Bridging Compositional and Distributional Semantics: A Survey on Latent Semantic Geometry via AutoEncoder

Anonymous ACL submission

Abstract

Integrating compositional and symbolic properties into distributional semantic spaces can enhance the interpretability, controllability, composition, and generalisation capabilities of Transformer Language Models. In this survey, we offer a novel perspective on latent space geometry through the lens of compositional semantics, a direction we refer to as *semantic representation learning*. This direction enables a bridge between symbolic and distributional semantics, thus improving semantic and geometrical interpretability and localised control. We review and compare three mainstream autoencoder architectures: Variational AutoEncoder (VAE), Vector Quantised VAE (VQ-VAE), and Sparse AutoEncoder (SAE), and examine the distinctive latent geometries they induce in relation to semantic structure and interpretability.

1 Introduction

Over the past decade, language models have progressed from n-gram methods and RNNs to large Transformer-based architectures capable of human-level language understanding and generation (Vaswani et al., 2017; Devlin et al., 2018; Radford et al., 2019). However, this remarkable progress has come at a cost: we understand little about the internal semantic mechanisms these models use. They operate as "black boxes", where the representations they learn are difficult to interpret and control. Specifically, ***Interpretability:*** Neural networks discover useful features automatically, and their latent spaces can capture the underlying features in data in terms of geometry (Bengio et al., 2013; Alshammari et al., 2025). For Transformer LMs (TLMs), however, this latent space is high-dimensional, complex, with generative features for natural language poorly defined and geometrically entangled (Scherlis et al., 2022), thereby lacking *semantic and geometrical* interpretation. ***Controllability:*** Without a disentangled latent space, *fine-grained, quasi-symbolic, localised* generative control is difficult (Yu et al., 2026).

A promising path toward interpretability and control lies in understanding and shaping the latent space geometry. An interpretable latent geometry exhibits meaningful clustering and separation aligned with underlying generative features, enabling localised control via targeted manipulations. The autoencoder architectures, such as VAE (Kingma and Welling, 2013), VQ-VAE (Van Den Oord et al., 2017), and SAE (Shu et al., 2025), offers a compelling framework for this. Their encoder–bottleneck–decoder structure encode inputs into a latent space, encouraging the capture of salient generative factors. When combined with an autoregressive TLM, these latent variables can be explicitly manipulated to guide generation. Crucially, we can intentionally shape this geometry for greater interpretability and controllability. E.g., latent generative factors can be potentially aligned with human-interpretable features, such as sentiment, semantic role, syntax, and reasoning behaviours, among others.

Overall, this survey focuses on *semantic representation learning* to render language representations or models more semantically and geometrically interpretable, and to enable localised, quasi-symbolic, compositional control through deliberate shaping of their latent space geometry. Our objective is to move toward TLMs whose internal semantic representations can be systematically interpreted, precisely shaped, and reliably directed. This type of representation can provide the foundation to shorten the gap between deep latent semantics and symbolic, linguistic representations, integrating the flexibility of distributional-neural models with the properties of linguistically grounded representations, facilitating both interpretability and generative control.

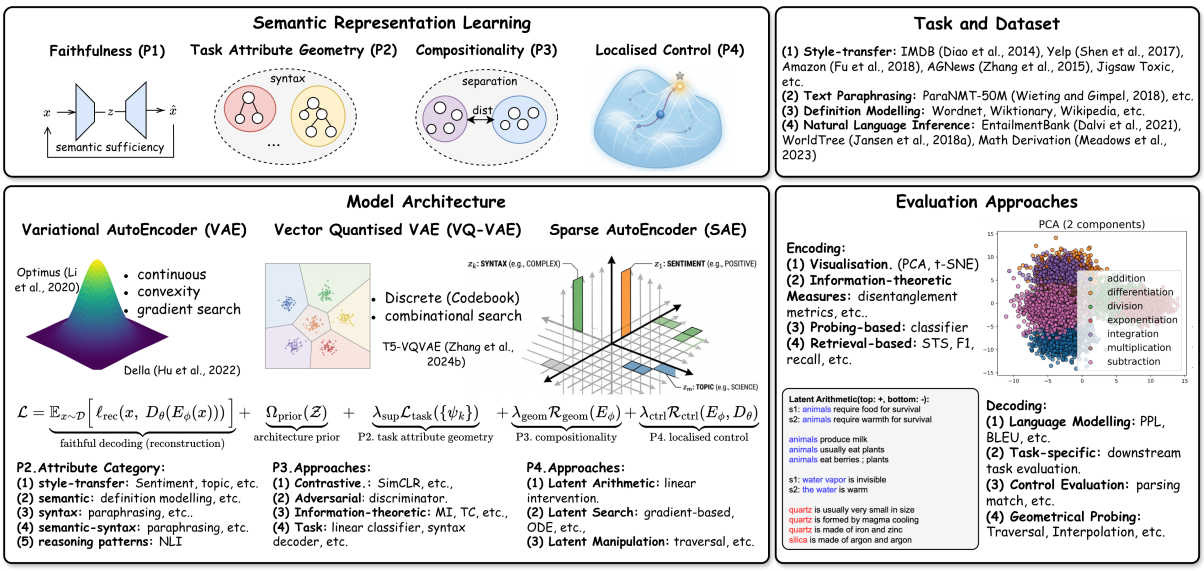

Figure 1: This survey focuses on the semantic and geometrical interpretability and fine-grained, localised, quasi-symbolic, compositional generative control of Transformer Language Models (TLMs) by shaping the latent space geometry within the AutoEncoder framework: VAE, VQVAE, and SAE.

2 Semantic Representation Learning

In this section, we introduce a unified perspective on semantic representation learning, grounded in the view of linguistic semantics as geometric structures in latent space. Rather than proposing a strict taxonomy, we synthesise insights from prior literature to identify four core properties that characterise desirable semantic representations.

Formally, let $\mathcal{X}$ be a space of linguistic expressions over a vocabulary $\mathcal{V}$. Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ denote a set of *semantic factors* (e.g., topic, sentiment, tense, semantic roles, reasoning rules), where each f_k takes values in a domain $\mathcal{A}_k$ (finite or continuous). A *semantic representation* consists of an encoder-decoder pair $E_\phi : \mathcal{X} \to \mathcal{Z}$, $D_\theta : \mathcal{Z} \to \Delta(\mathcal{X})$, with latent space $\mathcal{Z} \subseteq \mathbb{R}^d$ (continuous, discrete, or mixed), such that the tuple $(E_\phi, D_\theta, \mathcal{Z})$ satisfies the following properties:

P1. Faithfulness (semantic sufficiency). There exist readout maps $\{\psi_k : \mathcal{Z} \to \mathcal{A}_k\}_{k=1}^K$ (preferably linear or low-complexity[1]) for which $\psi_k(E_\phi(x)) \approx f_k(x)$ for each k, and decoding preserves semantics in expectation:

$$\mathbb{E}_{\tilde{x}\sim D_\theta(E_\phi(x))}[\mathbf{1}\{f_k(\tilde{x}) = f_k(x)\}] \geq 1 - \epsilon_{\text{rec}}, \quad \forall k.$$

This property serves as the foundational requirement, ensuring that semantic factors are sufficiently encoded and preserved, upon which higher-level properties (P2–P4) can be meaningfully defined.

P2. Task Attribute Geometry. For a given task, the latent space should be organised into task-relevant attribute spaces, where each semantic factor corresponds to a structured and interpretable subspace. Specifically, there is a decomposition of $\mathcal{Z}$ into factor-specific subspaces (e.g., sentiment-residual): $\mathcal{Z} = \bigoplus_{k=1}^K \mathcal{U}_k \oplus \mathcal{U}_{\text{res}}$, such that variation in f_k is confined to $\mathcal{U}_k$ (up to tolerance ϵ_{leak}). Writing $z = z_\parallel^{(k)} + z_\perp^{(k)}$ with $z_\parallel^{(k)} \in \mathcal{U}_k$ (projection of z on $\mathcal{U}_k$) and $z_\perp^{(k)} \in \bigoplus_{j\neq k} \mathcal{U}_j \oplus \mathcal{U}_{\text{res}}$, the readout ψ_k depends only on $z_\parallel^{(k)}$. This property yields meaningful directions, distances, and clusters aligned to semantic features.

P3. Compositionality. The latent representation supports compositional structure over semantic attributes, such that factor-wise variations can be consistently combined and manipulated. This property is typically facilitated by learning strategies that promote factor-wise separability. Formally, there exists a latent operation $\oplus$ (such as addition) such that for any factor assignments a, a': $E_\phi\big(\text{compose}(x; f_k \leftarrow a \odot a')\big) \approx E_\phi\big(\text{compose}(x; f_k \leftarrow a)\big) \oplus \Delta_k(a \to a')$ with "reusable" composition vectors $\Delta_k(a \to a') \in \mathcal{U}_k$. For instance, learning sentiment steering vector from positive to negative by calculating the difference between two sets of instances, no matter how residual semantics varied (Mercatali and Freitas,

[1]For instance, we can apply a "linear" classifier ψ_k to identify factor k, where higher classification performance indicates that this factor can be encoded in the latent space (Zhang et al., 2025b).

