iBERT: Interpretable Embeddings via Sense Decomposition

Vishal Anand[1], Milad Alshomary[2], Kathleen McKeown[2]

[1]Microsoft, Washington, USA
[2]Columbia University, New York, USA

Correspondence: vishal.anand@microsoft.com

Website: https://iBERT.io

Abstract

We present **iBERT** (interpretable-BERT), an encoder to produce inherently interpretable and controllable embeddings - designed to modularize and expose the discriminative cues present in language, such as semantic or stylistic structure. Each input token is represented as a sparse, non-negative mixture over k context-independent *sense vectors*, which can be pooled into sentence embeddings or used directly at the token level. This enables modular control over representation, before any decoding or downstream use.

To demonstrate our model's interpretability, we evaluate it on a suite of style-focused tasks. On the STEL benchmark, it improves style representation effectiveness by ∼8 points over SBERT-style baselines, while maintaining competitive performance on authorship verification. Because each embedding is a structured composition of interpretable senses, we highlight how specific style attributes get assigned to specific sense vectors. While our experiments center on style, iBERT is not limited to stylistic modeling. Its structural modularity is designed to interpretably decompose whichever discriminative signals are present in the data — enabling generalization even when supervision blends semantic or stylistic factors.

1 Introduction

Neural encoders increasingly serve as the backbone for tasks that rely on nuanced linguistic variation—ranging from authorship attribution and tone-controlled generation to stylistic retrieval and moderation. Yet, most popular encoders, such as SBERT (Reimers and Gurevych, 2019) or SimCSE (Gao et al., 2021), produce dense vector representations that offer no clear control over how style and meaning are encoded. This limits their reliability in domains where representational transparency and stylistic control are essential. These challenges are especially acute in stylistic tasks, where disentangling content from style is not optional but required. Without structured representations, it becomes difficult to attribute stylistic effects, audit model behavior, or intervene in generation pipelines. Prior approaches often rely on proxy supervision and post-hoc explanations, which yield entangled embeddings and incomplete attribution (Danilevsky et al., 2020; John et al., 2019; Elazar et al., 2021).

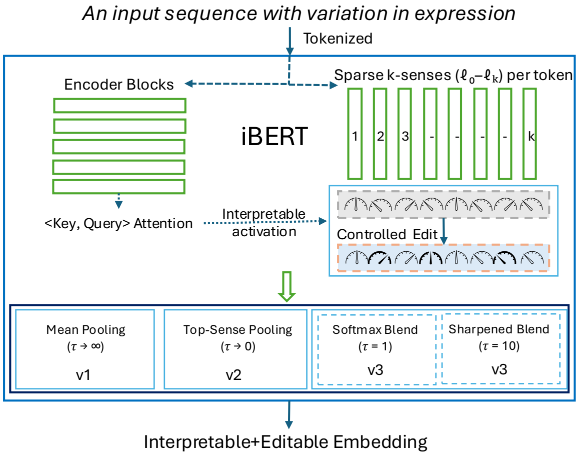


Figure 1: iBERT encodes tokens via k interpretable senses, producing editable and composable sense activations that are used either individually at the token level or pooled via configurable strategies into a global, interpretable embedding suitable for NLP pipelines.

In this work, we ask: *Can we design an encoder where the semantic and stylistic representations are made explicit and controllable within the embedding space—by design, not post-hoc?* To this end, we develop **iBERT**, an encoder architecture with representations that are interpretable and controllable by design. Each token is expressed as a sparse, non-negative mixture over k context-independent *sense vectors*, which can be pooled into sentence embeddings or used directly at the token level. This design enables modular control over specific dimensions of meaning and style that are learned during the training phase, enabling analysis, attribution,

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1413–1429
March 24-29, 2026 ©2026 Association for Computational Linguistics

and targeted edits in embedding space. Our architecture builds on the Backpack formulation of Hewitt et al. (2023), which modeled autoregressive decoding using sparse token-level senses. We adapt this formulation to an encoder-only setting with global pooling, allowing for sentence-level compositionality, bidirectional input, and plug-and-play use across classification and retrieval pipelines.

We train iBERT with a masked language modeling objective on a web corpus to produce sparse, interpretable sense embeddings. We evaluate iBERT on style analysis tasks—benchmarks that test a model's ability to isolate, attribute, and manipulate stylistic variation without entangling semantic content. In particular, by training iBERT on the style analysis task (Patel et al., 2025; Wegmann et al., 2022), we find that it improves style representation accuracy (STEL) by +8 points over SBERT-style baselines while remaining competitive on Style-or-Content (SoC) and PAN authorship verification. Because each embedding is a transparent mixture of sense vectors, we can directly identify stylistic axes (e.g., emoji use, sarcasm, lexical choices) and apply targeted edits in the embedding space.

Though evaluated on stylistic tasks, iBERT is not a style model. It is a general-purpose encoder with decomposable representations—suitable for domains where interpretability and control are prerequisites rather than afterthoughts.

Contributions.

- We present **iBERT**, an encoder with controllable decomposable representations, enabling interpretable token and sentence embeddings.
- We show strong performance across three benchmarks: STEL, SoC, and PAN - achieving up to **+8% STEL** gains over baselines.
- We demonstrate novel interpretability capabilities: probing sense vectors for stylistic traits, ablating senses to identify attribution, and editing embeddings for controlled style transfer.

2 Background and Related Work

Sentence encoders and interpretability. Transformer-based encoders such as SBERT (Reimers and Gurevych, 2019) and SimCSE (Gao et al., 2021) perform well on semantic retrieval and sentence similarity tasks, but their dense embeddings offer little visibility into what linguistic attributes are being captured. Efforts to explain these representations have focused on post-hoc methods such as probing-based explanations (Wallace et al., 2019), saliency and attribution techniques (Jacovi and Goldberg, 2020), and influence functions. As noted by Jacovi and Goldberg (2020), such methods are often applied to representations that were not designed for interpretability, and can yield incomplete or unfaithful explanations. The survey by Danilevsky et al. (2020) further emphasizes the need for *architecturally grounded* interpretability—where structure and transparency are embedded into the model itself.

Stylistic representation learning. Most prior work on stylistic representation relies on *proxy supervision*, typically using authorship identity as a surrogate for stylistic consistency. For example, Wegmann et al. (2022) construct contrastive triplets using same-author, different-topic texts to encourage content-invariant representations, and evaluate model performance using metrics such as STEL and Style-or-Content (SoC). While effective, these approaches still produce dense, entangled embeddings with no inherent interpretability or mechanism for stylistic control.

Patel et al. (2025) introduce a new paradigm: supervising *style directly*, independent of topic, by generating contrastive triplets with controlled lexical and syntactic variation. Their StyleDistance model, along with the SynthSTEL benchmark, enables direct evaluation along 40 distinct content-controlled style-axes. We adopt this direct-style supervision setup, and pair it with our new architectural formulation: an encoder that produces decomposable, interpretable embeddings by design.

Multi-sense representations and Backpack. Backpack-GPT (Hewitt et al., 2023) introduced sparse token-level representations via mixtures over static senses, enabling interpretable and locally steerable generation in autoregressive models. While their formulation supports tokenwise control, it operates within a decoder-centric setup and focuses on in-loop manipulation during generation.

iBERT repositions sparse sense mixtures as a representation-level mechanism, producing globally composable embeddings that enable interpretability and control across tasks. This allows for interpretable, task-agnostic control at the representation level—prior to any decoding—making iBERT compatible with retrieval, classification, and generation pipelines alike. Rather than serving as a generation-side control mechanism, iBERT reframes sense-based modeling as a general-purpose representational interface for controllable NLP.

3 Method

We present **iBERT**, a multi-sense encoder that produces interpretable embeddings by design. iBERT can operate as a token-level encoder (pretrained as MLM) or as a sentence encoder (v1–v3) trained with contrastive supervision. Each token is encoded as a sparse mixture over static *sense vectors*, which are then pooled to form globally composable sentence embeddings. This design enables explicit analysis, control, and manipulation of embeddings. In the paper, the usage of term BERT in a sentential context implies Sentence-BERT (Reimers and Gurevych, 2019).

3.1 iBERT Architecture

Figure 1 shows a high-level overview of our architecture; a full schematic is provided in the Appendix (Figure 5). Each token x_i is first mapped to a standard embedding $e_i = E\,x_i \in \mathbb{R}^d$. A feed-forward layer projects this into k **senses**:

$$C(x_i) = [c(x_i)_1, \ldots, c(x_i)_k] \in \mathbb{R}^{k \times d}$$

In parallel, a transformer encoder produces contextual states $H = [h_1, \ldots, h_n] \in \mathbb{R}^{n \times d}$. These are projected into sense-specific queries and keys: $Q^{(\ell)}, K^{(\ell)} \in \mathbb{R}^{(d/k) \times d}$, allowing us to compute mixture weights $\alpha_{\ell,i,j}$—the contribution of sense ℓ from token x_j to the contextual embedding of token x_i.

$$o_i = \sum_{j=1}^{n} \sum_{\ell=1}^{k} \alpha_{\ell,i,j} \cdot c(x_j)_\ell, \quad o_i \in \mathbb{R}^d.$$

The resulting o_i is a convex mixture of context-independent sense vectors. These representations can be inspected or edited directly. We pretrain this architecture as iBERT-MLM with k=8 senses using masked language modeling, yielding a 171M-parameter encoder, comparable in scale to BERT-base. At k=1, the sense construction block reduces to an embedding matrix, since each token is now mapped to a traditional one dimensional vector.

