Improving Semantic Control in Discrete Latent Spaces with Transformer Quantized Variational Autoencoders

Yingji Zhang[1†], **Danilo S. Carvalho**[1,3], **Marco Valentino**[2],
Ian Pratt-Hartmann[1], **André Freitas**[1,2,3]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] Idiap Research Institute, Switzerland
[3] National Biomarker Centre, CRUK-MI, Univ. of Manchester, United Kingdom
[1]{firstname.lastname}@[postgrad.][†]manchester.ac.uk
[2]{firstname.lastname}@idiap.ch

Abstract

Achieving precise semantic control over the latent spaces of Variational AutoEncoders (VAEs) holds significant value for downstream tasks in NLP as the underlying generative mechanisms could be better localised, explained and improved upon. Recent research, however, has struggled to achieve consistent results, primarily due to the inevitable loss of semantic information in the variational bottleneck and limited control over the decoding mechanism. To overcome these challenges, we investigate discrete latent spaces in Vector Quantized Variational AutoEncoders (VQVAEs) to improve semantic control and generation in Transformer-based VAEs. In particular, We propose T5VQVAE, a novel model that leverages the controllability of VQVAEs to guide the self-attention mechanism in T5 at the token-level, exploiting its full generalization capabilities. Experimental results indicate that T5VQVAE outperforms existing state-of-the-art VAE models, including Optimus, in terms of controllability and preservation of semantic information across different tasks such as auto-encoding of sentences and mathematical expressions, text transfer, and inference. Moreover, T5VQVAE exhibits improved inference capabilities, suggesting potential applications for downstream natural language and symbolic reasoning tasks.

1 Introduction

The emergence of deep generative neural networks supported by Variational AutoEncoders (VAEs) (Kingma and Welling, 2013) enables the localisation of syntactic and semantic properties within complex sentence latent spaces. By localising and manipulating these generative factors within the latent spaces, one can better control the properties of the textual output, enhancing performance on downstream tasks (Carvalho et al., 2023; John et al., 2019a), and providing mechanisms for representing and disentangling syntactic and semantic features within natural language (Zhang et al., 2023a, 2022; Mercatali and Freitas, 2021).

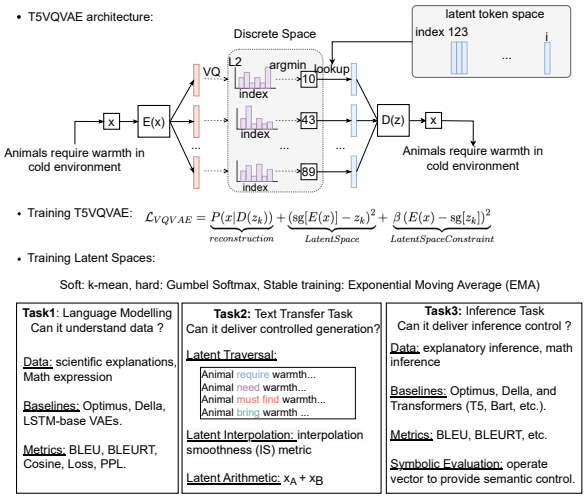

Figure 1: By controlling the token-level discrete latent space in VAEs, we aim to explicitly guide the cross-attention mechanism in T5 to improve the generation process. We focus on three challenging tasks to assess precise semantic control and inference.

Recent work (Carvalho et al., 2023; Zhang et al., 2022, 2023a) investigated controllable text generation via latent sentence geometry based on the canonical Optimus architecture (the first large pre-trained language VAE, Li et al. (2020)). However, the Optimus architecture brings its associated challenges since (i) the Optimus setup does not allow for a fine-grained (i.e., token-level) semantic control as sentence-level representation features are ignored by most attention heads especially in lower layers, where lexical-level semantics is captured (Hu et al., 2022); (ii) the sentence bottleneck in the VAE architecture leads to inevitable information loss during inference (Zhang et al., 2023b,d).