2021; Shen et al., 2020).

P4. Localised Control. The model enables targeted manipulation of specific semantic attributes during generation. For any k and target value a', define an intervention operator $\mathcal{I}_k(z, a') := z_\perp^{(k)} + \Gamma_k(a'), \Gamma_k : \mathcal{A}_k \to \mathcal{U}_k$, such that decoding $x' \sim D_\theta(\mathcal{I}_k(E_\phi(x), a'))$ achieves $\Pr[f_k(x') = a'] \geq 1 - \epsilon_{\text{set}}, \text{dist}(f_j(x'),\ f_j(x)) \leq \epsilon_{\text{hold}}\ \forall j \neq k$, That is, the success rate of factor intervention is higher than a probability, while the distance of unchanged information before and after intervention should less than a small value ϵ_{hold}.

Building on the four properties outlined above, semantic representation learning aims to shape the latent space to enhance both semantic and geometric interpretability, as well as to enable fine-grained, localised control over generation.

3 Architecture: A Latent Geometry Perspective

This section presents a brief overview of the principal autoencoder architectures.

(1) Transformer VAEs. VAEs (Kingma and Welling, 2013) learn continuous latent manifolds endowed with a known geometric prior, typically an isotropic Gaussian. The encoder defines a mapping from discrete text inputs onto a smooth, low-dimensional Riemannian manifold. The decoder then learns a reverse mapping, generating text conditioned on coordinates sampled from this manifold. Through the KL regularisation term, the Gaussian prior induces an approximately Euclidean latent geometry, promoting global smoothness and convexity. These properties facilitate gradient-based optimisation (Bonnet and Macfarlane, 2024) and enable latent arithmetic (Zhang et al., 2024e) within the representation space. Modern Transformer VAEs (e.g., Optimus (Li et al., 2020), Della (Hu et al., 2022), AdaVAE (Tu et al., 2022), LlamaVAE (Zhang et al., 2023), LangVAE (Carvalho et al., 2025)) combine this latent geometry with pretrained encoders and decoders. The resulting latent space provides an abstract representation of natural language, enabling more global control over generation. While non-Euclidean latent geometries (e.g., hyperbolic (Cho et al., 2023) or hyperspherical (Davidson et al., 2018)) have been explored, most models retain Gaussian priors due to its mathematical tractability, ease of optimisation.

(2) Vector Quantised VAEs. Vector Quantised VAEs (VQ-VAEs) (Van Den Oord et al., 2017) replace the continuous latent manifold with a discrete latent geometry. The model learns a finite set of prototype vectors (the codebook), which can be interpreted as discrete points partitioning the representation space. The encoder performs a nearest-neighbour projection, mapping each input to its closest codebook vector under a chosen distance metric (typically Euclidean), or alternatively employs the Gumbel–Softmax relaxation to approximate discrete latent distributions in a differentiable manner. This induces a piecewise-constant latent space, where the geometry is segmented into Voronoi regions around codebook entries. Owing to this discrete structure, the latent space can be explored using combinatorial search strategies, such as greedy or beam search, to probe and control generation (Hosking et al., 2022).

(3) Sparse AutoEncoder. In contrast to VAEs and VQ-VAEs, which primarily focus on learning compact latent representations, Sparse Autoencoders (SAEs) map inputs into higher-dimensional latent spaces, where representations are expressed as sparse combinations of basis feature vectors. For each input, only a small subset of these basis vectors is activated, transforming entangled representations into more disentangled and interpretable components. This design mitigates the issue of polysemanticity in large language models (Elhage et al., 2022), where individual neurons or dimensions encode multiple unrelated features.

Optimisation. Under specific optimisation, all architecture has the potential to learn the latent representation satisfied the four properties. Formally, given data $\mathcal{D} = \{x_i\}$, we minimise

$$\begin{aligned}\mathcal{L} = &\underbrace{\mathbb{E}_{x\sim\mathcal{D}}\big[\,\ell_{\text{rec}}(x,\ D_\theta(E_\phi(x)))\,\big]}_{\text{faithful decoding (reconstruction)}} \\ &+ \underbrace{\Omega_{\text{prior}}(\mathcal{Z})}_{\text{architecture prior}} + \underbrace{\lambda_{\text{sup}}\mathcal{L}_{\text{task}}(\{\psi_k\})}_{\text{P2. task attribute geometry}} \\ &+ \underbrace{\lambda_{\text{geom}}\mathcal{R}_{\text{geom}}(E_\phi)}_{\text{P3. compositionality}} + \underbrace{\lambda_{\text{ctrl}}\mathcal{R}_{\text{ctrl}}(E_\phi, D_\theta)}_{\text{P4. localised control}}\end{aligned} \tag{1}$$

Here Ω_{prior} instantiates the architectural choice (e.g., dual-encoder (Zhang et al., 2024e)); $\mathcal{L}_{\text{task}}$ is a supervised objective to ensure that specific semantic attributes are explicitly encoded in the latent representation; $\mathcal{R}_{\text{geom}}$ is geometrical constraint to promote subspace separation (e.g., contrastive learning (Zhang et al., 2024a), KL for VAE, sparsity for SAE); and $\mathcal{R}_{\text{ctrl}}$ enforces that $\mathcal{I}_k$ changes

only factor k (e.g., gradient searching (Bonnet and Macfarlane, 2024)).

Summary and Limitations. VAEs, VQ-VAEs, and SAEs exhibit distinct trade-offs in latent geometry. VAEs provide smooth, continuous spaces that can learn abstract-level representation and support interpolation and gradient-based optimisation. VQ-VAEs introduce discrete, clustered latent structures that improve interpretability, yet restrict compositionality and make optimisation less flexible. SAEs promote sparse, axis-aligned features that enable strong interpretability and localised control, though their linear structure may limit modelling of complex semantics. Overall, these approaches reflect a fundamental trade-off between continuity, discreteness, and interpretability in latent semantic geometry. SAE is unsupervised, which may be mis-aligned with task-specific feature.

4 Semantic Representation Properties

This section outlines the major architectural trends for learning semantic representation spaces and their integration with language models.

4.1 Task Attribute Geometry *(P2)*

Based on their distinct geometric properties with respect to task requirements, we categorise latent spaces into the following groups:

(1) Style-transfer Attribute Space. Style-transfer attribute spaces refer to latent representations in which non-structural linguistic attributes (e.g., sentiment, topic) are disentangled from core semantic content. Within VAE-based frameworks, most studies focus on controlled generation for short texts, particularly sentiment transfer, where models aim to modify specific attributes (e.g., from positive to negative) while preserving the underlying meaning (Li et al., 2022; Liu et al., 2020). This paradigm has been extended to multi-attribute control, enabling the simultaneous manipulation of multiple stylistic dimensions (Gu et al., 2023, 2022; Hu and Li, 2021).

Within VQ-VAE-based architectures, research has primarily focused on topic-level control. For example, Topic-VQVAE (Yoo and Choi, 2024) extends this framework to document-level generation by learning discrete topic codes. In parallel, SAEs enable feature-level interventions in LLMs, supporting the steering of attributes such as sentiment, truthfulness, and political polarity (He et al., 2025).

Building upon these attribute-based representations, we introduce four types of latent space geometries specifically designed to encode structural semantic features, such as syntax, semantics and reasoning logics.

(2) Syntax Attribute Space. We define syntax attribute spaces as latent representations where syntactic structure is encoded as independent generative factors, disentangled from lexical semantics. Most prior work focuses on modelling grammatical structures (e.g., subject–verb–object relations) to enable syntactically controlled generation, particularly in tasks such as style transfer and text paraphrasing, primarily within VAE-based frameworks (Mercatali and Freitas, 2021; Felhi et al., 2022b). In addition, HRQ-VAE (Hosking et al., 2022) models a hierarchical syntactic sketch space via residual quantisation, decomposing an input encoding into a sum of embeddings from multiple codebooks. Each codebook refines the residual of the previous one, enabling a coarse-to-fine representation of syntactic information.

(3) Syntax-Semantic Attribute Space. Syntax-semantic attribute space refers to a class of latent representations in which both syntactic structure and lexical semantics are explicitly encoded in the latent space. This representation enables independent control over syntactic form and semantic content, facilitating structure-induced generation and compositional generalisation. Several studies have explored methods for learning such disentangled latent spaces, with most work focusing on text paraphrasing within VAEs (Zhang et al., 2019; Hosking and Lapata, 2021). In addition, Zhang et al. (2024e) show that separating syntax and semantics via a graph-language dual-encoder architecture can mitigate the information bottleneck imposed by language encoders (e.g., BERT).

(4) Semantic Attribute Space. semantic attribute spaces refer to latent representations structured around semantic role structure. Most work focuses on definition modelling and semantic role modelling, where latent variables capture role-specific meaning contributions (Carvalho et al., 2023). Recent approaches formalise this geometry using Argument Structure Theory, representing semantic role-word interactions as structured regions (e.g., convex cones) in latent space, enabling fine-grained control and interpretability (Zhang et al., 2024b). Discrete models (e.g., T5-VQVAE (Zhang et al., 2024c)) and SAEs (Jing et al., 2025) further support controllable generation by introducing semantically interpretable latent features.