3.2 iBERT Sentence Embeddings

As a general idea, we take the two-dimensional sense vectors per token and construct a weighted structure to condense the senses together. We apply structured pooling over each token$_\text{i}$ representation o_i, resulting in three variants: v1 (uniform averaging), v2 (top-sense selection), and v3 (soft blending senses wired-into the network by a sense composition function).

iBERT-v1: Mean pooling. The sentence embedding is computed via uniform averaging:

$$s = \frac{1}{n} \sum_{i=1}^{n} o_i.$$

This produces a decomposable embedding, where each dimension reflects contributions from token-level sense activations—supporting interpretation and attribution (Figure 5b).

iBERT-v2: Top-sense pooling. For an input sequence, we compute the total activation per-sense:

$$S_\ell = \sum_{i=1}^{n} \left\| o_i^{(\ell)} \right\|_2, \text{where } o_i^{(\ell)} = \sum_j \alpha_{\ell,i,j} \cdot c(x_j)_\ell$$

We identify the dominant sense (*per sequence*) as $\ell^* = \arg\max_\ell S_\ell$, and retain only that sense across all tokens to compute the sequence embedding:

$$s = \frac{1}{n} \sum_{i=1}^{n} o_i^{(\ell^*)}.$$

This encourages stricter top-1 alignment between input sequences and individual senses. The process is repeated independently for each input sequence (Figure 5c).

iBERT-v3: Softmax-weighted pooling. We define a general family of models to interpolate behaviors between v1 and v2 using softmax over sense norms, controlled by a sense composition variable τ. Once τ is set, the v3 trained model is fixed with the pooling structure and is not replaced afterwards.

$$S_\ell = \sum_{i=1}^{n} \left\| o_i^{(\ell)} \right\|_2, \quad \pi_\ell(\tau) = \mathrm{softmax}(S_\ell/\tau),$$
$$s_\tau = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{k} \pi_\ell(\tau) \cdot o_i^{(\ell)}.$$

This general formulation defines v3, where:

- $\tau \to \infty$: recovers v1 (uniform averaging)
- $\tau \to 0$: recovers v2 (hard top-sense selection)
- $0 < \tau < \infty$: soft blending across senses

This generalized framework allows us to study how the sharpness of sense composition affects interpretability and performance (§4, Appendix A.3).

3.3 Training Pipeline

We train iBERT in two stages. Unlike prior work (Patel et al., 2025; Wegmann et al., 2022) which fine-tuned off-the-shelf encoders, our decomposable design requires MLM pretraining to instantiate meaningful sense vectors.

3.3.1 Stage 0: MLM Pretraining.

We pretrain iBERT-MLM from scratch on 5% of FineWeb (Penedo et al., 2025), a 15T-token web-scale corpus. This 750B-token slice offers broad linguistic and stylistic diversity while remaining computationally efficient. We use standard masked language modeling to learn token-level sense activations, producing a 171M-parameter encoder. For comparison, we also train a standard BERT model using identical settings and data.

3.3.2 Stage 1: Style Representaion Learning

To train iBERT on the task of style analysis, we use contrastive triplets $\langle s, s^+, s^- \rangle$ targeting stylistic variation and create iBERT-v1-v3. Training data is drawn from *StyleSynth* by Patel et al. (2025): content-controlled, synthetic style triplets; and from author-labeled triplets with topic control by Wegmann et al. (2022).

We apply the InfoNCE loss to encourage closeness between s and s^+, while pushing s away from s^-. Pooling weights $\pi_\ell(\tau)$ are shared between the anchor and positive, encouraging consistent sense-axis usage for stylistically similar inputs.

To ensure architectural parity, we also train *BERT* and then *SBERT* baselines using the two-stage setup: a BERT-base model pretrained on the same 5% FineWeb slice, followed by contrastive fine-tuning on anchored triplets. This controls for pretraining variance and isolates the contribution of our multi-sense architecture.

Model size. iBERT-v1–v3 has 171.4M parameters and SBERT has 149.7M. Inference is approximately 2% slower owing to sense composition. Tradeoffs are discussed in Section 7.

4 Experimental Setup

4.1 Pretraining and Fine-tuning

We reuse iBERT-MLM and BERT models from Stage 0 (§3.3.1), trained from scratch on 5% of FineWeb. Sentence-level variants—iBERT-v1–v3 are fine-tuned on contrastive triplets to learn style representation using either:

- **StyleDistance - SD** (Patel et al., 2025): 50k synthetic triplets (StyleSynth) with controlled lexical and syntactic variation. Median token-size is 19 (Appendix Table 7).
- **Wegmann - WG** (Wegmann et al., 2022): 40k author-anchored Reddit triplets with similar topics. Median token-length is 24.

Training follows the loss function of each dataset: for StyleSynth we use the InfoNCE objective as in Patel et al. (2025), while for Wegmann triplets we apply the margin-based contrastive loss used in Wegmann et al. (2022). In both cases, pooling weights $\pi_\ell(\tau)$ are shared between the anchor and positive to enforce consistent sense usage. Additional details are in Appendix A.4.

Baselines. To isolate the effect of our multi-sense architecture, we compare against contrastively trained versions of the same BERT encoder from Stage 0. We fine-tune this model to create:

- BERT$_{\text{SD}}$: trained on StyleSynth triplets with InfoNCE loss (as in Patel et al. (2025));
- BERT$_{\text{WG}}$: trained on Wegmann triplets with the margin-based triplet loss (as in Wegmann et al. (2022));
- BERT$_{\text{WG+SD}}$: trained on both datasets with InfoNCE loss (reflecting the strongest setup reported by Patel et al. (2025)), which is listed as the prior state-of-the-art performing model called StyleDistance.

BERT$_{\text{WG}}$ corresponds to the SBERT-style model of Wegmann et al. (2022), while BERT$_{\text{WG+SD}}$ mirrors the best-performing StyleDistance setup of Patel et al. (2025). All settings share the same backbone, initialization, and optimizer for a controlled comparison.

Backbone Consistency. Prior work used DistilRoBERTa-based encoders (e.g., Patel et al. (2025), Wegmann et al. (2022)). Since iBERT is trained from scratch, we also train a new BERT-base encoder using the same setup as iBERT-MLM. This serves as the backbone for BERT$_{\text{SD / WG / WG+SD}}$, replacing DistilRoBERTa in their respective setups and enabling a fair performance comparison. Full training protocol is detailed in Appendix A.4.

4.2 Evaluation Benchmarks

We evaluate all models across four direct style tasks (STEL-40$_{\text{SD}}$, SoC-40$_{\text{SD}}$, STEL-5$_{\text{WG}}$, SoC-5$_{\text{WG}}$)

and one proxy-style task (Authorship Verification - PAN), using both 128- and 512-token variants of the backbone BERT models.

Direct-style evaluation.

- **STEL**: multi-class classification over 40 style labels (StyleSynth) and 5 style classes (Wegmann), reported as weighted accuracy.
- **SoC**: average binary classification accuracy between positive/negative style polarities per class (e.g., formal vs. informal).

Proxy-style evaluation. **PAN** authorship verification: binary classification of same vs. different author pairs (PAN 2011/13/14/15), reporting AUC. Treated as a proxy for style consistency, though not grounded in specific stylistic dimensions.

5 Results

5.1 STEL Performance

Table 1 reports style representation effectiveness (STEL) across training regimes (Wegmann and Nguyen, 2021). Across the board, iBERT models outperform their BERT counterparts, underscoring the utility of structured, interpretable pooling for capturing stylistic signals. Under SD-ONLY training, iBERT-v3-1 achieves the highest STEL scores, yielding 38.3% (128 tokens) and 38.2% (512 tokens)—a substantial gain of +6.7 and +7.6 points over BERT. Close contenders include iBERT-v1 and iBERT-v3-10, both of which exhibit similarly strong performance, suggesting that soft, sense-weighted pooling is especially effective when guided by direct-style supervision.

Interestingly, iBERT-v2 consistently lags behind v1 and v3 across all settings. This is expected: its hard top-sense selection mechanism introduces sharp sparsity that likely harms generalization by collapsing compositional diversity. In contrast, iBERT-v3-10 is not only among the top performers but also the most stable across settings—offering a smoother inductive bias that bridges the extremes of v1 (uniform blend) and v2 (hard top-1 sense).

Even under joint supervision from WG+SD, where noisy author-level triplets may dilute sense-specific stylistic signal, iBERT variants maintain their edge. iBERT-v1 secures the best STEL at 128 tokens (38.6%), while iBERT-v3-1 leads at 512 tokens (36.4%), outperforming BERT by approximately 6 points in both cases.

