}$, respectively. The general form is,

$$\mathcal{L} = -\log\frac{\exp\left(\text{sim}(\mathbf{q}, \mathbf{k}_+)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(\mathbf{q}, \mathbf{k}_j)/\tau\right)}, \tag{4}$$

where $\mathbf{q}$ and $\mathbf{k}_+$ represent the embeddings of the query and the positive sample, respectively. $\mathbf{k}_j$ denotes the j-th candidate in a batch of size N.

We train the model with a joint objective that combines the relation-sensitive loss with the InfoNCE loss on SimCSE and MS MARCO datasets:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rel}} + \lambda_{\text{nli}}\mathcal{L}_{\text{nli}} + \lambda_{\text{ms}}\mathcal{L}_{\text{ms}}. \tag{5}$$

Here, λ_{nli} and λ_{ms} control the weights. Finally, we fine-tune the model using the AdamW optimizer by minimizing $\mathcal{L}_{\text{total}}$.

4 Benchmarking Results

4.1 Embedding Models

We evaluate a diverse set of open-source and commercial embedding models. The open-source models cover i) encoder-only architectures, such as Sentence-Transformer all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) (MiniLM), jina-embeddings-v3 (Sturua et al., 2024)(Jina-v3), BGE-M3 (Chen et al., 2024), e5-large-v2 (Wang et al., 2024c) (E5-Large) and EmbeddingGemma-300M (Vera et al., 2025) (EGemma); ii) decoder-only architectures Qwen3-Embedding-4B/8B (Zhang et al., 2025); iii) commercial models, such as OpenAI Embedding (text-embedding-3-large) (Neelakantan et al., 2022) (OpenAI-Emb). The model description is listed in Table 1.

Table 1: Model Description.

Name	#Dim	Type	Year
all-MiniLM-L6-v2	384	Mean	2021
BGE-M3	1024	CLS	2024
e5-large-v2	1024	Mean	2022
jina embedding v3	1024	Mean	2024
Qwen3-Embedding-4B	2560	EOS	2025
Qwen3-Embedding-8B	4096	EOS	2025
OpenAI text-embedding-3 (S)	1536	LAST	2024
OpenAI text-embedding-3 (L)	3072	LAST	2024
embeddinggemma-300M	768	Mean	2025

4.2 Similarity Distributions

Following previous studies (Chiang et al., 2023; Nagar et al., 2025), we use Kernel Density Estimation (KDE) plots to visualize the distribution of cosine similarities between the original triples and their perturbed counterparts, where the subject, relation, and object are perturbed separately.

As shown in Figure 3, the vertical lines indicate the average cosine similarity between the original triples and their perturbed counterparts, while the spread of each curve reflects the variability of similarity scores. Despite these semantic changes, all embedding models assign substantially higher similarities to relation-perturbed triples than to subject- or object-perturbed triples, even for strong models such as Qwen3-Embedding-8B and OpenAI Embedding. This suggests that embedding models are more sensitive to entity substitutions than to relation substitutions. One possible explanation is that subjects and objects are usually content-bearing entities with relatively explicit lexical semantics, making their replacements easier to detect in the embedding space. In contrast, relational words often function compositionally and derive their meaning from the surrounding arguments, rather than carrying standalone semantic content in isolation. Therefore, current embedding models may underrepresent relational changes, leading to high similarity scores even when the underlying triple semantics are substantially altered.

4.3 Relational Task Degradation

We evaluate whether existing embedding models capture human-annotated semantic similarity