(5) Reasoning Attribute Space. Reasoning attribute spaces refers to the latent representations that encode inferential structures or reasoning pat-

Literature	Attributes	Architecture	Cont Opt.	Adv Opt.	Information Opt.	Task Opt.	Arithmetic	Search	Manipulate
Hu et al. (2017)	sentiment	VAE	-	✓	-	✓ (cls)	-	-	✓
Fang et al. (2019)	sentiment	VAE	-	✓	✓ (posterior)	✓ (cls)	✓	-	✓
John et al. (2019)	sentiment	VAE	-	✓	-	✓ (BoW)	✓	-	-
Mai et al. (2020)	sentiment	AE	-	✓	-	✓ (cls)	-	✓	-
Shen et al. (2020)	sentiment	VAE	-	✓	-	-	✓	-	-
Rimsky et al. (2024)	sentiment	SAE	-	-	-	-	✓	-	-
Shi et al. (2021)	sentiment	AE (graph)	-	-	-	-	-	-	✓
Sha and Lukasiewicz (2024)	sentiment	VAE	✓	-	-	-	-	-	✓
Li et al. (2022)	sentiment	VAE	-	-	✓ (prior)	-	-	-	✓
Xu et al. (2021)	sentiment; code-switch;...	VAE	✓	-	-	-	✓	-	-
Hu and Li (2021)	sentiment, category	VAE	-	-	-	✓ (cls)	-	-	✓
Huang et al. (2020)	sentiment; topic	AE	-	✓	-	-	-	-	✓
Duan et al. (2020)	sentiment; topic; length	VAE	-	✓	✓ (prior)	-	-	-	✓
Gu et al. (2022)	sentiment, topic, detox	VAE	-	-	✓ (Euler distance)	✓ (cls)	-	✓	-
Liu et al. (2020)	sentiment; topic	VAE	-	-	-	✓ (BoW, cls)	-	✓	-
Gu et al. (2023)	sentiment, topic, detox	AE+flow	-	-	✓ (prior)	-	✓	✓	-
Fang et al. (2022)	sentiment, topic, tense	VAE	-	-	✓ (prior)	✓ (cls)	-	-	✓
Gupta and Zhang (2021)	topic	VQ-VAE	-	-	-	✓	-	-	
Hu et al. (2024)	topic	VQ-VAE	-	-	-	-	-	-	✓
Yoo and Choi (2024)	topic	VQ-VAE	-	-	-	-	-	-	✓
Liu et al. (2023)	negation, tense, formality	VAE	-	-	-	✓ (cls)	-	✓	-
Vasilakes et al. (2022)	negation, uncertainty	VAE	-	-	✓ (MI)	✓ (cls)	-	-	✓
He et al. (2025)	truthfulness,sentiment,...	SAE	✓	-	-	✓ (cls)	✓	-	-
Zhang et al. (2024a)	truthfulness	AE* (dual)	✓	-	-	-	✓	-	✓
Li et al. (2023)	truthfulness	SAE	-	-	-	-	✓	-	-
Deng et al. (2025)	code-switch	SAE	-	-	-	-	✓	-	-
Minder et al. (2025)	chat-specific features, ...	SAE	-	-	-	-	✓	-	-
Zhao et al. (2025)	context-memory	SAE	✓	-	-	-	✓	-	-
Yoshioka et al. (2022)	fluent spoken	VAE (dual)	-	-	✓ (posterior)	-	-	-	✓
Mercatali and Freitas (2021)	syntactic factors	VAE (discrete)	-	-	✓ (MI, TC)	-	✓	-	-
Felhi et al. (2022b)	syntactic roles	VAE		-	-	-	-	-	✓
Hosking et al. (2022)	hierarchical syntax	VQ-VAE	-	-	-	✓(cls)	-	✓	-
Huang and Chang (2021)	semantic-syntax	AE (dual)	-	-	-	-	-	-	✓
Zhang et al. (2019)	semantic-syntax	VAE (dual)	-	-	✓ (prior)	✓ (treecoder)	-	-	
Bao et al. (2019)	semantic-syntax	VAE	-	✓	-	✓ (cls,treecoder)	-	-	✓
Chen et al. (2019)	semantic-syntax	VAE (dual)	✓	-	✓(prior,posterior)	-	-	-	-
Huang et al. (2021)	semantic-syntax	Transformer (dual)	-	✓	-	-	-	-	✓
Felhi et al. (2022a)	semantic-syntax	VAE	-	-	-	-	-	-	✓
Zhang et al. (2024e)	semantic-syntax	VAE (dual)	-	-	-	✓ (cls)	-	-	✓
Hosking and Lapata (2021)	semantic-syntax	VQ-VAE	-	-	-	-	-	-	✓
Carvalho et al. (2023)	semantic role	VAE	-	-	✓ (prior)	-	✓	-	-
Zhang et al. (2024b)	semantic role	AE+flow	-	-	✓ (prior)	-	✓	-	-
Zhang et al. (2025a)	semantic role	VAE	-	-	✓ (prior)	-	✓	-	✓
Roy and Grangier (2019)	semantics (concept)	VQ-VAE*	-	-	-	-	-	-	✓
Zhang et al. (2024c)	semantics (concept)	VQ-VAE	-	-	-	-	✓	-	✓
Garg et al. (2025)	semantics (concept)	VQ-VAE*	-	-	-	-	-	-	✓
Jing et al. (2025)	semantics	SAE	-	-	-	-	✓	-	-
Zhang et al. (2024d)	reasoning patterns	Transformer	-	-	-	-	-	-	✓
Zhang et al. (2025b)	reasoning patterns	VAE	-	-	✓ (prior)	✓ (cls)	-	-	✓
Sanchez et al. (2023)	reasoning patterns	VAE	-	-	✓ (prior)	-	-	-	✓
Venhoff et al. (2025)	reasoning patterns	SAE	-	-	-	-	✓	✓	-

Table 1: Existing work on **P2. task attribute geometry** and **P3. compositionality** optimisation spans contrastive (Cont), adversarial (Adv), information-theoretic (Information), and task-specific (Task) paradigms, while **P4. localised control** is achieved via latent manipulation, arithmetic, or searching. The summarisation of dataset and evaluation metric used in those studies is provided in Table 3.

terns as controllable factors. Recent studies focus on tasks such as rule-based natural language inference (NLI), where latent spaces capture logical relationships (Zhang et al., 2024d, 2025b). Related work also encodes symbolic graphs or reasoning trajectories (e.g., via random walks or sparse features), allowing models to reproduce structured reasoning behaviours during generation (Sanchez et al., 2023; Venhoff et al., 2025). For instance, Venhoff et al. (2025) use SAE to learn low-dimensional latent reasoning behaviours in LLMs, such as *problem restatement*, *subgoal setting*, and *knowledge retrieval*. Applying these feature vectors during generation can guide base models to reproduce structured reasoning chains.

Summary and Limitations. Overall, while these approaches mark a shift from unstructured latent representations toward geometrically organised spaces grounded in linguistic structures, structured and discourse-level representations remain underexplored. Moreover, compared to VAEs and VQ-VAEs, SAEs have received relatively limited attention in modeling structural representations. By encoding generative factors such as syntax, semantics, and reasoning, modern approaches enable more controllable, interpretable, and compositional language generation.

4.2 Compositionality *(P3)*

In this section, we review the principal optimisation strategies used to induce compositional structure in latent representations. Specifically, these methods aim to enforce factor-wise separability, thereby enabling consistent manipulation and combination of

semantic attributes in the latent space.

(1) Contrastive Learning. To encourage the latent space to capture contrastive features (e.g., positive-negative in style transfer), contrastive learning (Sohn, 2016) is widely adopted as an effective objective. It operates on latent representations by promoting similarity between samples sharing the same underlying factor while separating those corresponding to different factors.

$$\mathcal{L}(z, Z^+, Z^-) = -\log \frac{\sum_{z' \in Z^+} \exp\left(\mathrm{sim}(z, z')/\tau\right)}{\sum_{z' \in Z^+ \cup Z^-} \exp\left(\mathrm{sim}(z, z')/\tau\right)}$$

Here, z denotes the anchor representation, Z^+ and Z^- represent the sets of positive and negative samples, respectively. The function $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature parameter controlling the concentration of the distribution.

(2) Adversarial Learning. Adversarial learning is employed to enforce invariance to non-target factors by removing undesired information from the latent representation. Typically, a discriminator is trained to predict non-target attributes from the latent code, while the encoder is optimised to prevent such predictions:

$$\min_E \max_D \mathcal{L}_{\text{adv}}(z, y) = \mathbb{E}_z\big[\log D(y \mid z)\big],$$

where $z = E(x)$ and y denotes the non-target factor. This min–max objective encourages the latent space to be invariant to y, thereby promoting factor-wise disentanglement.

However, adversarial objectives may introduce instability during training and can inadvertently remove information that is correlated with the target factor, leading to a trade-off between invariance and reconstruction fidelity.