Finally, in the WG-ONLY setting, where labels reflect author identity rather than explicit styles, all models see degraded performance. Nevertheless, iBERT-v3 and iBERT-v1 is at par with BERT, signaling that even without style supervision, iBERT retains competitive structure-aware generalizability. These results affirm that iBERT's compositional inductive bias enables robustness without compromising direct-style effectiveness.

5.2 SoC Performance

For the SoC metric—evaluating separability of stylistic polarity—iBERT-v3-1 achieves the highest overall scores: 92.8% (128 tokens) and 92.0% (512 tokens) under SD-ONLY, outperforming all baselines. These results show that iBERT can preserve or even enhance polarity sensitivity while maintaining interpretability.

In contrast, iBERT-v2 underperforms across settings (e.g., 88.7% under SD-ONLY), often lagging behind both v1 and BERT. We attribute this to its top-1 sense pooling: by forcing each sequence to align with a single dominant sense, iBERT-v2 severely restricts the number of senses that can specialize for the same data-points observed. This architectural bottleneck reduces expressive capacity, especially for fine-grained polarity distinctions that require distributed compositionality.

Under WG-ONLY training—where supervision is provided weakly via author labels rather than direct stylistic polarity—all models see a dramatic drop in SoC performance. BERT, for example, falls from 91.5% (SD-ONLY) to just 27.5%, while iBERT variants drop even further (e.g., 8.0% for v1). Despite this collapse, BERT appears to outperform iBERT in this setting. We hypothesize this is due to *content leakage*: BERT implicitly entangles semantic and stylistic attributes, using topic or lexical cues as proxies for style. While this can yield fragile SoC signal in weakly supervised settings, it sacrifices interpretability and modularity. In contrast, iBERT's explicit disentanglement fails to recover polarity distinctions without true stylistic supervision—a tradeoff aligned with its core design goals.

Taken together, these results affirm that iBERT excels when trained with style-specific signals, achieving strong polarity separability without compromising compositional control or interpretability. Further analysis of the pooling sharpness mechanism (§3.2) translating to performance and stability is discussed in Appendix A.3.

Data	Models	128 tokens			512 tokens		
		STEL↑	SoC↑	PAN↑	STEL↑	SoC↑	PAN↑
SD-only	BERT	31.6 ± 0.78	91.5 ± 0.56	*60.57*	30.6 ± 0.86	*91.7* ± 0.33	**63.52**
	iBERT-v1	*37.5* ± 0.41	*92.3* ± 0.33	**60.88**	**38.9** ± 0.81	*92.0* ± 0.29	59.49
	iBERT-v2	28.5 ± 1.13	88.7 ± 0.68	58.95	31.4 ± 1.46	91.1 ± 0.68	59.41
	iBERT-v3-1	**38.3** ± 0.62	**92.8** ± 0.48	58.15	*38.2* ± 0.75	*92.0* ± 0.32	59.39
	iBERT-v3-10	*38.1* ± 0.44	*92.5* ± 0.30	58.29	36.0 ± 0.21	**92.0** ± 0.30	59.41
WG+SD	BERT	30.6 ± 0.93	88.9 ± 0.53	58.96	30.4 ± 0.94	*89.6* ± 0.38	**62.76**
	iBERT-v1	**38.6** ± 1.57	*90.0* ± 0.22	58.96	*35.7* ± 1.76	*89.5* ± 0.70	*61.53*
	iBERT-v2	33.4 ± 0.54	89.1 ± 0.50	58.47	32.3 ± 1.13	89.0 ± 0.54	59.54
	iBERT-v3-1	*38.0* ± 0.85	**90.2** ± 0.30	58.56	**36.4** ± 0.71	**90.1** ± 0.58	*61.43*
	iBERT-v3-10	*38.0* ± 0.88	*89.6* ± 0.65	**60.96**	34.8 ± 0.93	*89.8* ± 0.32	60.34
WG-only	BERT	27.1 ± 0.70	**27.5** ± 1.18	58.20	**27.3** ± 0.97	**28.9** ± 1.53	**61.92**
	iBERT-v1	*28.5* ± 0.86	8.0 ± 0.62	*59.29*	*27.2* ± 0.74	8.0 ± 0.73	60.74
	iBERT-v2	25.9 ± 0.76	6.7 ± 0.65	57.77	*26.8* ± 0.56	6.3 ± 0.75	58.63
	iBERT-v3-1	*28.1* ± 0.59	7.4 ± 0.42	*59.47*	*27.1* ± 0.73	7.4 ± 0.50	60.87
	iBERT-v3-10	**28.5** ± 0.58	7.3 ± 0.41	**59.87**	26.5 ± 0.96	7.4 ± 0.68	*61.07*

Table 1: Direct-style: STEL and SoC reported on the **45 style groups** (40 of StyleSynth + 5 of Wegmann STEL). Proxy-style: PAN reports average AUC% on PAN11/13/14/15 authorship verification tasks. The models are trained on corresponding Data triplets for **5 runs**. Median token-lengths: 19 for SD, and 24 for WG (Appendix Table 7). $\text{BERT}_{\text{WG+SD}}$, named StyleDistance, is the prior state-of-the-art setup.

5.3 Authorship Verification (PAN)

To assess generalizability of learning style to solve the authorship attribution task, we use the PAN11–15 benchmarks, a suite of proxy-style tasks requiring models to distinguish between same-author and different-author text pairs. Unlike STEL or SoC, PAN tasks blend stylistic and semantic variation, providing a rich probe of compositional generalization. The results are in Tables 1 and 4.

WG training dataset is a mixed dataset, containing author identity supervision with intertwined semantic and stylistic signals. Unlike BERT, which entangles all available cues, iBERT is architecturally designed to tease apart these differentials in an interpretable fashion. This enables it to maintain competitive performance even when the supervision is noisy or not explicitly style-specific—while still excelling on clean, style-focused tasks like STEL and SoC.

At 128 tokens, all iBERT variants match or slightly outperform BERT, with iBERT-v3-10 achieving the highest AUC (59.87%) for WG. This is notable not because iBERT ignores semantic information, but because it selectively organizes it in disentangled form—generalizing well even when stylistic cues are weakly defined.

At 512 tokens, BERT slightly outperforms iBERT (61.92% vs. 61.07% for v3), with the largest margins on PAN13 and PAN15, that are characterized by semantic drift and cross-topic variation. This suggests that BERT benefits more from extended context, due to its architecture's tight coupling of semantic and stylistic features. In contrast, iBERT's modular structure—designed to isolate stylistic axes—does not directly exploit semantic continuity. Still, the gap stays slim, highlighting the semantic signals acquired in MLM stage remain accessible even in disentangled representations.

Interestingly, even iBERT-v2, which lagged on STEL and SoC, performs competitively on PAN. This underscores that semantic signals are preserved, even when stylistic specialization is weak. Looking across datasets, iBERT performs best on PAN11 and PAN14, where stylistic consistency is high and maintains near-parity on PAN13 and PAN15, where semantic variation dominates.

While WG provides author identity supervision, its examples entangle stylistic and semantic cues: offering iBERT no clean signals to isolate. In contrast, SD supplies disentangled stylistic supervision but lacks authorial labels entirely. This creates a supervision-task mismatch for iBERT, which is architecturally primed to extract interpretable signals. Yet it adapts: from WG, it organizes conflated authorial traits (e.g., tone, topic, lexical patterns) into structured axes; from SD, it learns clean stylistic subspaces that can approximate the mixed style-content clusters found in PAN's test sets. That iBERT performs strongly across PAN bench-

Sense	Top-Aligned Style Axes	Emergent Theme
$\ell = 0$	*With Emojis / No Emojis* *Frequent / Infrequent Conjunctions* *Frequent / Infrequent Personal Pronouns*	Surface-level markers
$\ell = 1$	*All Upper Case / Proper Capitalization* *Text Emojis / No Emojis* *Long / Short Average Word Length* *With / Without Number Substitution*	Orthographic and visual style
$\ell = 2$	*Humorous / Non-Humorous* *Sarcastic / Non-Sarcastic* *Metaphoric / Literal* *Offensive / Non-Offensive*	Affect and expressive tone
$\ell = 3$	*All Lower Case / Proper Capitalization* *With Misspellings / Normal Sentence*	Textual correctness and noise
$\ell = 4$	*More / Less Frequent Function Words* *With / Without Nominalizations*	Functional grammar
$\ell = 5$	*Active / Passive* *Contracted / Non-Contracted*	Syntactic voice and register
$\ell = 6$	*Frequent / Infrequent Pronouns* *More / Less Frequent Verbs*	Pronoun and verbal focus
$\ell = 7$	*With / Without Determiners* *Certain / Uncertain*	Grammatical commitment

Table 2: Representative sense-style alignments in iBERT-v3-10, based on sense activation (Table 6). All listed axes are the top-aligned style for their respective sense (i.e., highest probing activation).

marks—despite this misalignment—demonstrates its ability to decompose and generalize whichever discriminative signals are available, even when style and semantics are blended.