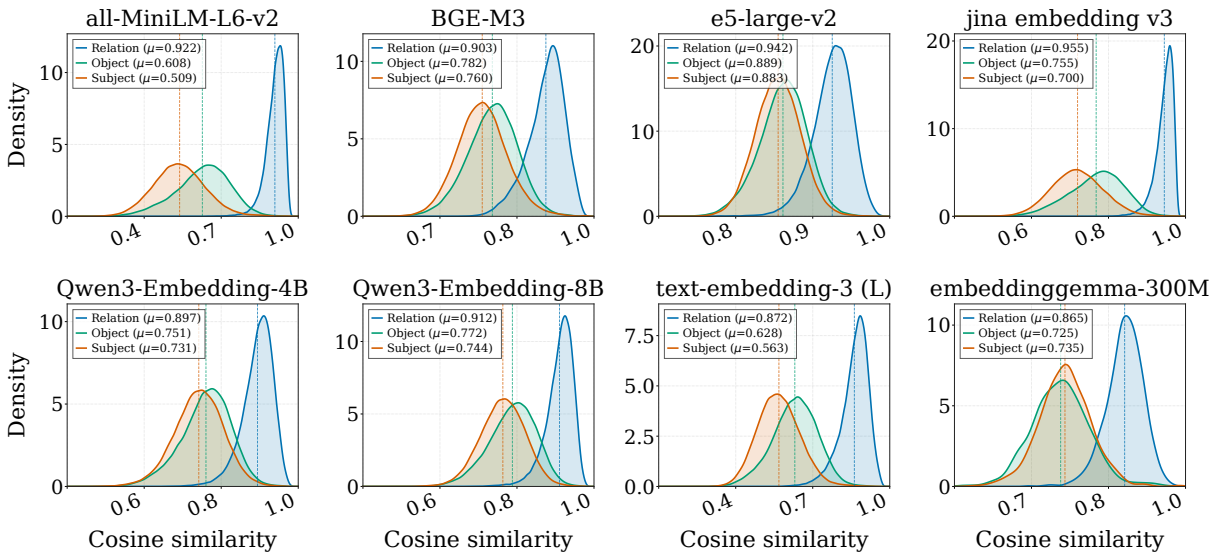

Figure 3: Kernel density estimation of cosine similarities between original triples and their perturbed counterparts, where the relation, object, and subject are perturbed separately.

on public benchmarks. We report performance on relation-sensitive datasets, including SICK-R (the relational subsets) (Marelli et al., 2014) from MTEB (Muennighoff et al., 2023). We further select controlled subsets from SICK-R in which only one component of each triple differs between the paired examples. The entity- and relation-difference subsets are denoted as **EN** and **RE**, with 551 and 75 examples, respectively. We also identify a subset of antonymic relation pairs, denoted as **RE*** with 65 examples.

As shown in Table 2, embedding models perform well on general STS benchmarks but show a lower performance drop on SICK-R, which requires more relational reasoning. On our selected relation-sensitive subsets, the degradation becomes more substantial. For example, Qwen3-Embedding-8B drops from 84.97% on SICK-R to 45.32% on RE*. Models exhibit a moderate performance drop on the entity-difference subset, but their scores decline substantially on the relation-difference subset and further decrease on the antonymic relation subset. This pattern suggests that embedding models are less effective at capturing relational changes than entity-level semantic changes, even though such relational changes can substantially alter the meaning of the original triples.

4.4 Fine-Grained Relation Analysis

We further conduct a fine-grained analysis by grouping relation perturbations into different semantic categories, including spatial, possession, coordination, temporal, comparison, reference, and other relational predicates. For each category, we compute the cosine similarity between the original triples and their perturbed counterparts, and report the average and standard deviation across embedding models in Table 3.

Table 2: Performance on standard embedding benchmarks and relation-sensitive subsets. **EN** denotes entity-difference pairs, **RE** denotes relation-difference pairs, and **RE*** denotes antonymic relation pairs. MTEB $\supset$ STS $\supset$ SICK-R $\supset$ **EN** $\supset$ **RE** $\supset$ **RE***.

Models	MTEB	STS	SICK-R	EN	RE	RE*
MiniLM	56.30	78.95	77.58	72.23	39.15	19.97
Jina-v3	65.50	83.00	77.25	77.22	52.1	33.96
BGE-M3	65.60	80.40	79.71	74.57	46.54	26.93
E5-Large	62.25	83.22	76.49	71.67	37.30	18.81
EGemma	62.00	81.50	80.16	71.58	50.44	35.37
Qwen3-4B	74.60	88.72	83.89	79.01	83.89	44.53
Qwen3-8B	75.40	89.10	84.97	81.25	59.79	45.32
OpenAI Emb(S)	62.26	81.04	76.78	60.05	42.23	26.19
OpenAI Emb(L)	64.59	81.70	79.00	56.80	39.85	24.94

The results show that relation-perturbed triples consistently receive high cosine similarities across all categories and models, suggesting that embedding models are generally insensitive to all relational changes even when the perturbed relation introduces an opposite or substantially different meaning. Although temporal and possession relations yield slightly lower similarities for some models, the overall scores remain high, which suggests that embedding models still fail to clearly separate them well in the embedding space.

Table 3: Average and standard deviation of cosine similarities between original triples and perturbed counterparts across different relation categories.