(3) Information-theoretic Optimisation. In contrast to contrastive and adversarial objectives, which operate at the sample level, information-theoretic optimisation regulates latent representations by explicitly controlling statistical dependencies at the distribution level. Representative approaches include: (i) Mutual Information (MI); (ii) Total Correlation (TC); and (iii) structured priors and posteriors, such as trainable priors (Zhang et al., 2025a) (i.e., Conditional VAE), that go beyond fixed prior assumptions. For instance, Zhang et al. (2025a) incorporate semantic role structure into the prior distribution of VAE using a BERT encoder. Under this formulation, the joint distribution is expressed as: $P_\theta(x, r, z) = \underbrace{P_\theta(x|z, r)}_{likelihood} \times \underbrace{P_\theta(z|r)}_{prior} \times P(r)$ where the prior is conditioned on structured semantic role r. This design enables the latent space to encode structural information in a geometrically meaningful manner, thereby enhancing representational expressiveness.

Nevertheless, information-theoretic optimisation typically incurs higher computational cost, as it requires estimating global statistical dependencies between variables, often through variational approximations or auxiliary networks.

(4) Task Optimisation. Task-specific objectives are widely employed to induce latent separation by explicitly supervising representations with respect to target factors. Typical approaches include linear classifiers (He et al., 2025), syntax-aware decoders (Zhang et al., 2019), Bag-of-Words predictors (Bao et al., 2019), which impose structural or semantic constraints on the latent space.

Nevertheless, this approach relies on the assumption that such information is already encoded in the latent space in a form that can be separated by a simple predictor. Consequently, deeper linguistic structures may not be adequately represented.

Summary and Limitations. Overall, under different optimisation strategies, the latent space can represent specific geometry to target attributes. The structure of the latent space under different training objectives can be evaluated by visualising feature clusters and their separability.

4.3 Localised Control *(P4)*

Having reviewed the principal techniques for learning latent representations, we now turn to the generation stage, focusing on methods for achieving localised control through the manipulation of latent representations. We categorise these approaches into three main classes.

(1) Latent Manipulation. Latent manipulation refers to modifying latent representations, encompassing both continuous traversal (Higgins et al., 2017; Kim and Mnih, 2018) and discrete interventions on semantically grounded variables (e.g., sentiment shifts). For instance, we can perform a random walk, such as *Ornstein-Uhlenbeck*, across the VAE latent space. If the latent space exhibits well-structured geometrical properties, the traversed examples are expected to display consistent patterns during generation.

(2) Latent Arithmetic. Latent arithmetic provides a lightweight mechanism for localised control

by exploiting linear structure in latent space. The underlying assumption is that semantic attributes are geometrically separable and approximately linear, allowing attribute manipulation via vector operations (Bengio et al., 2013).

Concretely, given an input x with latent representation $z = E(x)$, attribute control can be achieved by decoding a perturbed code $D(z \pm v)$, where v is a steering vector corresponding to the desired semantic shift (Shen et al., 2020). This paradigm has been extended to LLMs by intervening on residual activations (Turner et al., 2023), with methods such as CAA (Rimsky et al., 2024), RePe (Zou et al., 2023), Top PC (Im and Li, 2025), and ITI (Li et al., 2023), and SAE-based steering approaches (Zhao et al., 2025; He et al., 2025).

However, this approach relies on the assumption that the latent space is approximately linear and convex, which limits its applicability.

(3) Latent Searching. Latent searching refers to obtaining desired latent representations by iteratively optimising or navigating the latent space under a predefined objective.

For example, Bonnet and Macfarlane (2024) map input-output exemplars $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ into a continuous latent space and perform test-time adaptation by iteratively refining z via gradient ascent: $z_{k+1} = z_k + \eta \nabla_z \sum_{i=1}^n \log p_\theta(y_i \mid x_i, z_k)$. Alternatively, Liu et al. (2023) model latent navigation as an ODE driven by attribute-specific energy functions E_i (e.g., attribute classifier), where a vector field guiding movement in latent space, yielding: $dz = \frac{1}{2}\beta(t) \sum_i \nabla_z E_i(a_i \mid z)\, dt$. For VQ-VAE, Hosking et al. (2022) performs beam search over hierarchical latent syntax codes conditioned on semantic information, and uses the resulting latent sketch to control generation.

Although latent searching offers greater flexibility than latent manipulation and arithmetic, it typically requires higher computational cost and longer inference time.

Summary and Limitations. Latent manipulation and latent arithmetic offer computationally efficient control mechanisms, but they rely on strong assumptions such as linearity and disentanglement. In contrast, latent search provides a more flexible and expressive framework, albeit with higher computational cost. VAEs with continuous latent space enable more efficient control and search than discrete latent spaces due to their smooth and differentiable geometry, which supports gradient-based optimisation. In contrast, discrete latent spaces require combinatorial or non-differentiable search, making optimisation over inherently discrete text representations particularly challenging.

5 Evaluation Approaches

In this section, we review evaluation approaches from two perspectives: the latent space and the decoding process. These correspond to assessing four properties. We also summarise the benchmark datasets used in prior work in Tables 3 and 4.

Latent Space Evaluation. Based on prior literature, latent representations can be evaluated along four key dimensions, each capturing different aspects of structure, information, and accessibility.

(1) Visualisation. For VAE and VQ-VAE, dimensionality reduction techniques such as linear *Principal Component Analysis (PCA)*, and non-linear *t-SNE* (Van der Maaten and Hinton, 2008) and *UMAP* (McInnes et al., 2018) are commonly employed to project high-dimensional latent representations into low-dimensional spaces. These visualisations provide qualitative evidence of clustering, separation, and continuity with respect to underlying semantic factors.

(2) Information-theoretic Measures. Quantitative evaluation of latent information and structure can be performed using statistical measures such as MI and TC, as well as disentanglement metrics (as illustrated in Table 5) including Mutual Information Gap (MIG), Modularity, Completeness, and Explicitness. Together, these metrics assess the extent to which individual latent dimensions capture distinct generative factors while remaining invariant to others (Carvalho et al., 2023).

Compared to visualisation-based approaches, these metrics provide more rigorous and quantitative evaluation. However, they often rely on strong assumptions (e.g., factor independence or known ground-truth factors) and may not fully reflect the complexity of real-world linguistic representations.

(3) Probing-based Evaluation. Linear prober (e.g., classifier) are trained on latent representations to predict target attributes (e.g., syntax depth in paraphrasing). High probing accuracy indicates that the corresponding factor is explicitly encoded and accessible in the latent space (Zhang et al., 2024e).

(4) Retrieval-based Evaluation. Latent representations can also be evaluated through semantic similarity and retrieval tasks. For example, performance on Semantic Textual Similarity (STS) benchmarks is typically measured using Pearson or Spearman correlation coefficients (Zhang et al., 2023), while retrieval-specific metrics such as Re-

call and F1 score can be applied under appropriate experimental settings (Vasilakes et al., 2022). It can better evaluate complex semantic structure information in latent space. Its effectiveness, however, depends on the careful design of the target corpus and evaluation tasks.

Decoding Evaluation. While latent space evaluation focuses on representational structure, decoding evaluation assesses how such structure translates into generation quality and controllability.

(1) Language Modelling Metrics. Standard metrics such as BLEU, ROUGE, and Perplexity (PPL) are used to evaluate fluency and reconstruction fidelity (Li et al., 2020). These metrics primarily reflect P1 (faithfulness).

(2) Task-specific Evaluation. Task-specific metrics evaluate performance with respect to downstream objectives. For example, in truthfulness steering, question answering performance can be measured using Exact Match (EM) (Zhang et al., 2024a). Compared to general language modelling metrics, these evaluations are more aligned with practical goals, but they are task-dependent and may lack comparability across settings.

(3) Control Evaluation. The effectiveness of attribute control is typically measured using pretrained classifiers that assess whether generated outputs satisfy desired attributes (Gu et al., 2022). For structured tasks such as paraphrasing, controllability can also be quantified using metrics such as tree edit distance (Zhang and Shasha, 1989). These approaches directly evaluate P4 (localised control).

(4) Geometrical Probing. Geometrical probing evaluates controllability by applying latent operations such as traversal, interpolation, and arithmetic, and observing their effects on generated outputs (Zhang et al., 2024c). Zhang et al. (2024c) further propose Interpolation Smoothness metric to quantify their evaluation. This provides intuitive evidence of latent structure and compositionality.

Summary and Limitations. Overall, latent space evaluation assesses the structural quality of representations (e.g., via visualisation, probing, and information-theoretic metrics), while decoding evaluation measures generation quality and controllability. Together, they provide a complementary view of whether latent variables are both meaningful and useful. However, current evaluation methods have notable limitations. Many metrics rely on simplifying assumptions (e.g., independence or linearity), qualitative analyses are subjective, and control evaluation often depends on biased pretrained classifiers. Moreover, improvements in latent disentanglement do not always translate to better generation performance. These issues highlight the need for more robust and standardised evaluation frameworks.

6 Challenges and Future Directions

In this section, we outline the potential challenges and future directions.

(1) Latent Compositional Reasoning. Variational Autoencoders (VAEs) have been extensively investigated for generative drug design, where they are used to encode and decode chemical structures, thereby enabling the synthesis of novel molecular candidates (Ochiai et al., 2023). In the neurosymbolic domain, latent representations have also shown promise in facilitating the discovery of mathematical formulas (Li et al., 2026).