5.4 Interpretability Analysis of iBERT

As highlighted before, the main goal of developing iBERT is to encode inputs as mixtures of interpretable sense vectors. To examine whether specific senses specialize in capturing coherent stylistic structure, we analyze the alignment between sense dimensions and styles via probing and controlled ablations. Table 2 groups the style pairs most strongly aligned with each sense for iBERT-v3-10, identified by first finding the highest activation senses per style, and then performing combinatorial optimization to group them (Appendix Table 6). Clear thematic clustering is observed: i.e., sense ℓ=2 aligns with affective and expressive tone (sarcasm, metaphor), while ℓ=0 captures surface-level markers like emoji use and personal pronouns. These patterns suggest that individual senses structurally specialize in distinct stylistic attributes.

To validate these groupings, Table 3 reports the change in cosine distance between opposing style centroids before and after removing each sense contribution in final representation. When ablating a style-aligned sense, the separability between style polarities drops substantially (e.g., 66% for $\ell = 3$), confirming that key stylistic information is isolated within that sense. In contrast, most non-target styles exhibit low or negligible change in distance, indicating minimal collateral disruption: a hallmark of localized specialization.

Interestingly, for ℓ=6 and ℓ=7, ablating these senses harms separation among non-aligned styles (i.e., negative ΔDist), suggesting they may encode semantic content that generalizes across style boundaries. This aligns with the style groups they capture (e.g., pronoun and verb usage, determiners, certainty), which plausibly reflect broader discourse semantics. These observations highlight a nuanced balance between stylistic and semantic organization within iBERT's modular space.

Taken together, these results demonstrate that sense dimensions in iBERT naturally organize around interpretable stylistic clusters, enabling explicit, localized editing of style without disrupting unrelated attributes.

5.5 Visualize Targeted Editing via Ablation

To visualize iBERT's ability to localize and disentangle style, we conduct controlled ablations over its sense vectors. Figure 2 shows t-SNE projections for three representative styles ablated along their most aligned senses (Table 2). In all cases, ablating the target sense causes the positive samples (red) to move toward the negative centroid (gray), with relative distance dropping up to 84%. These reflect localized edits validating iBERT's controllability.

5.6 Visualize Non-Target Style Ablation

To ensure that iBERT's sense editing visualizations (subsection 5.5) are not generic artifacts of perturbation, Figure 3 ablates unrelated senses for unrelated (non-target) style contrasts. In all cases, positive and negative samples remain well-separated, even as the positive cluster slightly shifts. This affirms that sense vectors act locally, without broad entanglement across unrelated styles.

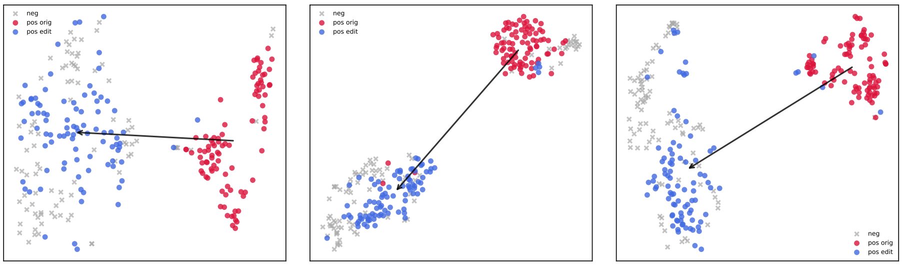
Note that t-SNE primarily preserves local topology rather than absolute displacement; thus, these visualizations should be interpreted qualitatively. The clear contrast between targeted collapse (Figure 2) and stable separation in control (Figure 3) supports our claims of modularity and editability.

5.7 Quantifying Locality of Sense-Editing

A core utility for controllable representations is *locality*: editing a target attribute should only affect that attribute while minimally disturbing unrelated dimensions. While Figure 2 illustrates targeted edits qualitatively, we now formalize this behavior.

SENSE	TARGET-ALIGNED STYLES				NON-TARGET STYLES		
	#STYLES	ORIG	EDIT	ΔDIST (%)	ORIG	EDIT	ΔDIST (%)
$\ell = 0$	3	0.812	0.459	**38.5**	0.405	0.278	1.2
$\ell = 1$	4	0.725	0.321	**58.0**	0.390	0.250	15.2
$\ell = 2$	4	0.563	0.310	**40.4**	0.412	0.299	-2.2
$\ell = 3$	2	1.333	0.469	**65.7**	0.414	0.324	-4.3
$\ell = 4$	2	0.336	0.168	**46.3**	0.392	0.232	7.7
$\ell = 5$	2	0.326	0.237	**26.1**	0.452	0.271	8.7
$\ell = 6$	2	0.531	0.395	**25.5**	0.403	0.272	-10.7
$\ell = 7$	2	0.353	0.328	**7.0**	0.488	0.347	-2.6

Table 3: Impact of sense-level ablation in iBERT-v3-10, using style groups from Table 2. We report mean cosine distance to the opposite style centroid, before and after ablating sense ℓ. Left: styles aligned with ℓ; Right: all others. ORIG = original; EDIT = post-ablation; ΔDIST = relative distance reduction (↑ = stronger disentanglement).



(a) Text Emojis / No Emojis ($\ell = 1$, ΔDist: 84 %)

(b) With/Without Nominalizations (ℓ=4, ΔDist: 67%)

(c) All Lowercase / Proper Caps ($\ell = 7$, ΔDist: 70%)

Figure 2: Style-edit t-SNEs for iBERT-v3-10, ablating the most aligned sense ℓ for a given style. Red: original positive samples; Blue: edited ablated positive samples; Gray: negative samples. Arrows are from original-positive-centroid, to: edited-positive-centroid. Edits control semantics along: (a) visual form, (b) syntactic function, and (c) grammatical commitment. ΔDist: relative decrease in mean distance of positive samples to the negative centroid.

For a given style x (e.g., *Metaphorical / Literal*), we identify the sense ℓ^* with the highest probing activation and ablate it during encoding by zeroing the corresponding sense gain. Let P_x denote the positive examples for style x, and P_x^{orig}, P_x^{edit} be their embeddings before and after ablating ℓ^*, respectively. We compute cosine distance between a set of embeddings A and a centroid μ as:

$$d(A, \mu) = \frac{1}{|A|} \sum_{a \in A} (1 - \cos(a, \mu)) .$$

The *target editability* is then:

$$\Delta_x = \frac{d(P_x^{\text{orig}}, \mu_x^-) - d(P_x^{\text{edit}}, \mu_x^-)}{d(P_x^{\text{orig}}, \mu_x^-)},$$

where μ_x^- is the centroid of negative examples for style x.

To measure *non-target editing* (collateral effects), we evaluate how the edits for x impact *other* stylistic properties $y \neq x$, using μ_y^-, which is the centroid of non-target y's negative examples:

$$\Delta_y = \frac{d(P_x^{\text{orig}}, \mu_y^-) - d(P_x^{\text{edit}}, \mu_y^-)}{d(P_x^{\text{orig}}, \mu_y^-)}.$$

We report the following aggregate locality metrics:

$$\text{Avg. Other Shift} = \frac{1}{|\{y \neq x\}|} \sum_{y \neq x} \Delta_y$$

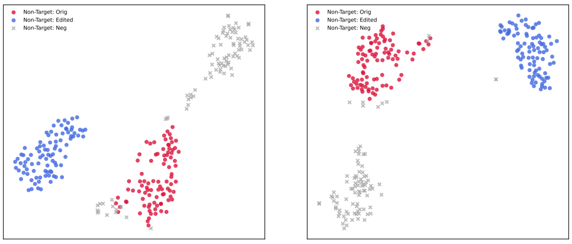
A useful and disentangled iBERT sense should ideally yield a large, positive Δ_x while keeping Avg. Other Shift low, indicating minimal collateral influence on unrelated style attributes.

Table 3 shows that ablating a style-aligned sense yields large reductions in target separability (up to 65%), while inducing minimal average shift on non-target styles. This confirms that stylistic attributes are concentrated in specific sense dimensions rather than distributed diffusely across the embedding.

As an added control, we refer again to Figure 3, which shows that ablating a non-primary sense (i.e., not the top sense ℓ^*) for a given style leaves its attribute structures intact. This supports the view that iBERT's sense representations are modular, with sense edits affecting the intended attributes, rather than causing broad changes across the space.

6 Discussion

Interpretability without performance tradeoff. iBERT delivers inherently interpretable embeddings, while maintaining competitive performance



(a) $\ell = 0$ ablated that aligns with More/Less conjunctions

(b) $\ell = 2$ ablated that aligns with Sarcasm / No sarcasm

Figure 3: All Upper Case/Proper Capitalization style (ℓ^*=1) maintains separation on ablating other senses (ℓ={0, 2}). Colors as in Fig. 2- Red: original, Blue: edited, Gray: negatives.

across both style-intensive and mixed-style tasks. Because each input is represented as a sparse mixture of sense vectors, users can inspect, edit, or ablate specific dimensions with explicit attribution to stylistic or semantic features. Unlike post-hoc attribution methods, these controls are embedded in model's structure that directly alters representation.