Model	Spatial	Possession	Discourse	Temporal	Comparison	Causal	Others	All
MiniLM	0.91±0.05	0.88±0.07	0.90±0.05	0.88±0.07	0.90±0.05	0.94±0.05	0.91±0.03	0.91±0.05
BGE-M3	0.91±0.04	0.90±0.05	0.89±0.05	0.91±0.05	0.90±0.05	0.92±0.05	0.91±0.04	0.91±0.04
E5-Large	0.94±0.02	0.93±0.03	0.93±0.02	0.93±0.02	0.94±0.02	0.94±0.03	0.93±0.02	0.94±0.02
Jina-v3	0.95±0.03	0.93±0.04	0.95±0.03	0.94±0.04	0.95±0.04	0.96±0.02	0.93±0.03	0.95±0.03
Qwen3-4B	0.91±0.05	0.91±0.05	0.93±0.04	0.92±0.05	0.93±0.05	0.94±0.04	0.91±0.04	0.91±0.05
Qwen3-8B	0.92±0.04	0.92±0.06	0.93±0.04	0.91±0.05	0.94±0.04	0.95±0.04	0.93±0.03	0.92±0.04
OpenAI-Emb(L)	0.87±0.06	0.84±0.07	0.88±0.05	0.88±0.05	0.89±0.05	0.91±0.05	0.87±0.05	0.87±0.06
Gemma-300M	0.89±0.06	0.87±0.06	0.91±0.04	0.86±0.07	0.89±0.08	0.91±0.06	0.88±0.04	0.89±0.06

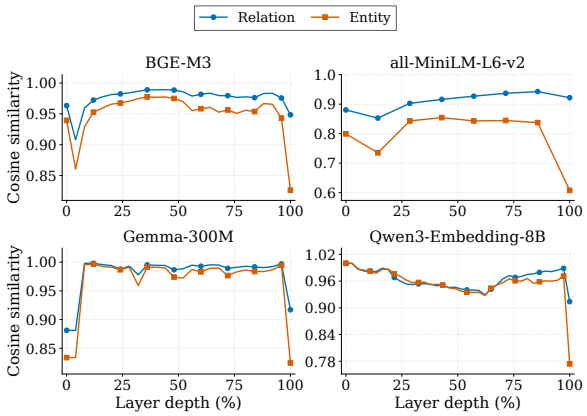

Figure 4: Layer-wise similarity between the original triples and their entity- or relation-perturbed variants.

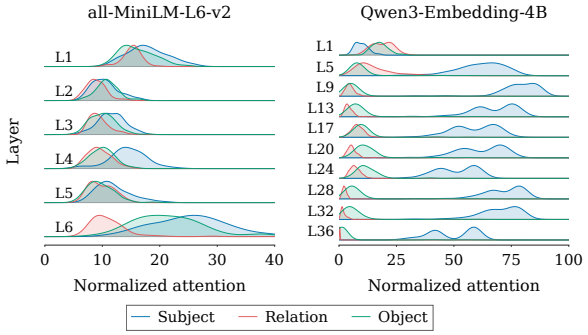

Figure 5: Layer-wise attention distribution for relation, subject and object tokens. We use the relation and LAST token as the query for all-MiniLM-L6-v2 and Qwen3-Embeddding-4B respectively.

4.5 Layer-wise Representation Analysis

Figure 4 shows the layer-wise cosine similarity between the original triples and their entity- or relation-perturbed variants. We can observe that semantic similarity remains consistently high across most layers, from the early layers to the penultimate layers, before dropping sharply in the final one or two layers. These results suggest that fine-grained perturbation sensitivity is largely concentrated near the output layers, and that the resulting embeddings capture relation-induced semantic changes less effectively than entity-induced changes.

4.6 Layer-wise Attention Visualization

Figure 5 visualizes the layer-wise distribution of attention values. We use the last token and the relation token as queries, since the Qwen embedding model uses the last-token representation as the sentence embedding, whereas MiniLM adopts average pooling. For MiniLM, the first five layers show relatively balanced attention across subject, relation, and object tokens, suggesting that relation representations are not strongly biased toward any specific component in early layers. In the final layer, however, the relation query assigns substantially higher attention to subject and object tokens, indicating that relation information is integrated with entity semantics near the output layer. In contrast, for Qwen3-Embedding-4B, the last-token query consistently assigns low attention to relation tokens across layers, while attending much more strongly to subject and object tokens. These results suggest that the final sentence embedding is dominated by entity-level information and receives insufficient relational signals, providing a plausible explanation for the model's weak sensitivity to relation-induced semantic changes.