Nevertheless, the application of VAEs to scientific NLI, such as biomedical question answering that require multi-step entailment reasoning (Jullien et al., 2023), remains relatively underexplored. This gap motivates the following research question: ***RQ:*** *How can prior knowledge, reasoning structures, and logical constraints be efficiently encoded into latent representations to enhance NLI models?* Answering this question allows us to improve reasoning interpretability and safety.

(2) Compositional Diffusion LMs. diffusion-based language models provide a promising paradigm for capturing compositional semantics (Li et al., 2025). In particular, syntactic structure can be explicitly integrated into the design of the diffusion process, such as through structured noise schedules that progressively corrupt and reconstruct representations in alignment with the hierarchical organization of language. This motivates the following research question: ***RQ:*** *How can compositional language modelling be systematically achieved through the incorporation of diffusion processes?*

For instance, the model may first establish a coarse syntactic backbone (e.g., subject–verb–object structure), and then incrementally recover phrases, modifiers, and finer-grained lexical details. Under this paradigm, diffusion time steps can be aligned with the stages of language generation (from random noise → syntactic structure → semantic content → phrases → complete sentences). Such alignment introduces an inductive bias toward linguistic structure, thereby enhancing grammatical consistency, improving long-range coherence in sentence generation, and increasing controllability over the generation process.

Limitations

This survey primarily focuses on autoencoder-based frameworks (VAE, VQ-VAE, and Sparse Autoencoders) for shaping latent semantic geometry, and therefore does not comprehensively cover alternative interpretability or control paradigms outside this design space. In addition, the surveyed studies are predominantly based on English-language datasets, limiting insights into multilingual or low-resource settings. Finally, given the rapid evolution of language model interpretability research, some recent developments may not be fully captured.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Shaden Naif Alshammari, John R Hershey, Axel Feldmann, William T Freeman, and Mark Hamilton. 2025. I-con: A unifying framework for representation learning. In *The Thirteenth International Conference on Learning Representations*.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Clément Bonnet and Matthew V Macfarlane. 2024. Searching latent program spaces.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Danilo Carvalho, Yingji Zhang, Harriet Unsworth, and Andre Freitas. 2025. LangVAE and LangSpace: Building and probing for language model VAEs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 749–759, Suzhou, China. Association for Computational Linguistics.

Danilo S. Carvalho, Yingji Zhang, Giangiacomo Mercatali, and Andre Freitas. 2023. Learning disentangled representations for natural language definitions. In *Findings of the European chapter of Association for Computational Linguistics (Findings of EACL)*. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. 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 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. 2018. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2615–2625.

Seunghyuk Cho, Juyong Lee, and Dongwoo Kim. 2023. Hyperbolic vae via latent gaussian distributions. *Advances in Neural Information Processing Systems*, 36:569–588.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. 2018. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Boyi Deng, Yu Wan, Yidan Zhang, Baosong Yang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. *arXiv preprint arXiv:2505.05111*.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 193–202, New York, NY, USA. Association for Computing Machinery.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li. 2020. Pre-train and plug-in: Flexible conditional text generation with variational auto-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 253–262, Online. Association for Computational Linguistics.

Cian Eastwood and Christopher KI Williams. 2018. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. *arXiv preprint arXiv:1908.11527*.

Xianghong Fang, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Dit-Yan Yeung. 2022. Controlled text generation using dictionary prior in variational autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 97–111, Dublin, Ireland. Association for Computational Linguistics.

Ghazi Felhi, Joseph Le Roux, and Djamé Seddah. 2022a. Exploiting inductive bias in transformers for unsupervised disentanglement of syntax and semantics with VAEs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5763–5776, Seattle, United States. Association for Computational Linguistics.

Ghazi Felhi, Joseph Le Roux, and Djamé Seddah. 2022b. Towards unsupervised content disentanglement in sentence representations via syntactic roles. *arXiv preprint arXiv:2206.11184*.

Victoria Fromkin, Robert Rodman, Nina M Hyams, Mengistu Amberber, Felicity Cox, and Rosalind Thornton. 2017. *An Introduction to Language with Online Study Tools 12 Months*. Cengage AU.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018, Miami, Florida, USA. Association for Computational Linguistics.

Ankur Garg, Xuemin Yu, Hassan Sajjad, and Samira Ebrahimi Kahou. 2025. Cross-layer discrete concept discovery for interpreting language models.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long Papers), pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Amulya Gupta and Zhu Zhang. 2021. Vector-quantization-based topic modeling. *ACM Trans. Intell. Syst. Technol.*, 12(3).

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. 2025. Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models. *arXiv preprint arXiv:2505.16188*.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495, Mexico City, Mexico. Association for Computational Linguistics.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1418, Online. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2489–2501, Dublin, Ireland. Association for Computational Linguistics.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. 2022. Fuse it more deeply! a variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 697–716, Seattle, United States. Association for Computational Linguistics.

Qiumei Hu, Jingwu Xiao, Junting Lu, Xiankui Zhang, Xueyang Liu, and Wenhui Hu. 2024. Short text topic modeling with vector quantization. In *International Conference on Data Mining and Big Data*, pages 206–217. Springer.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1372–1379, Online. Association for Computational Linguistics.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. *arXiv preprint arXiv:2101.10579*.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. *arXiv preprint arXiv:2010.00735*.

Shawn Im and Yixuan Li. 2025. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018a. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. 2018b. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*.

Yi Jing, Zijun Yao, Lingxu Ran, Hongzhu Guo, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. Sparse auto-encoder interprets linguistic features in large language models. *arXiv preprint arXiv:2502.20344*.

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.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly

supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Qian Li, Yuxiao Hu, Juncheng Liu, and Yuntian Chen. 2026. Gensr: Symbolic regression based in equation generative space. *arXiv preprint arXiv:2602.20557*.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*.

Zhuang Li, Lizhen Qu, Qiongkai Xu, Tongtong Wu, Tianyang Zhan, and Gholamreza Haffari. 2022. Variational autoencoder with disentanglement priors for low-resource task-specific natural language generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10335–10356, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023. Composable text controls in latent space with ODEs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16543–16570, Singapore. Association for Computational Linguistics.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for

paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190, Montréal, Canada. Association for Computational Linguistics.

Kikuo Maekawa. 2000. Spontaneous speech corpus of japanese. In *Proc. LREC2000 (Second International Conference on Language Resources and Evaluation)*.

Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A. Smith, and James Henderson. 2020. Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6076–6092, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2023. A symbolic framework for systematic evaluation of mathematical reasoning with transformers. *arXiv preprint arXiv:2305.12563*.

Giangiacomo Mercatali and André Freitas. 2021. Disentangling generative factors in natural language with discrete variational autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3547–3556, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chugtai, and Neel Nanda. 2025. Robustly identifying concepts introduced during chat fine-tuning using crosscoders. *arXiv preprint arXiv:2504.02922*.

Francesco Maria Molfese, Simone Conia, Riccardo Orlando, and Roberto Navigli. 2024. ZEBRA: Zero-shot example-based retrieval augmentation for commonsense question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22429–22444, Miami, Florida, USA. Association for Computational Linguistics.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

Amine M'Charrak. 2018. Deep learning for natural language processing (nlp) using variational autoencoders (vae).

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2016. Definition modeling: Learning to define word embeddings in natural language.

Toshiki Ochiai, Tensei Inukai, Manato Akiyama, Kairi Furui, Masahito Ohue, Nobuaki Matsumori, Shinsuke Inuki, Motonari Uesugi, Toshiaki Sunazuka, Kazuya Kikuchi, et al. 2023. Variational autoencoder-based chemical latent space for large molecular structures with 3d complexity. *Communications Chemistry*, 6(1):249.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Ridgeway and Michael C Mozer. 2018. Learning deep disentangled embeddings with the f-statistic loss. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 185–194.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6033–6039, Florence, Italy. Association for Computational Linguistics.

Ramses Sanchez, Lukas Conrads, Pascal Welke, Kostadin Cvejoski, and Cesar Ojeda Marin. 2023. Hidden schema networks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4764–4798, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. 2022. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*.

Lei Sha and Thomas Lukasiewicz. 2024. Text attribute control via closed-loop disentanglement. *Transactions of the Association for Computational Linguistics*, 12:190–209.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *International Conference on Machine Learning*, pages 8719–8729. PMLR.

Yukai Shi, Sen Zhang, Chenxing Zhou, Xiaodan Liang, Xiaojun Yang, and Liang Lin. 2021. Gtae: graph transformer–based auto-encoders for linguistic-constrained text style transfer. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(3):1–16.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Haoqin Tu, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2022. Adavae: Exploring adaptive gpt-2s in variational auto-encoders for language modeling. *arXiv preprint arXiv:2205.05862*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jake Vasilakes, Chrysoula Zerva, Makoto Miwa, and Sophia Ananiadou. 2022. Learning disentangled representations of negation and uncertainty. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8380–8397, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. Base models know how to reason, thinking models learn when.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human*

Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Haoran Xu, Sixing Lu, Zhongkai Sun, Chengyuan Ma, and Chenlei Guo. 2021. Vae based text style transfer with pivot words enhancement learning. *arXiv preprint arXiv:2112.03154*.