Disentanglement through structure and signal. Our results confirm that architectural inductive bias alone is insufficient: meaningful disentanglement emerges only with high-quality supervision (e.g., SynthSTEL). Under proxy-style signals (e.g., WG-only), sense specialization weakens, and style polarity collapses. This co-dependence of structure and supervision underscores a key design insight: iBERT is not style-specific, but learns to modularize whichever axes are discriminative in the data.

Localized edits, global control. Ablating a single sense (e.g., ℓ=1 for emoji, or ℓ=4 for nominalization) shifts the representation toward the negative style without disrupting unrelated styles. These transformations are visible both numerically (Table 3) and spatially (Fig. 2), confirming that iBERT learns axis-aligned subspaces that can be controlled with precision. This kind of intervention is infeasible in dense encoders like SBERT or BERT.

Semantic resilience in modular space. Despite modularizing stylistic signals, iBERT retains generalization capacity in semantically entangled tasks (e.g., PAN13/15). Even sense vectors associated with stylistic axes (e.g., ℓ=6, 7) appear to encode semantic scaffolding, since ablating them harms non-target style separability. This indicates that the modularity does not rigidly isolate content, but supports a soft balance between style and semantics.

Bridging representation learning and sociolinguistics. Emergent sense specializations align with sociolinguistic constructs such as formality, register, modality, and expressive tone. This alignment arises without explicit annotation, suggesting iBERT's decomposition mirrors meaningful axes of human communication. These findings open pathways for interpretable modeling in domains like forensics, social analysis, and fairness auditing.

Applications. iBERT is not a style model, but a general architecture to tease out discriminative signals into interpretable senses, and it supports a range of feature—conditioned retrieval, classifier debiasing via sense ablation, latent data augmentation, and embedding-level control in RAG pipelines. See Appendix A.1 for more details.

7 Conclusion

We introduced **iBERT** (interpretable-BERT), a modular encoder that produces sparse, interpretable, and controllable embeddings without compromising performance. By structuring each input as a mixture over reusable sense vectors, iBERT supports inspection, editing, and attribution of linguistic properties within the representation.

iBERT matches or outperforms dense baselines across stylistic and semantic tasks like STEL, SoC, and PAN, while enabling: (a) disentangled encoding of style and semantics via sense-probing, (b) precise edits (e.g., formality), and (c) alignment of emergent senses with sociolinguistic axes.

Rather than relying on post-hoc interpretability methods, iBERT is designed to be *explainable by construction*. Its structure reveals how representations evolve and interact—making it a viable backbone for transparent, modular NLP.

We release our models to encourage further exploration of compositional embeddings, zero-shot control, and sense-aware generation. iBERT offers a step toward embedding models that are not only powerful, but inherently understandable.

Model and Code

We release the models, demonstration, and code for other researchers to use. These are made available at https://github.com/vishalanand/iBERT and https://iBERT.io .

Limitations

While iBERT enables modular, interpretable representations with strong empirical performance, a few limitations remain. The model's design introduces a tradeoff between disentangled structure and tight semantic coupling: in tasks requiring deep contextual integration, this can lead to slight performance gaps compared to dense encoders.

In addition, sense specialization can be sensitive to sequence length. Our training datasets: SynthSTEL (median 19 tokens) and WG (24 tokens) offer limited context, which may constrain the model's ability to fully activate or separate sense subspaces. This could partially explain weaker gains in tasks involving longer-form semantics.

Lastly, while we focused on English, extending sense-level modularity to morphologically rich or low-resource languages remains an open direction.

These limitations offer promising opportunities to expand iBERT into a broader, language-agnostic foundation for interpretable and controllable representation learning.

Ethics Statement

iBERT is developed as a general-purpose encoder that produces interpretable, controllable sentence representations by decomposing inputs into structured mixtures over sense vectors. This modular design supports greater transparency in how linguistic features—whether semantic, syntactic, or stylistic—contribute to downstream decisions.

We believe such structured interpretability fosters accountability in language technologies. However, we recognize potential misuse: modular representations that expose linguistic traits may be leveraged for tasks like profiling or authorship attribution, which can carry privacy risks in sensitive settings. These risks are pronounced when models are applied without consent, or outside the distribution of the data used during training.

iBERT is trained on public data, including synthetic edits (StyleSynth) and large-scale web corpora. As with all pretrained models, underlying biases in this data may influence the dimensions that emerge. We encourage careful evaluation before deployment in applications with societal or demographic impact.

We release our models and code for research use, with the goal of encouraging safe, transparent, and interpretable alternatives to opaque neural representations.

Acknowledgments

We thank John Hewitt (Columbia University) for helpful discussions during experiment and model design. We thank Dr. Cheng Wu (Microsoft) for his insights during modeling and experiment design, and supporting our research. We thank Ajay Patel (University of Pennsylvania), Anna Wegmann (Utrecht University), Nicholas Andrews (Johns Hopkins University), Joel Tetreault (Dataminr), and Sudha Rao (Microsoft Research) for granting access to relevant datasets. We also thank David M. Rothschild (Microsoft Research), Vishal Chowdhary (Microsoft Office AI), Yasaman Ameri (Visa Inc.), and Aparna Balagopalan (MIT) for early conversations that helped shape the first author's thinking around research problem formulation and scoping.

References

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hewitt, John Thickstun, Christopher Manning, and Percy Liang. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9103–9125, Toronto, Canada. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings* *of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2025. The fineweb datasets: decanting the web for the finest text data at scale. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

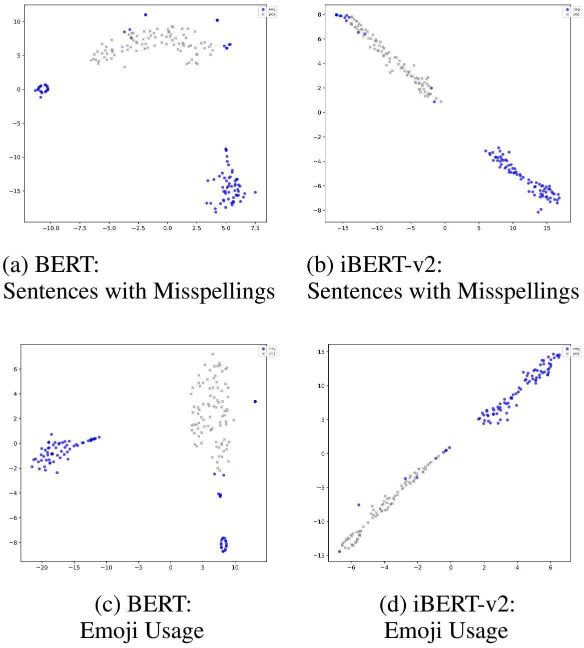
Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. 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): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Anna Wegmann and Dong Nguyen. 2021. Does it capture STEL? a modular, similarity-based linguistic style evaluation framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7130, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.



(a) BERT:
Sentences with Misspellings

(b) iBERT-v2:
Sentences with Misspellings

(c) BERT:
Emoji Usage

(d) iBERT-v2:
Emoji Usage

Figure 4: t-SNE projections of sentence embeddings for SynthSTEL. iBERT consistently separates contrastive style variants (e.g., misspelled sentences, emoji usage) better than BERT, showing clearer margins and lower entanglement. Blue points represent positive samples and gray crosses represent negative samples. Despite iBERT-v2 being the most underperforming iBERT variant it still matches BERT's performance, and the separation of positive and negative styles is cleaner.

A Appendix

A.1 Broader Applications

iBERT enables embedding-level interpretability and control, unlocking several downstream capabilities beyond benchmarking.

Style-Conditioned Retrieval. Rather than relying on keyword overlap or dense similarity, queries can be modified in latent space to match stylistic attributes (e.g., formality). This enables retrieval that is both semantically relevant and stylistically aligned—particularly useful in assistant and RAG pipelines where tone coherence matters.