5 Fine-Tuning Results

In this section, we fine-tune MiniLM-L6-v2 and BGE-M3 to evaluate whether additional supervision can mitigate relational blindness. The fine-tuning dataset combines 40K antonymous relation pairs with public datasets, including SimCSE (274,951 training samples) (Gao et al., 2021) and MS MARCO (400,782 training samples) (Bajaj et al., 2016). We fine-tune the model using AdamW

Table 4: Performance comparison with model finetuning on standard embedding benchmarks and relation-sensitive subsets. **EN** denotes entity-difference pairs, **RE** denotes relation-difference pairs, and **RE*** denotes antonymic relation pairs. MTEB $\supset$ STS $\supset$ SICK-R $\supset$ **EN** $\supset$ **RE** $\supset$ **RE***.

Models	MTEB	STS	SICK-R	EN	RE	RE*
Before finetuning						
MiniLM	53.77	77.99	77.58	72.23	39.15	19.97
BGE-M3	60.36	80.71	79.71	74.57	46.82	26.78
After finetuning						
MiniLM	54.25	79.82	78.68	71.07	52.62	35.00
BGE-M3	60.25	82.73	81.00	76.17	55.74	39.77

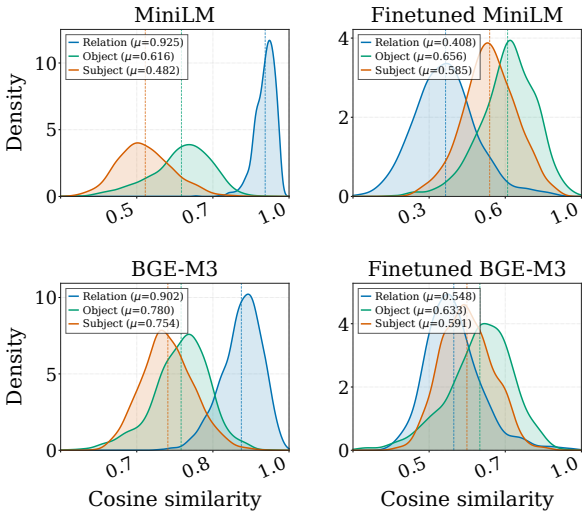

Figure 6: Distribution shift in cosine similarity over original triples and their perturbed counterparts. The left panels show the distributions before finetuning, while the right panels show the distributions after finetuning.

optimizer (2×10^{-6} for BGE, 2×10^{-5} for MiniLM) for up to 50 epochs. The batch size is set to 512 for BGE and 1024 for MiniLM. More training details are provided in Appendix B.2.

5.1 Performance after finetuning

Following Section 4.3, we evaluate the fine-tuned models on both standard embedding benchmarks and relation-sensitive subsets. It allows us to examine whether fine-tuning improves relational awareness and preserve general embedding performance.

Table 4 compares model performance before and after fine-tuning. As expected, fine-tuning with our antonymic relation data substantially improves performance on relation-sensitive tasks. For example, MiniLM-L6-v2 improves by 16% on **RE*** and by 13.47% on **RE**. On the general SICK-R dataset, relation-aware fine-tuning improves performance by 1.10% and 1.29 % for MiniLM-L6-v2 and BGE, respectively. On the STS benchmark, relation-aware fine-tuning also yields marginal gains, with improvements of 1–2%. Meanwhile, performance on general MTEB benchmarks is largely preserved, suggesting that incorporating public datasets during fine-tuning helps maintain general embedding capacity and prevents semantic distortion.

5.2 Distribution Shifts in Relational Representations

In this study, we compare the distribution of cosine similarities on our curated test set before and after fine-tuning. As shown in Figure 6, before fine-tuning, the embedding models are relatively insensitive to relation changes, with high average similarities of 0.925 for MiniLM and 0.902 for BGE-M3. After fine-tuning, the similarities decrease substantially to 0.408 and 0.548, respectively, indicating improved sensitivity to relation-induced semantic changes as expected. However, we do not claim that lower similarity is always better, as the ideal similarity range depends on the downstream task and the expected semantic granularity. Future work should further validate how such similarity shifts translate into performance gains on downstream applications.