Jiacheng Xu and Greg Durrett. 2018. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4503–4513.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pages 3881–3890. PMLR.

YoungJoon Yoo and Jongwon Choi. 2024. Topic-vq-vae: leveraging latent codebooks for flexible topic-guided document generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19422–19430.

Daiki Yoshioka, Yusuke Yasuda, Noriyuki Matsunaga, Yamato Ohtani, and Tomoki Toda. 2022. Spoken-text-style transfer with conditional variational autoencoder and content word storage. In *INTERSPEECH*, pages 4576–4580.

Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, et al. 2026. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024a. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Xinyuan Zhang, Yi Yang, Siyang Yuan, Dinghan Shen, and Lawrence Carin. 2019. Syntax-infused variational autoencoder for text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2069–2078, Florence, Italy. Association for Computational Linguistics.

Yingji Zhang, Danilo Carvalho, and Andre Freitas. 2024b. Learning disentangled semantic spaces of explanations via invertible neural networks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2113–2134, Bangkok, Thailand. Association for Computational Linguistics.

Yingji Zhang, Danilo Carvalho, and Andre Freitas. 2025a. Quasi-symbolic semantic geometry over transformer-based variational autoencoder. In *Proceedings of the 29th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Yingji Zhang, Danilo Carvalho, Marco Valentino, Ian Pratt-Hartmann, and Andre Freitas. 2024c. Improving semantic control in discrete latent spaces with transformer quantized variational autoencoders. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1434–1450, St. Julian's, Malta. Association for Computational Linguistics.

Yingji Zhang, Danilo S. Carvalho, and Andre Freitas. 2024d. Guiding explanatory inference through inference types.

Yingji Zhang, Danilo S Carvalho, Ian Pratt-Hartmann, and André Freitas. 2023. Llamavae: Guiding large language model generation via continuous latent sentence spaces. *arXiv preprint arXiv:2312.13208*.

Yingji Zhang, Marco Valentino, Danilo Carvalho, Ian Pratt-Hartmann, and Andre Freitas. 2024e. Graph-induced syntactic-semantic spaces in transformer-based variational AutoEncoders. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 474–489, Mexico City, Mexico. Association for Computational Linguistics.

Yingji Zhang, Marco Valentino, Danilo S. Carvalho, and André Freitas. 2025b. Learning to disentangle latent reasoning rules with language vaes: A systematic study. *The Fortieth AAAI Conference on Artificial Intelligence: AAAI 2026*.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. 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 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.

Shu Zhong, Mingyu Xu, Tenglong Ao, and Guang Shi. 2025. Understanding transformer from the perspective of associative memory. *arXiv preprint arXiv:2505.19488*.

Ying Zhou, Xinyao Wang, Yulei Niu, Yaojie Shen, Lexin Tang, Fan Chen, Ben He, Le Sun, and Longyin Wen. 2024. Difflm: Controllable synthetic data generation via diffusion language models. *arXiv preprint arXiv:2411.03250*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Architecture Details

VAE. VAE comprises two primary components: an encoder and a decoder. The encoder maps an input x to a distribution over a latent space, typically parameterised as a Gaussian, while the decoder reconstructs the input by sampling from this latent distribution. Training optimises a variational objective consisting of a reconstruction term and a regularisation term that enforces proximity between the approximate posterior and a predefined prior over the latent variables, thereby encouraging a smooth and structured latent space:

$$\mathcal{L}_{\text{VAE}}(x;\theta,\phi) = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathrm{KL}(q_\phi(z|x)\,\|\,p(z))\,. \quad (2)$$

A central geometric mechanism in such models is latent space injection, which determines how a sampled latent variable z influences the generative process. A representative instantiation is the Optimus architecture, which combines BERT (Devlin et al., 2018) as the encoder with GPT-2 (Radford et al., 2019) as the decoder. In this setup, BERT encodes the input sentence x into a fixed-length representation via the `[CLS]` token. This representation parameterises a Gaussian latent distribution $\mathcal{N}(\mu,\Sigma)$, where both μ and Σ are learned. A latent sample $z \sim \mathcal{N}(\mu,\Sigma)$ is then injected into the decoder through different mechanisms, as summarised in Table 2.

VQ-VAE. The VQ-VAE extends the VAE framework by introducing discrete latent representations. It consists of three components: (i) an encoder that maps the input x to continuous latent vectors, (ii) a quantisation module that replaces each vector with its nearest neighbour in a learned codebook, and (iii) a decoder that reconstructs the input from these quantised representations. The training objective combines a reconstruction loss, a codebook loss that aligns encoder outputs with codebook embeddings, and a commitment loss that encourages encoder outputs to remain close to the selected embeddings:

$$\mathcal{L}_{\text{VQ-VAE}} = \underbrace{\|x-\hat{x}\|_2^2}_{\text{reconstruction loss}} + \underbrace{\|\text{sg}[z_e(x)]-e\|_2^2}_{\text{codebook loss}} + \underbrace{\beta\|z_e(x)-\text{sg}[e]\|_2^2}_{\text{commitment loss}}\,. \quad (3)$$

SAE. SAE consists of a linear encoder that projects the input into a higher-dimensional latent space and a linear decoder that reconstructs the input from this representation. To enforce sparsity, an additional regularisation term—commonly an ℓ_1 penalty or a KL-divergence-based constraint—is applied to the latent activations. This encourages most neurons to remain inactive for any given input, thereby promoting the learning of meaningful and disentangled features while preventing trivial identity mappings. A comprehensive survey of SAE architectures and optimisation strategies is provided by Shu et al. (2025). The objective is given by:

$$\mathcal{L}_{\text{SAE}} = \underbrace{\|x-\hat{x}\|_2^2}_{\text{reconstruction loss}} + \underbrace{\lambda\|z\|_1}_{\text{sparsity penalty}}\,. \quad (4)$$

B Dataset and Benchmark

In this section, we introduce the main dataset and task used for semantic representation learning. A summary of the datasets for each study is provided in Tables 3 and 4.

Style Transfer. Style transfer datasets capture attribute-driven semantic variations, such as sentiment, topic, or tone, and are widely used in controlled text generation tasks. Representative corpora include Yelp (Shen et al., 2017), Yahoo Answers (Yang et al., 2017), dSentences (M'Charrak, 2018), IMDb (Maas et al., 2011), AGNews (Zhang et al., 2015), and Jigsaw. In addition, SAEs have been explored as a mechanism for steering truthfulness in LLMs. Benchmarks used to evaluate this capability include TruthfulQA (Lin et al., 2022), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and FACTOR (Muhlgay et al., 2024).

Latent injection setup	Summarisation
$\text{softmax}\left(\frac{Q[z;K]^T}{\sqrt{d}}\right)[z;V]$	**KV Memory**: LM decoder only rely on attached token z which is more likely to be ignored by self-attention, even exacerbating KL vanishing (Li et al., 2020).
$\text{softmax}\left(\frac{Q[z+K]^T}{\sqrt{d}}\right)[z+V]$	**KV Addition**: Simple and global interventions in the sequence of attention of LM decoder, which can mitigate the KL vanishing without additional parameters (Zhang et al., 2024e).
$\text{softmax}\left(\frac{[Q+z]K^T}{\sqrt{d}}\right)V$	**Query Addition**: Integrating the latent representation z into the decoder via query can deliver effective searching over Key and Value in decoders' parametric space (Zhang et al., 2025b; Zhong et al., 2025).
$\hat{K} = \left(\sum_{i=1}^{r} W_k^i[K;1]\right) \circ \left(\sum_{i=1}^{r} W_z^i[z;1]\right)$	**KV Fusion**: Using tensor fuse operation (Liu et al., 2018) to deeply fuse latent variables with hidden states in the LM decoder, but requiring additional training parameters (Hu et al., 2022).
$h_{\text{cache}} = W_m z$	**KV Cache**: Flexibly integrating z into the attention of various LM decoders with the cost of perturbing pretrained knowledge and more training parameters (Carvalho et al., 2025).
$P = \text{MLP}(z)$ (where P are k soft prompt embeddings)	**Latent Prompt**: Seamlessly and flexibly integrating z into LM decoders, but work on high-dimensional token embeddings. High dimensionality and complexity of the sequence space makes the search/optimisation difficult (Zhou et al., 2024).

Table 2: Latent space injection methods in Transformer-based Language VAE.

Text Paraphrasing. Paraphrasing datasets focus on semantic equivalence under surface variation, providing pairs of sentences that convey the same meaning with different lexical or syntactic realisations. Prominent resources include ParaNMT-50M (Wieting and Gimpel, 2018), the Penn Treebank (PTB) (Marcus et al., 1993), SNLI (Bowman et al., 2015), and the Quora Question Pairs dataset[2].

Definition Modelling. Definition modelling datasets capture mappings from lexical items to their corresponding conceptual descriptions, often following canonical patterns such as *"X is a Y that Z"*. These resources are used to study the transformation between a word (or its embedding) and its textual definition (Noraset et al., 2016; Mickus et al., 2022). Commonly used corpora include Wikipedia, WordNet, and Wiktionary.