	Models	128 tokens					512 tokens				
		Avg	PAN11	PAN13	PAN14	PAN15	Avg	PAN11	PAN13	PAN14	PAN15
SD-only	BERT	*60.57*	79.34	**56.61**	**52.84**	**53.47**	**63.52**	*78.48*	**69.89**	**56.40**	49.31
	iBERT-v1	**60.88**	**87.64**	53.50	51.37	50.99	59.49	**79.19**	55.95	52.52	50.31
	iBERT-v2	58.95	83.49	48.70	51.03	*52.57*	59.41	72.83	57.88	52.86	**54.08**
	iBERT-v3-1	58.15	80.42	50.56	50.98	50.64	59.39	*78.54*	57.04	52.60	49.37
	iBERT-v3-10	58.29	80.73	49.61	50.47	52.36	59.41	77.30	56.19	53.62	50.51
WG+SD	BERT	58.96	68.82	63.04	52.55	**51.43**	**62.76**	70.35	**71.46**	**57.00**	*52.23*
	iBERT-v1	58.96	63.75	**66.87**	**54.45**	*50.77*	61.53	**71.64**	67.92	55.91	50.64
	iBERT-v2	58.47	70.25	59.64	53.10	*50.88*	59.54	66.60	65.32	54.75	51.50
	iBERT-v3-1	58.56	69.08	61.33	*53.90*	49.92	61.43	70.01	66.15	*56.82*	**52.72**
	iBERT-v3-10	**60.96**	**79.02**	60.72	52.91	*51.19*	60.34	*71.03*	63.72	55.95	50.65
WG-only	BERT	58.20	65.59	62.03	53.55	**51.62**	**61.92**	65.05	**70.18**	*58.19*	**54.27**
	iBERT-v1	*59.29*	64.69	**67.99**	**55.80**	48.69	60.74	*67.14*	66.89	*57.77*	51.17
	iBERT-v2	57.77	66.93	62.27	54.17	47.74	58.63	62.41	65.75	56.58	49.78
	iBERT-v3-1	*59.47*	66.12	66.98	*55.68*	49.11	60.87	65.92	68.87	**58.29**	50.38
	iBERT-v3-10	**59.87**	**69.70**	65.15	*55.23*	49.40	*61.07*	**68.02**	67.73	*57.92*	50.59

Table 4: AUC (%) on PAN 11/13/14/15 authorship verification tasks for **128 vs. 512 token-size models**. Rows are grouped by training setup (**SD**, **WG**, **WG+SD**). **Bold** indicates the best mean per group-token combination. Underline highlights global best per token. *Italics* mark values reasonably close to the best per group.

Model	τ	Aggregation	STEL↑	SoC↑
v1	∞	Mean (uniform blend)	37.5	92.3
v3-10	10	Peaked blend	38.1	**92.8**
v3-1	1	Softmax blend	**38.3**	**92.8**
v2	0	Top-sense only	28.5	88.7

Table 5: Effect of aggregation sharpness τ on iBERT, trained on StyleSynth triplets (128 token size).

Classifier Debiasing via Sense Ablation. Masking specific sense dimensions (e.g., sarcasm, punctuation cues) neutralizes stylistic noise before classification. This supports robust models that focus on core semantics—offering interpretability by construction.

Latent Data Augmentation. Controlled sense edits allow creation of style-shifted variants without decoding. These augmentations improve robustness and style-invariance for downstream classifiers.

Style-Aware Personalization. Sense-level edits adapt outputs to user preferences (e.g., concise vs. elaborate tone) without retraining—enabling personalization in retrieval and generation pipelines.

RAG and Retrieval Robustness. iBERT supports query normalization, document style filtering, and attribution for retrieval decisions—addressing failure modes like tone mismatch or prompt drift in RAG systems.

Forensic and Monitoring Applications. iBERT enables tracking tone shifts, authorial style evolution, or discourse change over time—useful for moderation, attribution, or sociolinguistic analysis.

Controller for Generation. Sense vectors from iBERT can guide decoder-based models (e.g., T5, GPT) as structured, interpretable control signals—supporting style-conditioned generation without prompt engineering.

Representation Auditing. Because embeddings decompose into interpretable senses, iBERT supports bias audits, attribution analysis, and fairness enforcement via projection or adversarial masking.

Zero-Decode Semantic Transfer. Sense-level edits enable style transfer, paraphrasing, or domain adaptation directly in latent space—without requiring any generation.

A.2 Effect of Aggregation Sharpness

As described in Section 3, τ controls how sharply the model aggregates across its k=8 sense vectors—ranging from uniform averaging ($\tau \to \infty$) to top-sense-only pooling (τ=0). Intermediate values apply softmax pooling over sense magnitudes.

Sharper but soft aggregation version of v3 created by τ=10, v3-10 yields the best STEL and SoC performance, as reported in the full results (Table 1). Appendix Table 5 highlights a representative subset across τ values, to draw attention to how aggregation sharpness affects performance

STYLE	$\ell = 0$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$	$\ell = 7$	BEST
Active / Passive	0.6222	0.6222	0.5278	0.6278	0.6111	0.6444	0.5944	0.6333	5
Affective process / Perceptual process	0.6389	0.5444	0.6500	0.5889	0.6222	0.5833	0.6444	0.6278	2
Affective processes / Cognitive processes	0.6722	0.5833	0.6833	0.6111	0.6833	0.6389	0.7000	0.6722	6
All Lower Case / Proper Capitalization	0.9778	0.7611	0.9667	0.8444	0.9667	0.8778	0.9833	0.9833	7
All Upper Case / Proper Capitalization	0.9278	0.9444	0.9278	0.9333	0.9333	0.9389	0.9167	0.9222	1
Certain / Uncertain	0.6444	0.6056	0.6167	0.6444	0.6389	0.6500	0.6833	0.6944	7
Cognitive process / Perceptual process	0.6444	0.6167	0.5500	0.6500	0.6500	0.6500	0.6444	0.6667	7
Complex / Simple	0.5944	0.5444	0.6278	0.5611	0.6000	0.5556	0.6333	0.5944	6
Fluent sentence / Disfluent sentence	0.6278	0.5889	0.5111	0.6167	0.6111	0.6167	0.6444	0.6000	6
Formal / Informal	0.7833	0.7389	0.8167	0.7667	0.8167	0.7667	0.7667	0.7778	2
Long average word length / Short average word length	0.8389	0.8444	0.7944	0.8444	0.8389	0.8333	0.8278	0.8389	1
Offensive / Non-Offensive	0.9111	0.7389	0.9444	0.7611	0.9056	0.7722	0.9111	0.9000	2
Polite / Impolite	0.6389	0.4167	0.7333	0.4833	0.6667	0.4889	0.6444	0.5833	2
Positive / Negative	0.5944	0.5111	0.6944	0.5222	0.5944	0.5111	0.6056	0.5556	2
Present-focused / Future-focused	0.7000	0.6278	0.5556	0.6556	0.6833	0.6556	0.6778	0.7000	7
Present-focused / Past-focused	0.6778	0.5778	0.6333	0.5944	0.6278	0.5833	0.6944	0.6667	6
Self-focused / Audience-focused	0.8111	0.5778	0.7944	0.6444	0.7889	0.6389	0.8278	0.8111	6
Self-focused / Inclusive-focused	0.7500	0.5222	0.7556	0.5611	0.7000	0.5667	0.7333	0.7222	2
Self-focused / Third-person singular	0.6333	0.6056	0.5833	0.6167	0.6167	0.6111	0.6222	0.6444	7
Self-focused / You-focused	0.7944	0.5500	0.7778	0.6278	0.7556	0.6333	0.7889	0.7667	0
Sentence With a Few Misspelled Words / Normal Sentence	0.8889	0.8722	0.8389	0.8889	0.8944	0.8889	0.9111	0.9056	6
Text Emojis / No Emojis	0.9278	0.9333	0.9167	0.9278	0.9333	0.9333	0.9167	0.9167	1
With Emojis / No Emojis	0.9389	0.9389	0.9222	0.9389	0.9389	0.9389	0.9278	0.9389	0
With Humor / Without Humor	0.7556	0.6944	0.8000	0.7167	0.7667	0.7167	0.7778	0.7222	2
With articles / Less frequent articles	0.6889	0.6889	0.6000	0.6889	0.6833	0.6833	0.7278	0.7167	6
With common verbs / Less frequent common verbs	0.7889	0.7278	0.7611	0.7111	0.7444	0.7111	0.8111	0.7444	6
With conjunctions / Less frequent conjunctions	0.7944	0.7667	0.7944	0.7500	0.7833	0.7500	0.7833	0.7444	0
With contractions / Without contractions	0.8167	0.5778	0.7056	0.6722	0.7722	0.7000	0.8333	0.8278	6
With determiners / Less frequent determiners	0.6778	0.7222	0.6389	0.7056	0.6889	0.7000	0.7167	0.7278	7
With digits / Less frequent digits	0.8889	0.8389	0.8889	0.8556	0.8833	0.8667	0.8667	0.8778	0
With frequent punctuation / Less Frequent punctuation	0.6667	0.6167	0.7111	0.6389	0.6778	0.6389	0.7278	0.7000	6
With function words / Less frequent function words	0.7056	0.6944	0.6944	0.7056	0.7111	0.7056	0.7056	0.6944	4
With metaphor / Without metaphor	0.7278	0.6889	0.8222	0.6944	0.7667	0.7056	0.7389	0.7222	2
With nominalizations / Without nominalizations	0.8167	0.8056	0.8222	0.8111	0.8222	0.8222	0.7889	0.8000	4
With number substitution / Without number substitution	0.9278	0.9111	0.9389	0.9167	0.9222	0.9167	0.9222	0.9167	2
With personal pronouns / Less frequent pronouns	0.7389	0.7389	0.7111	0.7111	0.7167	0.7056	0.7222	0.7000	0
With prepositions / Less frequent prepositions	0.6333	0.6722	0.6111	0.6889	0.6611	0.6833	0.6667	0.6778	3
With pronouns / Less frequent pronouns	0.7611	0.7222	0.7444	0.7111	0.7278	0.7056	0.7889	0.7222	6
With sarcasm / Without sarcasm	0.7778	0.7278	0.8167	0.7278	0.7889	0.7333	0.7833	0.7278	2
With uppercase letters / Without uppercase letters	0.6556	0.5333	0.6722	0.5444	0.6611	0.5444	0.6500	0.6278	2

Table 6: [iBERT-v3-10 (128-tokens)] Probing activations for style classes across all senses ($\ell = 0$ to 7). The BEST column indicates the latent sense with the highest activation for each style. Style names in this table reflect the original labels from the StyleSynth dataset. In the main text, shorter style names are used for presentation clarity.