6 Conclusion

In this work, we investigate whether modern embedding models faithfully capture semantic compositionality, particularly in relational contexts. We construct a large-scale subject-relation-object (SRO) triplet dataset and use controlled perturbations to evaluate model sensitivity to fine-grained semantic changes. Our results reveal a consistent failure mode, which we term Relational Myopia: existing embedding models often preserve high similarity even when relational predicates induce to the underlying semantics change. Through layer-wise and mechanistic analyses, we further show that relational information is weakly preserved in the latent space and is often dominated by entity-centric lexical cues. Finally, we demonstrate that incorporating relation-aware triplets into fine-tuning substantially improves performance on relation-intensive tasks while maintaining competitive results on general embedding benchmarks. These findings highlight the limitations of current dense embedding models and point toward the need for more relation-aware representation learning.

Limitations

The first limitation is the discrete evaluation design, unlike STS datasets that use human similarity scores on a graded scale. Our curated dataset only ensures semantic changes, but does not capture the full nuance of human similarity judgments.

Second, due to the unavailability of training data for existing embedding models, our fine-tuning experiments are conducted on two small-scale models using a mixture of curated and public datasets. While the results are promising, further experiments across larger models, diverse architectures, and broader training settings are needed to assess the generality of the improvements.

Third, our benchmark comprises concise phrases rather than typical lengthy sentences. Though these propositions remain grammatically intact and semantically natural, this design choice precludes the assessment of document-level contextual dynamics typically encountered in real-world RAG scenarios.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Danilo S. Carvalho, Edoardo Manino, Julia Rozanova, Lucas Cordeiro, and André Freitas. 2025. Montague semantics and modifier consistency measurement in neural language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5515–5529, Abu Dhabi, UAE. Association for Computational Linguistics.

Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. 2024. Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems. *arXiv preprint arXiv:2407.08275*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Cheng-Han Chiang and Hung-yi Lee. 2023. Are synonym substitution attacks really synonym substitution attacks? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1853–1878, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang, Hung-yi Lee, Yung-Sung Chuang, and James Glass. 2023. Revealing the blind spot of sentence encoder evaluation by HEROS. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 289–302, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Irene Heim and Angelika Kratzer. 1998. Semantics in generative grammar.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 114–123.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jabez Magomere, Emanuele La Malfa, Manuel Tonneau, Ashkan Kazemi, and Scott A. Hale. 2025. When claims evolve: Evaluating and enhancing the robustness of embedding models against misinformation edits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22374–22404, Vienna, Austria. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

Marcin Morzycki. 2016. *Modification*. Cambridge University Press.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Aishik Nagar, Ishaan Singh Rawal, Mansi Dhanania, and Cheston Tan. 2025. How do transformer embeddings represent compositions? a functional analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21444–21461.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM 2016 - 2016 IEEE Military Communications Conference*, page 49–54. IEEE Press.

Francis Jeffry Pelletier. 2017. Compositionality and concepts—a perspective from formal semantics and philosophy of language. In *Compositionality and concepts in linguistics and psychology*, pages 31–94. Springer.

Bryan Pickel and Zoltán Gendler Szabó. 2025. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2025 edition. Metaphysics Research Lab, Stanford University.

Paul H Portner and Barbara H Partee. 2008. *Formal semantics: The essential readings*. John Wiley & Sons.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Guanyu Wang, Yuekang Li, Yi Liu, Gelei Deng, Tianlin Li, Guosheng Xu, Yang Liu, Haoyu Wang, and Kailong Wang. 2024a. Metmap: Metamorphic testing for detecting false vector matching problems in llm augmented generation. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, pages 12–23.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024c. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024d. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Data Collection

A.1 Perturbation Strategies

The space of possible subject and object replacements is extremely large. To constrain substitutions to meaningful candidates, we use WordNet supersense labels to guide entity replacement. Specifically, we first link each subject or object of Visual Genome annotations to its corresponding WordNet synset and obtain its supersense label. As shown in Table 5, the Visual Genome dataset covers 26 supersense categories. During perturbation, we restrict entity substitutions to candidates from the same supersense category, and only use the top 25% most frequent entities in that category as replacement candidates. This helps preserve semantic coherence and naturalness.

We manually construct a relation replacement lexicon for relation-level perturbations. The lexicon is organized as a three-level hierarchy, consisting of 7 top-level relation types, 42 fine-grained subcategories, and 345 surface phrases. During perturbation, a relation phrase is replaced only by phrases from the same subcategory, which serves as a controlled replacement table and helps preserve syntactic compatibility while inducing fine-grained semantic changes. Table 6 provides an overview of the top-level relation types.