Natural Language Inference. Datasets with structured explanations, such as atomic and conditional statements, encode multi-step reasoning in natural language and serve as key benchmarks for compositional inference. Representative resources include WorldTree (Jansen et al., 2018b), EntailmentBank (Dalvi et al., 2021), and Zebra (Molfese et al., 2024), all of which provide explicit reasoning chains that facilitate the study of compositional generalisation.

In addition, the Math Derivation dataset (Meadows et al., 2023) introduces a highly controlled corpus for symbolic reasoning. In this setting, models are required to generate derived mathematical expressions given an input expression and a specified operation (e.g., differentiation, addition, or subtraction). To evaluate out-of-distribution (OOD) generalisation and assess whether models capture underlying reasoning patterns, the dataset defines five evaluation splits:

(1) EVAL: mathematical expressions following the training set's distribution (like $U + cos(n)$), (2) VAR: mathematical expressions composed of a different set of variables (like $U + cos(beta)$), (3) EASY: mathematical expressions with a lower number of variables (like $cos(n)$), (4) EQ: mathematical derivations with equality insertions (like $E = U + cos(n)$), (5) LEN: mathematical derivations with a higher number of variables (like $U + cos(n)) + A + B$).

Moreover, SyllogisticNLI (Valentino et al., 2025) introduces a controlled benchmark for syllogistic deductive reasoning, enabling the evaluation of logical inference under minimal confounding linguistic variation.

C Disentanglement Metrics

Disentanglement metrics provide quantitative measures of the alignment between latent dimensions and underlying generative factors, as well as the degree of independence among latent variables. These metrics can be broadly categorised into two groups: *dimension-based* and *factor-based* metrics.

[2] https://www.kaggle.com/c/quora-question-pairs

Literature	Task, Dataset, and Benchmark	Encoding	Decoding
Hu et al. (2017)	Sentiment Transfer: IMDB (Diao et al., 2014), SST-full (Socher et al., 2013), Lexicon (Wilson et al., 2005)	-	Classifier Acc
Fang et al. (2019)	Sentiment Transfer: PTB (Marcus et al., 1993), Yahoo, Yelp (He et al., 2019; Yang et al., 2017)	KL, MI, AU	Classifier Acc, BLEU, PPL, RPPL, ...
John et al. (2019)	Sentiment Transfer: Yelp Review (Shen et al., 2017), Amazon Review (Fu et al., 2018)	Cosine Sim T-SNE	Classifier Acc, Word Overlap, PPL, ...
Mai et al. (2020)	Sentiment Transfer: WikiLarge (Zhang and Lapata, 2017), Yelp Review (Shen et al., 2017)	-	Classifier Acc, BLEU
Shen et al. (2020)	Sentiment Transfer: Yelp, Yahoo (Yang et al., 2017)	Recall	Parser Acc, BLEU, PPL
Rimsky et al. (2024)	LLM Steering: AI Risk and Sycophancy (Perez et al., 2023)	PCA	QA match
Shi et al. (2021)	Sentiment Transfer: Yelp review (Shen et al., 2017), Political Slant (Prabhumoye et al., 2018)	-	WMD, PPL, Classifier Acc, BLEU
Sha and Lukasiewicz (2024)	Sentiment Transfer: Yelp review (Shen et al., 2017), Amazon Review (Fu et al., 2018)	MIG, T-SNE	Classifier Acc, PPL, BLEU
Li et al. (2022)	Sentiment Transfer: EMPATHETIC, TACRED	T-SNE	Classifier Acc, WMD, BLEU
Xu et al. (2021)	Sentiment/Code-switch/Formality Transfer: Yelp, GYAFC (Rao and Tetreault, 2018), LinCE (Aguilar et al., 2020)	-	Classifier Acc, PPL, BLEU
Hu and Li (2021)	Sentiment/Category Transfer: Yelp, BIOS (De-Arteaga et al., 2019)	-	Classifier Acc, Fluency, Diversity
Huang et al. (2020)	Sentiment/Topic Transfer: Yelp, Yahoo QA	-	Classifier Acc, BLEU, PPL, RPPL
Duan et al. (2020)	Sentiment/Topic/length Transfer: Yelp, News Titles (Fu et al., 2018)	-	Classifier Acc, Diversity (Li et al., 2016)
Gu et al. (2022)	Sentiment/Topic/Detox Transfer: IMDb (Maas et al., 2011), AGNews (Zhang et al., 2015), Jigsaw Toxic	PCA	Classifier Acc, Distinctness (Li et al., 2016)
Liu et al. (2020)	Sentiment/topic Transfer: Yelp, Amazon reviews (He and McAuley, 2016)	-	Classifier Acc, PPL, Word Overlap, BLEU
Gu et al. (2023)	Sentiment/Topic/Detox Transfer: IMDb (Maas et al., 2011), AGNews (Zhang et al., 2015), Jigsaw Toxic	-	Classifier Acc, Distinctness (Li et al., 2016)
Fang et al. (2022)	Sentiment/Topic/Tense Transfer: PTB (Marcus et al., 1993), SNLI (Bowman et al., 2015), Yahoo Answers (Xu and Durrett, 2018), Yelp (Yang et al., 2017)	MI, AU	Classifier Acc, PPL, Distinctness (Li et al., 2016)
Gupta and Zhang (2021)	Topic Transfer: 20 Ng (Lang, 1995), Conala (https://conala-corpus.github.io/),	-	PPL, Topic Coherence, Topic Diversity, Topic Interpretability
Hu et al. (2024)	Topic Transfer:		
Yoo and Choi (2024)	Topic Transfer: 20 Ng (Lang, 1995), New York Times	T-SNE	Topic Coherence, Topic Diversity
Liu et al. (2023)	Sentiment/Negation/Tense/Formality Transfer: Yelp (Shen et al., 2017), Amazon Review (Fu et al., 2018)	-	Classifier Acc, PPL, BLEU, CTC (Deng et al., 2021)
Vasilakes et al. (2022)	Negation/Uncertainty Transfer: Yelp (Shen et al., 2017), Amazon Review (Fu et al., 2018)	T-SNE, Precision, recall, F1	PPL, BLEU
He et al. (2025)	Truthfulness/Sentiment Steering: TruthGen (Fulay et al., 2024)	-	Steering Success Rate, Lexical Diversity, Entropy
Zhang et al. (2024a)	Truthfulness Steering: TruthfulQA (Lin et al., 2022), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), FACTOR (Muhlgay et al., 2024)	Kernel density estimate	QA match
Li et al. (2023)	Truthfulness Steering: TruthfulQA (Lin et al., 2022)	-	QA match, Cross Entropy, KL
Deng et al. (2025)	Code-switch Steering: Flores-200 (Costa-Jussà et al., 2022)	-	Adversarial Language Identification, CrossLingual Continuation
Minder et al. (2025)	chat-specific features, ...	SAE	
Zhao et al. (2025)	Context-Memory Steering: NQSwap (Longpre et al., 2021), Macnoise (Hong et al., 2024)	-	QA match
Yoshioka et al. (2022)	Text-to-speech Synthesis: CSJ (Maekawa, 2000), KVJ	-	Classifier Acc, BLEU

Table 3: Dataset and evaluation metrics for task-specific attribute latent space model.

(1) Dimension-based metrics evaluate disentanglement by analysing how individual latent dimensions respond to variations in generative factors. Common examples include $z_{\text{min-var}}$ error (Kim and Mnih, 2018), *Modularity* (Ridgeway and Mozer, 2018), and the *Disentanglement Score* (Eastwood and Williams, 2018). These metrics assess whether a given latent dimension re-