DATASET	SPLIT	MEDIAN TOKENS	>128 TOKENS
StyleSynth	Train	19	0.00%
	Test	10	0.00%
Wegmann	Train	24	6.78%
	Test	24	6.59%

Table 7: Token-statistics across datasets and splits.

and stability. This trend suggests that axis-selective pooling supports better alignment with stylistic dimensions. In addition to strong performance, v3-10 also exhibits consistently lower variance across runs (see Table 1), indicating more stable convergence during training. In contrast, hard top-sense selection degrades SoC, suggesting that some degree of sense blending is beneficial for generalization. These trends mirror regularization analogies: peaked pooling acts like L1, while smoother blending behaves L2-like.

A.3 Effect of Pooling Sharpness τ

To study the impact of pooling sharpness, we evaluated iBERT-Sentence (iSBERT) under four values of $\tau \in \{0, 1, 10, \infty\}$ referred to as iBERT-v1, iBERT-v2, iBERT-v3-1 and iBERT-v3-10.

We find that τ=0 (v2, top-sense pooling) yields the highest sense-level sparsity, with sharper sense specialization and more effective style edits (§5.5). τ=1 (v3, soft pooling) balances performance and interpretability. τ=10 leans towards uniform weighting (v1) across senses. $\tau \rightarrow \infty$ (v1, mean pooling) exhibits minimal sense-selectivity and lower editability. This confirms that sharper pooling enhances sense-disentanglement, critical for controllable edits and style probing.

	Variant	Frozen Phase		Unfrozen Phase		Total	
		Time	Loss / Epochs	Time	Loss / Epochs	Time	Epochs
128	iBERT	23h	1.6912 / 57	3d 6h	1.3365 / 85	4d 5h	142
	BERT	2h	1.3607 / 5	1d 2h	1.2228 / 31	1d 5h	36
512	iBERT	2d 9h	1.4418 / 46	1d 7h	1.2941 / 16	3d 16h	62
	BERT	14h	1.1876 / 6	3d 8h	1.0940 / 46	3d 22h	52

Table 8: Training runtimes, epochs, and final losses for iBERT-MLM and BERT (MLM) under 128- and 512-token limits. Epoch counts exclude early-stopping patience rounds.

A.4 Training and Implementation Details

All iBERT models use k=8 senses, 768-dim hidden states, and FlashAttention-2 (Dao, 2024) for fast attention computation. AdamW optimizer with learning rate of (2×10^{-5}), a batch size of 99 is used. The models have an early stopping with 3-epoch patience over validation loss and Flash-Attention-2 is used for faster compute. We perform both masked language modeling and sentence embedding training at input lengths of 128 and 512 tokens. MLM uses 5% of FineWeb (750B tokens), and contrastive fine-tuning is performed using anchor-positive-negative triplets from StyleSynth / StyleDistance (SD) and Wegmann (WG) datasets.

Training. All models are initialized with encoder weights from ModernBERT (Warner et al., 2025), replacing into either BERT or our proposed iBERT architecture. Due to the larger Context Sense Block in iBERT, we freeze the encoder weights in an initial warmup stage (for both baseline and iBERT to maintain parity) and allow the heads to converge. Once stabilized and converged, we unfreeze the full model and resume end-to-end training until convergence. This process is used in both MLM and contrastive phases. The details of the MLM stage are listed in Table 8.

Contrastive training is supervised with either InfoNCE or triplet loss, depending on the dataset. Triplets are grouped as (anchor, pos, neg) and processed using sentence encoders that implement SBERT-style wrappers over both BERT and iBERT backbones. Our training code supports sense ablation via `-sense_gain`, and different SBERT pooling strategies (`v1`, `v2`, `v3-1`, `v3-10`).

Training is run on 4×NVIDIA A100 (80GB) GPUs using mixed-precision (`torch.bfloat16`). We use AdamW with a learning rate of 2×10^{-5}, batch size 99, and early stopping with 3-epoch patience on validation loss. All training stages are fully configurable via CLI or YAML using a shared interface, and can resume from checkpoints.

Software. Our implementation stack includes PyTorch (2.7.0), Transformers (v4.51.3), Sentence-Transformers (5.0.0), FlashAttention-2, Datasets (3.6.0), and Accelerate (1.7.0). Tokenization is handled via HuggingFace Tokenizers (0.21.1), and logging is supported through Terminal, JSONL, TensorBoard, and Weights-and-Biases (W&B).

Data Statistics For FineWeb, we split the 5% slice into a 90:10 train:dev split, without needing a test set, given it is used in an MLM. For the sentence-embedding training stages, we reuse the predefined train/test split provided by the dataset authors, and use 90:10 split of original train dataset for training and validation correspondingly.

Datasets, Licensing, and Safety Our study is based on publicly available datasets, which have been used in prior research on author attribution and stylistic analysis. These datasets originate from online platforms and may inherently contain offensive or personally identifiable content. However, we do not apply any additional filtering beyond what has been done in previous studies. To ensure ethical usage, we follow the guidelines set by the original sources that published these datasets and acknowledge any potential biases or content-related concerns that may arise.

Additionally, we discuss the licensing terms associated with the datasets used in our experiments. The datasets are provided under open-access licenses, allowing their use for academic research. We ensure compliance with these licenses and properly attribute the sources in our work. Detailed information on dataset licensing is included in the final version of the paper.

A.4.1 Sentence Embedding Training and Replications

For sentence-level contrastive training, we train a total of five distinct sentence encoder models (per data configuration): a baseline SBERT built on top of BERT, and four sentence encoders built on iBERT-MLM checkpoints using different pooling frameworks. These correspond to mean-pooling (v1, $\tau{=}\infty$), top-sense (v2, $\tau{=}0$), softmax-blend (v3-1, $\tau{=}1$), and peaked-blend (v3-10, $\tau{=}10$). Each is treated as a separate model, not as a parameter sweep. All models are trained at two input lengths (128 and 512) and evaluated on StyleSynth and Wegmann datasets.

To account for training variability, we perform 5 independent training replications for each sentence encoder: (S)BERT and all iBERT variants, at both token lengths (128 and 512). Reported results reflect mean and standard deviation across the runs.

$$\begin{aligned}
\text{Training runs} &= 5 \quad \left\{\begin{array}{l}\text{model types:}\\ \text{(S)BERT, iBERT-v1, v2}\\ \text{iBERT-v3-1, v3-10}\end{array}\right.\\
&\times 3 \quad \left\{\begin{array}{l}\text{datasets:}\\ \text{SD, WG+SD, WG}\end{array}\right.\\
&\times 2 \quad \left\{\begin{array}{l}\text{token lengths:}\\ \text{128, 512}\end{array}\right.\\
&\times 5 \quad \left\{\begin{array}{l}\text{replications per}\\ \text{configuration}\end{array}\right.\\
&= \mathbf{150}
\end{aligned}$$

A.5 Style Contrast Visualization

We visualize the separation of contrastive style pairs (e.g., number substitution, personal pronoun frequency) using t-SNE projections. For each example, we compare (S)BERT embeddings (left) and iBERT-v2 embeddings (right). In both cases, sentence embeddings were obtained after training on StyleSynth.