A.2 LLM-based Commonsense Filtering

Figure 7 presents the LLM-judge prompt used to identify and filter nonsensical triples. To evaluate the reliability of this filtering process, we asked three human annotators to assess the commonsense plausibility of 200 randomly sampled instances from the raw dataset. We then compared the LLM judgments with the human annotations. As shown in Table 7, the LLM filter achieves a precision of 0.969, indicating that it can reliably retain plausible triples and improve the quality of the curated dataset.

Table 5: WordNet supersense categories used to constrain subject and object substitutions. Counts denote the number of entities in each category, and examples show the top three most frequent entities.

#	Lexname	Counts	Top-3 Examples
1	noun.Tops	325	person, food, animal
2	noun.act	776	writing, design, path
3	noun.animal	1,639	giraffe, dog, cat
4	noun.artifact	13,013	building, window, wall
5	noun.attribute	475	distance, ring, patch
6	noun.body	1,929	head, hand, leg
7	noun.cognition	484	thought, reflection, idea
8	noun.communication	2,066	sign, logo, number
9	noun.event	308	wave, waves, ripples
10	noun.feeling	37	vanity, relish, alarm
11	noun.food	1,317	pizza, cake, sandwich
12	noun.group	887	people, line, court
13	noun.location	1,355	field, post, spot
14	noun.object	1,113	sky, ground, beach
15	noun.person	2,347	man, woman, boy
16	noun.phenomenon	407	snow, light, clouds
17	noun.plant	1,270	tree, grass, trees
18	noun.possession	163	lettuce, slice, land
19	noun.process	87	icing, flame, flames
20	noun.quantity	337	whiskers, meter, lot
21	noun.relation	193	part, scale, portion
22	noun.shape	391	doughnut, circle, square
23	noun.state	332	shadow, shade, shadows
24	noun.substance	1,252	water, glass, sand
25	noun.time	229	stage, cycle, calendar

A.3 Semantic Change Filtering

To ensure that each perturbation induces a meaningful semantic change, we apply different filtering strategies for entity and relation replacements. For subject and object perturbations, we measure the semantic distance between the original and replaced entities using Wu–Palmer (WUP) similarity in WordNet. We retain only the top 100K pairs with the largest WUP distances, ensuring that the replacement changes the entity semantics.

For relation perturbations, we construct an antonym-based relation replacement table. Each relation phrase is paired with its antonymous counterpart, such as *above–below* or *before–after*. We then filter the relation-perturbed dataset to retain pairs whose relational semantics are explicitly inverted. After semantic change filtering, the final dataset contains 100K subject-replacement pairs, 100K object-replacement pairs, and 69K relation-replacement pairs.

B Additional Experiments

B.1 Fine-Grained Entity Analysis

Table 8 reports the cosine similarities between original triples and entity-perturbed variants across different WordNet supersense categories. Overall, entity perturbations lead to substantially lower similarities than relation perturbations across all embedding models, indicating that current embedding models are generally more sensitive to entity changes than to relation changes. This pattern is consistent across most entity categories: although the absolute similarity varies by model, the differences across supersense categories are relatively small. For example, MiniLM produces entity-perturbation similarities mostly around 0.50–0.64, while BGE-M3 and Qwen models remain in the range of roughly 0.73–0.83 and 0.71–0.79, respectively. In contrast, relation perturbations yield much higher average similarities, ranging from 0.87 to 0.95 across models. These results

Table 6: Top-level relation types in the relation lexicon and representative examples.

type	count	examples
Spatial	81	*in*, *inside*, *on*, *under*, *between*, *through*, *from*, *to*, *near*, *in front of*
Possession	35	*of*, *belonging to*, *among*, *part of*, *made of*, *member of*, *composed of*
Discourse	49	*and*, *or*, *but*, *however*, *although*, *despite*, *moreover*, *like*, *first*
Temporal	47	*before*, *after*, *while*, *during*, *since*, *until*, *then*, *from . . . to*, *beforehand*, *as of*
Comparison	70	*more than*, *as . . . as*, *most*, *enough*, *all*, *some*, *not*, *per*, *approximately*
Causal	48	*because*, *due to*, *therefore*, *so that*, *if*, *unless*, *except for*, *as a result*
Others	15	*cause*, *lead to*, *contain*, *include*, *consist of*, *relate to*, *depend on*
Total	**345**	

Prompt

System Prompt: You are a strict commonsense relation evaluator. You will receive a short phrase composed of two entities connected by a relational word or phrase. Your task is to decide whether this relation is plausible, natural, and consistent with real-world spatial, physical, or semantic logic. Focus on real-world feasibility, not grammar.
Output only compact JSON:
{
"label":0/1,
"confidence": from 0 to 1
}.
If unsure, use lower confidence.