Literature	Task, Dataset, and Benchmark	Encoding	Decoding
	structural attribute space		
Mercatali and Freitas (2021)	Syntax Disentanglement: Yelp, dSentences	Z-min, Z-diff, MIG	Parser Acc
Felhi et al. (2022b)	Syntax Disentanglement: SNLI (Bowman et al., 2015)	MIG	qualitative eval
Hosking et al. (2022)	Question Paraphrasing: Paralex (Fader et al., 2013), Quora Question Pairs (https://www.kaggle.com/c/quora-question-pairs), MSCOCO 2017 (Lin et al., 2014)	T-SNE	iBLEU, BLEU, Self-B
Huang and Chang (2021)	Paraphrasing: ParaNMT-50M (Wieting and Gimpel, 2018), Quora, PAN (Madnani et al., 2012), MRPC (Dolan et al., 2004)	-	BLEU, Parsing Match
Zhang et al. (2019)	Paraphrasing: PTB (Marcus et al., 1993), Wikipedia (Gulordava et al., 2018)	-	Relevance, Readability, Diversity, PPL
Bao et al. (2019)	Paraphrasing: PTB (Marcus et al., 1993), Quora Question Pairs (https://www.kaggle.com/c/quora-question-pairs)	-	BLEU, PPL, RPPL, Tree Edit Distance (Zhang and Shasha, 1989)
Chen et al. (2019)	Paraphrasing: ParaNMT-50M (Wieting and Gimpel, 2018)	Pearson correlation (STS bench)	Tree Edit Distance, Constituent Parsing F1, POS Tagging
Huang et al. (2021)	Paraphrasing: ParaNMT-50M (Wieting and Gimpel, 2018), Quora Question Pairs (https://www.kaggle.com/c/quora-question-pairs)	Pearson correlation (STS bench), Classifier Acc (BShift, TreeDepth, TopConst), Cosine Sim	-
Felhi et al. (2022a)	Semantic-syntax Disentanglement: ParaNMT-50M (Wieting and Gimpel, 2018)	-	tree edit distance, constituency tree (TMA2), ParaBART (consine sim), Meteor score
Zhang et al. (2024e)	Semantic-syntax Disentanglement: EntailmentBank (Dalvi et al., 2021), WorldTree (Jansen et al., 2018a), Math Derivation (Meadows et al., 2023)	MI, KL, Wasserstein Distance, MSE, T-SNE	BLEU, BLEURT, Cosine Sim, PPL
Hosking and Lapata (2021)	Question Paraphrasing: Paralex (Fader et al., 2013), Quora Question Pairs (https://www.kaggle.com/c/quora-question-pairs)	-	BLEU, Self-BLEU, iBLEU
Carvalho et al. (2023)	Semantic Disentanglement: Wordnet, Wiktionary, Wikipedia	T-SNE, UMAP, z-diff z-min-var, MIG Modularity, Explicitness, Disentanglement, Completeness, Informativeness	BLEU, PPL
Zhang et al. (2024b)	Semantic Disentanglement: EntailmentBank (Dalvi et al., 2021), WorldTree (Jansen et al., 2018a)	T-SNE, PCA, Classifier Acc	Interpolation Smoothness
Zhang et al. (2025a)	Semantic Disentanglement: EntailmentBank (Dalvi et al., 2021), WorldTree (Jansen et al., 2018a)	T-SNE, PCA, Classifier Acc	Arithmetic Ratio
Roy and Grangier (2019)	Semantics (concept): WMT'17 English-German (Bojar et al., 2017), WMT-Newscrawl		Classifier Acc, F1, BLEU, Regression
Zhang et al. (2024c)	Semantics (concept): EntailmentBank (Dalvi et al., 2021), WorldTree (Jansen et al., 2018a), Math Derivation (Meadows et al., 2023)	T-SNE	BLEU, PPL, Cosine, BLEURT, Interpolation Smoothness
Garg et al. (2025)	Semantics (concept): ERASER (Pang and Lee, 2004), Jigsaw (https://kaggle.com/ competitions/jigsaw-toxic-comment-classification-challenge), AGNEWS	-	Fleiss' Kappa, Average Confidence, Model Alignment Rate
Jing et al. (2025)	Semantics: (Fromkin et al., 2017)	Feature Representation Confidence (FRC)	Feature Intervention Confidence (FIC)
Zhang et al. (2024d)	Reasoning Disentanglement: EntailmentBank (Dalvi et al., 2021)	PCA	BLEU, BLEURT, Cosine, LLM AutoEval
Zhang et al. (2025b)	Reasoning Disentanglement: EntailmentBank (Dalvi et al., 2021), Math Derivation (Meadows et al., 2023), Syllogism	PCA	BLEU, Exact Match
Sanchez et al. (2023)	Reasoning Graph: PTB (Marcus et al., 1993), Yahoo, Yelp, ATOMIC (Sap et al., 2019)	MI	BLEU, BERT Score (Zhang* et al., 2020), PPL
Venhoff et al. (2025)	reasoning patterns	SAE	

Table 4: Dataset and evaluation for structural attribute latent space model.

mains invariant to changes in unrelated factors, indicating that it encodes a distinct and isolated

Category	Metrics	Summarisation
Dimension-based	z_{min_var} *error*	Identify dimensions with the smallest variance across different samples for a given factor.
	Modularity	Assess how exclusively a dimension is associated with a single factor. If MI for one factor dominates a dimension and others are close to zero, modularity is high.
	Disentanglement Score	Use feature importances from a classifier to compute entropy for each dimension; lower entropy = better disentanglement.
Factor-based	*Mutual Information Gap*	Measure how much more one latent dimension knows about a factor compared to others. A larger gap between the top two mutual information (MI) scores for a factor indicates stronger disentanglement.
	Completeness Score	Similar to Disentanglement Score, but entropy is computed over factors rather than dimensions.
	Informativeness Score	Use a classifier to predict factor labels from latent codes; score is based on prediction error (lower error = higher informativeness).
	Explicitness	Determine how linearly separable the factor information is from the latent representation. Use ROC AUC to quantify separability.

Table 5: Disentanglement metrics quantify the feature-dimension alignment. An implementation in Python is provided by (Carvalho et al., 2025).

feature.

(2) Factor-based metrics assess disentanglement from the perspective of individual generative factors by measuring how latent dimensions respond when a single factor varies. Representative metrics include the *Mutual Information Gap* (MIG) (Chen et al., 2018), *Completeness* and *Informativeness* (Eastwood and Williams, 2018), and *Explicitness* (Ridgeway and Mozer, 2018).

We summarise the definitions of the commonly used metrics below:

1. z_{diff} $accuracy$ (Higgins et al., 2017): The accuracy of a predictor for $p(y|z^b_{diff})$, where z^b_{diff} is the absolute linear difference between the inferred latent representations for a batch B of latent vectors, written as a percentage value. Higher values imply better disentanglement.

2. z_{min_var} $error$ (Kim and Mnih, 2018): For a chosen factor k, data is generated with this factor fixed but all other factors varying randomly; their representations are obtained, with each dimension normalised by its empirical standard deviation over the full data (or a large enough random subset); the empirical variance is taken for each dimension of these normalised representations. Then the index of the dimension with the lowest variance and the target index k provide one training input/output example for the classifier. Thus, if the representation is perfectly disentangled, the empirical variance in the dimension corresponding to the fixed factor will be 0. The representations are normalised so that the arg min is invariant to rescaling of the representations in each dimension. Since both inputs and outputs lie in a discrete space, the optimal classifier is the majority-vote classifier, and the metric is the error rate of the classifier. Lower values imply better disentanglement.

3. Mutual Information Gap (MIG) (Chen et al., 2018): The difference between the top two latent variables with the highest mutual information. Empirical mutual information between a latent representation z_j and a ground truth factor v_k, is estimated using the joint distribution defined by $q(z_j, v_k) = \sum_{n=1}^{N} p(v_k)p(n|v_k)q(z_j|n)$. A higher mutual information implies that z_j contains a more information about v_k, and the mutual information is maximal if there exists a deterministic, invertible relationship between z_j and v_k. MIG values are in the interval [0, 1], with higher values implying better disentanglement.

4. *Modularity* (Ridgeway and Mozer, 2018): The deviation from an ideally modular case of latent representation. If latent vector dimension i is ideally modular, it will have high mutual information with a single factor and zero mutual information with all other factors. A deviation δ_i of 0 indicates perfect modularity and 1 indicates that this dimension has equal mutual information with every factor. Thus, $1 - \delta_i$ is used as a modularity score for vector dimension i and the mean of $1 - \delta_i$ over i as the modularity score for the overall representation. Higher values imply better disentanglement.

5. *Explicitness* (Ridgeway and Mozer, 2018): Mean of the ROC area-under-the-curve (AUC_{jk}) of a one-versus-rest logistic-regression classifier that takes the latent vec-

tors as input and has factor values as targets, over a factor index j and an index k on values of factor j. Represents the coverage of the representation, in other words, how well each factor is represented. Higher values imply better disentanglement.

6. *Disentanglement Score* (Eastwood and Williams, 2018): The degree to which a representation factorises or disentangles the underlying factors of variation, with each variable (or dimension) capturing at most one generative factor. It is computed as a weighted average of a disentanglement score $D_i = (1 - H_K(P_i.))$ for each latent dimension variable c_i, on the relevance of each c_i, where $H_K(P_i.)$ denotes the entropy and P_{ij} denotes the 'probability' of c_i being important for predicting z_j. If c_i is important for predicting a single generative factor, the score will be 1. If c_i is equally important for predicting all generative factors, the score will be 0. Higher values imply better disentanglement.

7. *Completeness Score* (Eastwood and Williams, 2018): The degree to which each underlying factor is captured by a single latent dimension variable. For a given z_j it is given by $C_j = (1 - H_D(\tilde{P}.j))$, where $H_D(\tilde{P}.j) = -\sum_{d=0}^{D-1} \tilde{P}_{dj} log_D \tilde{P}_{ij}$ denotes the entropy of the $\tilde{P}.j$ distribution. If a single latent dimension variable contributes to z_j's prediction, the score will be 1 (complete). If all code variables contribute equally to z_j's prediction, the score will be 0 (maximally over-complete). Higher values imply better disentanglement.

8. *Informativeness Score* (Eastwood and Williams, 2018): The amount of information that a representation captures about the underlying factors of variation. Given a latent representation c, It is quantified for each generative factor z_j by the prediction error $E(z_j, \hat{z}_j)$ (averaged over the dataset), where E is an appropriate error function and $\hat{z}_j = f_j(c)$. Lower values imply better disentanglement.