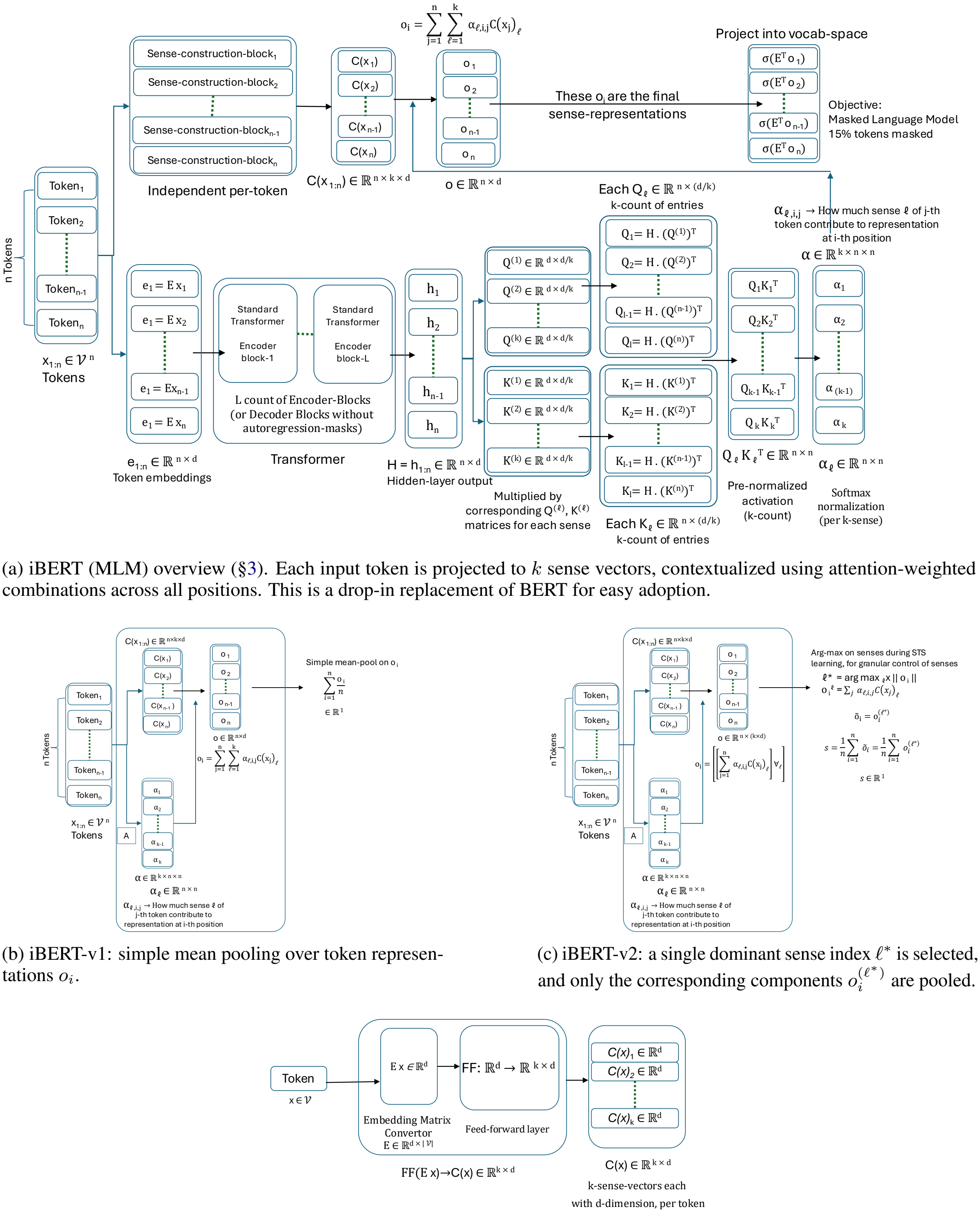
Figure 4 shows that iBERT exhibits stronger axis-wise separation between stylistic variants, with clearer clusters and lower embedding overlap, consistent with our claims about modularity and disentanglement.

A.6 Global t-SNE for Embedding Clustering

We visualize sentence embeddings across all 40 style contrasts using t-SNE, comparing BERT and iBERT-v3 in Figure 6. The iBERT model exhibits sharper geometric structure, with more linear and segregated style clusters. This supports our hypothesis that modular pooling (v3) induces interpretable and axis-aligned representations, in contrast to the entangled embedding space of BERT.

A.7 Model Schematic and Embeddings

We conclude the appendix with Figures 5 and 6, which illustrate the full model architecture and the global embedding structure.



(a) iBERT (MLM) overview (§3). Each input token is projected to k sense vectors, contextualized using attention-weighted combinations across all positions. This is a drop-in replacement of BERT for easy adoption.

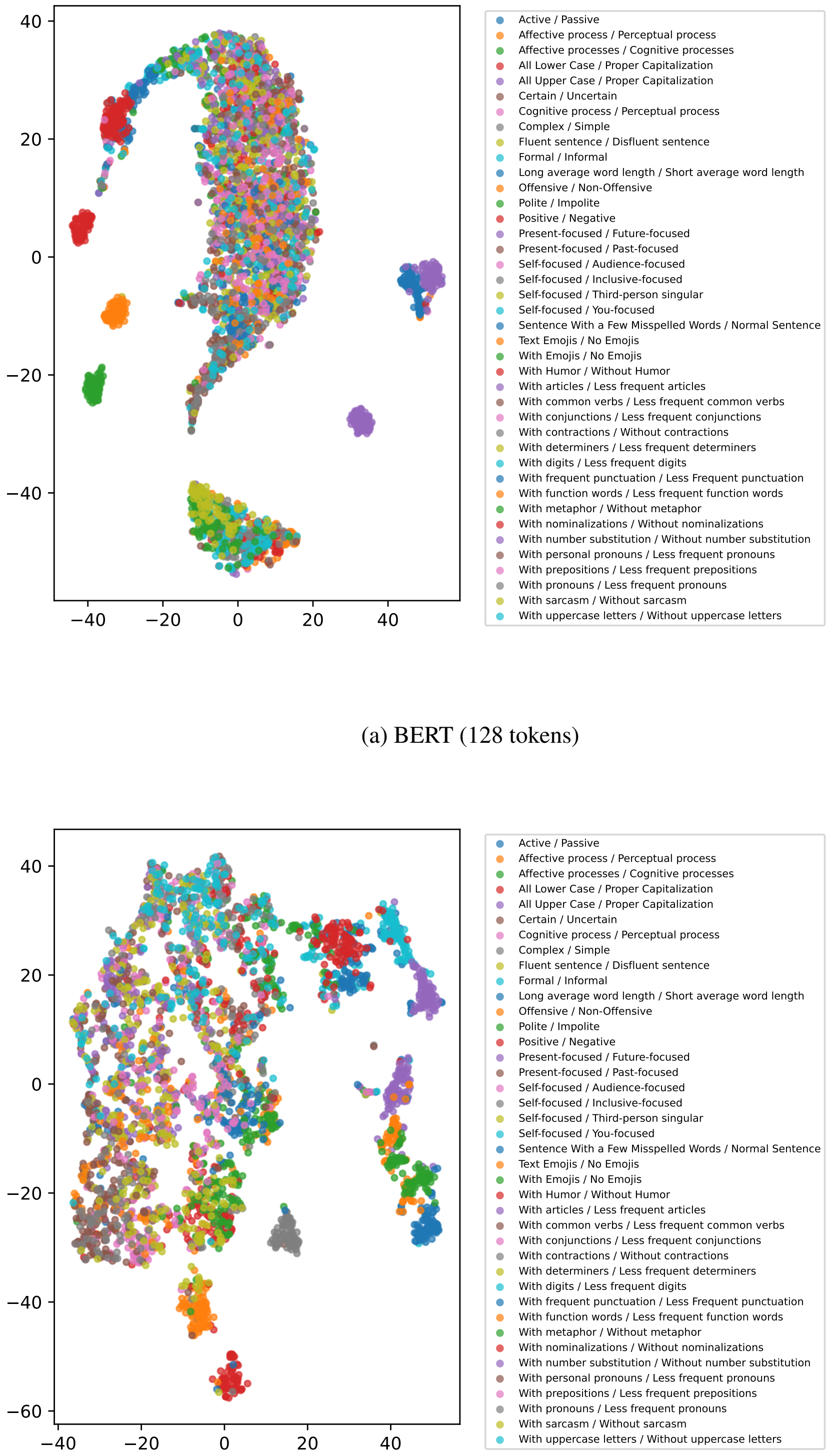
C(x1:n) ∈ ℝ n×k×d
o ∈ ℝ n×d
oi = Σ_{j=1}^{n} Σ_{ℓ=1}^{k} αℓ,i,j C(xj)ℓ
Simple mean-pool on oi
Σ_{i=1}^{n} oi / n
∈ ℝ 1
A
α ∈ ℝ k × n × n
αℓ ∈ ℝ n × n
αℓ,i,j → How much sense ℓ of j-th token contribute to representation at i-th position
o ∈ ℝ n × (k×d)
oi = [[Σ_{j=1}^{n} αℓ,i,j C(xj)ℓ] ∀ℓ]
Arg-max on senses during STS learning, for granular control of senses
ℓ* = arg max ℓ x || oi ||
oiℓ = Σj αℓ,i,j C(xj)ℓ
ōi = oi(ℓ*)
s = 1/n Σ_{i=1}^{n} ōi = 1/n Σ_{i=1}^{n} oi(ℓ*)
s ∈ ℝ 1

(b) iBERT-v1: simple mean pooling over token representations o_i.

(c) iBERT-v2: a single dominant sense index ℓ^* is selected, and only the corresponding components $o_i^{(\ell^*)}$ are pooled.

(d) Sense Construction Block: each token is mapped via a feedforward layer to k sense vectors of dimension d.

Figure 5: (a) shows the detailed technical iBERT architecture; (b–c) illustrate two pooling strategies used in sentence encoding (v1 and v2). The softmax-blend pooling variant (v3) lies between these and is described in Section subsection 3.2. (d) details the sense construction block.



(a) BERT (128 tokens)

(b) iBERT-v3-10 (128 tokens)

Figure 6: t-SNE projections of sentence embeddings for all 40 style contrast pairs. Each point represents a sentence, color-coded by its style label. BERT shows more dispersed and overlapping clusters, while iBERT-v3-10 shows tighter, aligned groupings — indicating improved style modularity and disentanglement. Both these language models were first trained on FINEWEB and then trained on StyleSynth triplets.