User Prompt: {sentence}

Figure 7: The prompt used to filter nonsensical triples.

Table 7: Performance of the LLM-based commonsense filter against human annotations.

Metric	**Precision**	**Recall**	**F1-score**
Score	0.969	0.778	0.863

suggest that embedding models to encode entity-level substitutions more strongly than relation-level semantic changes.

B.2 Training Details

Training data: We use three data sources.

- **Antonymous relation pairs:** We use relation-perturbed pairs from our constructed dataset, where each instance consists of a Visual Genome annotation and its antonymous relation perturbation. We use the first 40,000 pairs for training and hold out the remaining 29,244 pairs for validation.
- **SimCSE triplets:** We use 274,951 entailment/contradiction triplets in total. We randomly split them into training and validation sets with an 8:2 ratio, resulting in 219,961 training examples and 54,990 validation examples.
- **MS MARCO:** We use 400,782 query–passage pairs. Using the same 8:2 split, we obtain 320,626 training examples and 80,156 validation examples.

The final objective combines all loss terms with equal weights. We use objective-specific temperature parameters, setting $\tau = 0.10$ for the anti-relation loss, $\tau = 0.05$ for the NLI InfoNCE loss,

Table 8: Average and standard deviation of cosine similarities between original triples and perturbed counterparts across different entity categories.

Supersense	MiniLM	BGE-M3	E5-large	Qwen3-8B	Qwen3-4B	OpenAI-Emb(L)	Jina-v3
artifact	0.53 ± 0.12	0.75 ± 0.06	0.88 ± 0.02	0.74 ± 0.07	0.72 ± 0.07	0.57 ± 0.09	0.70 ± 0.08
plant	0.57 ± 0.12	0.79 ± 0.06	0.90 ± 0.02	0.79 ± 0.06	0.77 ± 0.07	0.63 ± 0.09	0.75 ± 0.08
person	0.56 ± 0.12	0.77 ± 0.06	0.89 ± 0.03	0.73 ± 0.07	0.71 ± 0.08	0.60 ± 0.10	0.73 ± 0.08
food	0.58 ± 0.12	0.77 ± 0.05	0.90 ± 0.02	0.78 ± 0.06	0.76 ± 0.07	0.61 ± 0.08	0.72 ± 0.08
animal	0.57 ± 0.11	0.78 ± 0.06	0.89 ± 0.02	0.75 ± 0.07	0.75 ± 0.07	0.60 ± 0.08	0.70 ± 0.08
communication	0.56 ± 0.12	0.75 ± 0.06	0.89 ± 0.03	0.76 ± 0.07	0.74 ± 0.07	0.61 ± 0.09	0.75 ± 0.08
object	0.59 ± 0.12	0.79 ± 0.06	0.89 ± 0.03	0.79 ± 0.07	0.77 ± 0.07	0.61 ± 0.09	0.75 ± 0.08
substance	0.56 ± 0.12	0.77 ± 0.06	0.89 ± 0.02	0.77 ± 0.06	0.75 ± 0.07	0.60 ± 0.09	0.73 ± 0.08
body	0.60 ± 0.12	0.77 ± 0.06	0.89 ± 0.03	0.77 ± 0.07	0.75 ± 0.07	0.61 ± 0.10	0.75 ± 0.08
location	0.60 ± 0.12	0.77 ± 0.06	0.88 ± 0.03	0.78 ± 0.07	0.76 ± 0.07	0.62 ± 0.09	0.76 ± 0.08
group	0.55 ± 0.12	0.76 ± 0.06	0.88 ± 0.03	0.75 ± 0.08	0.74 ± 0.08	0.59 ± 0.10	0.73 ± 0.08
Tops	0.64 ± 0.14	0.83 ± 0.07	0.90 ± 0.03	0.78 ± 0.09	0.77 ± 0.09	0.66 ± 0.11	0.78 ± 0.10
phenomenon	0.55 ± 0.11	0.78 ± 0.06	0.89 ± 0.02	0.77 ± 0.06	0.76 ± 0.06	0.62 ± 0.09	0.73 ± 0.07
act	0.56 ± 0.13	0.74 ± 0.06	0.88 ± 0.03	0.75 ± 0.07	0.74 ± 0.07	0.59 ± 0.10	0.73 ± 0.08
shape	0.56 ± 0.11	0.75 ± 0.05	0.89 ± 0.02	0.79 ± 0.06	0.76 ± 0.06	0.61 ± 0.09	0.74 ± 0.07
cognition	0.56 ± 0.12	0.75 ± 0.06	0.88 ± 0.02	0.75 ± 0.08	0.75 ± 0.07	0.59 ± 0.10	0.75 ± 0.08
state	0.56 ± 0.11	0.76 ± 0.06	0.89 ± 0.02	0.77 ± 0.07	0.76 ± 0.07	0.59 ± 0.09	0.73 ± 0.08
attribute	0.53 ± 0.12	0.75 ± 0.06	0.88 ± 0.02	0.77 ± 0.07	0.75 ± 0.07	0.58 ± 0.10	0.72 ± 0.08
event	0.52 ± 0.12	0.75 ± 0.06	0.88 ± 0.03	0.75 ± 0.08	0.74 ± 0.08	0.57 ± 0.10	0.70 ± 0.09
possession	0.53 ± 0.12	0.75 ± 0.06	0.88 ± 0.03	0.76 ± 0.08	0.74 ± 0.08	0.57 ± 0.10	0.71 ± 0.09
quantity	0.53 ± 0.12	0.73 ± 0.07	0.88 ± 0.02	0.76 ± 0.07	0.74 ± 0.07	0.57 ± 0.09	0.72 ± 0.08
time	0.61 ± 0.14	0.78 ± 0.07	0.88 ± 0.03	0.77 ± 0.08	0.75 ± 0.08	0.62 ± 0.11	0.76 ± 0.09
relation	0.59 ± 0.12	0.73 ± 0.06	0.88 ± 0.02	0.75 ± 0.08	0.76 ± 0.08	0.62 ± 0.09	0.77 ± 0.08
process	0.50 ± 0.12	0.73 ± 0.06	0.88 ± 0.02	0.77 ± 0.07	0.74 ± 0.07	0.58 ± 0.09	0.71 ± 0.08
feeling	0.52 ± 0.11	0.77 ± 0.05	0.89 ± 0.02	0.74 ± 0.07	0.74 ± 0.07	0.58 ± 0.09	0.72 ± 0.08
Entity (avg.)	0.55 ± 0.12	0.76 ± 0.06	0.89 ± 0.03	0.76 ± 0.07	0.74 ± 0.07	0.59 ± 0.10	0.72 ± 0.08
Relation (avg.)	0.91 ± 0.05	0.91 ± 0.04	0.94 ± 0.02	0.92 ± 0.04	0.91 ± 0.05	0.87 ± 0.06	0.95 ± 0.03

and $\tau = 0.02$ for the MS MARCO retrieval loss. We train the model using AdamW with a learning rate of 2×10^{-5} for MiniLM and 2×10^{-6} for BGE. The batch size is set to 512 for BGE and 1024 for MiniLM.

B.3 Ablation Studies

Table 9 reports the effect of different training data combinations. Training with STS-style data and our antonymous relation pairs improves STS and SICK-R performance, but slightly degrades the overall MTEB score, suggesting that relation-aware training alone may hurt general embedding ability. After adding MS MARCO, the model achieves more balanced performance, improving relation-sensitive test sets while maintaining stable MTEB performance. We also experimented with training only on our antonymous relation pairs, but the model collapsed in the early training stage; therefore, we omit these results from the table. Overall, this ablation shows that relation-aware fine-tuning should be combined with diverse semantic and retrieval data to preserve general-purpose embedding quality.

Table 9: Training ablation results on MTEB and STS benchmarks. (SimCSE) denotes training with SimCSE-style data and antonymous relation pairs, while All further includes MS MARCO retrieval data.

Model	**MTEB**	**STS**	**SICK-R**	**En**	**RE**	**RE***
MiniLM	53.78	78.00	77.58	72.23	39.15	19.97
MiniLM (SimCSE)	53.28	80.38	79.69	73.42	51.60	33.38
MiniLM (All)	54.25	79.82	78.68	71.07	52.62	35.00
BGE-M3	60.37	80.72	79.71	74.57	46.82	26.78
BGE-M3 (SimCSE)	59.46	82.84	80.76	74.57	46.82	26.78
BGE-M3 (ALL)	60.25	82.73	81.00	76.17	52.74	35.77