This work concentrates on addressing these architectural limitations by aiming to minimise the information loss in the latent space and effectively control the decoder and its attention mechanism.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1434–1450

The Vector Quantized Variational AutoEncoder (VQVAE) (Van Den Oord et al., 2017), as a discrete latent variable model, can be considered an ideal mechanism to alleviate these issues since it preserves and closely integrates both a coarse-grained continuous latent sentence space and a fine-grained latent token space that can preventinformation loss. More importantly, its latent token space can directly work on the cross-attention module (Vaswani et al., 2017) to guide the generation in seq2seq models, such as T5 (Raffel et al., 2020). Therefore, we hypothesise that such a mechanism can enable better generalisation and semantic control in Transformer-based VAEs.

Following these insights, we propose a novel approach named T5VQVAE, a model that leverages the controllability of VQVAE to guide the token-level self-attention mechanism during the generation process. We evaluate T5VQVAE on three challenging and diverse downstream tasks including (1) language modelling, (2) text transfer (guided text generation via the movement of latent vectors), and (3) natural language and symbolic inference tasks. An illustration of the complete model architecture and experimental setup can be found in Figure 1.

The overall contribution of the paper can be summarised as follows:

1. We propose T5VQVAE, the first pre-trained language Vector-Quantised variational Autoencoder, bridging the gap between VAEs and token-level representations, improving sentence-level localisation, controllability, and generalisation under VAE architectures. The experiments reveal that the proposed model outperforms previous state-of-the-art VAE models, including Optimus (Li et al., 2020), on three target tasks, as well as delivering improved semantic control when compared to the previous state-of-the-art.

2. We propose the Interpolation Smoothness (IS) metric for quantitatively evaluating sentence interpolation performance, a fundamental proxy for measuring the localisation of syntactic and semantic properties within sentence latent spaces. The experimental results indicate that T5VQVAE can lead to better interpolation paths (suggesting better interpretability and control).

3. Experiments on syllogistic-deductive NLI and mathematical expression derivation reveal that a quasi-symbolic behaviour may emerge in the latent space of T5VQVAE, and that the model can be explicitly controlled to achieve superior reasoning capabilities.

Our experimental code is available online[1] to encourage future work in the field.

2 Methodology

In this section, we first present our model, T5VQVAE, whose primary goal is to learn a latent space by reconstructing input sentences. Next, we illustrate its objective function, which consists of three parts designed to improve semantic control: reconstruction term, latent space optimization term, and encoder constraint term. Finally, we highlight the architectural advantages of T5VQVAE compared to Transformer-based VAEs.

Model architecture. Van Den Oord et al. (2017) first proposed the VQVAE architecture for learning a discretised latent space of images, showing that it can alleviate the issue of *posterior collapse*, in which the latent representations produced by the Encoder are ignored by the Decoder (Kingma and Welling, 2013). In this work, we propose to integrate T5 encoder/decoder into the VQVAE architecture for representation learning with natural language. T5 was selected due to its consistent performance across a large range of NLP tasks and its accessibility. To cast T5 into a VQVAE model, we first establish a latent token embedding space, denoted as the codebook, represented by $z \in \mathbb{R}^{K \times I}$. Here, K refers to the number of tokens in the codebook, and I represents the dimensionality of each token embedding. When given a token x, the Encoder E maps it into a vector representation, denoted as $E(x)$. Then, the nearest latent representation z_k from the codebook z is selected based on the $L2$ distance. The input of the cross-attention module can then be formalised as follows:

$$\hat{x} = \text{MultiHead}\left(D(x)W^q, z_k W^k, z_k W^v\right)$$

Here, z_k is the key and value and $D(x)$, which represents the input token embedding of the decoder, is the query. $\hat{x}$ represents the reconstructed token, while W^q, W^k, and W^v are trainable weights of **q**uery, **k**ey, and **v**alue.

[1] https://github.com/SnowYJ/T5VQVAE

Training T5VQVAE The training of T5VQVAE can be then considered as the optimisation of three independent parts, including $D(z_k)$, z_k, and $E(x)$. Starting from D, the model can be trained by maximising the reconstruction probability $P(x|D(z_k))$ via the teach-forcing scheme. Next, the z_k is optimised by minimising the $L2$ distance between $E(x)$ and z_k, which can be described as $(\text{sg}[E(x)]-z_k)^2$ where sg is the stop gradient operation. Finally, $E(x)$ can be trained via the $L2$ distance. By ensuring that $E(x)$ can learn the latent embedding under the constraint of $R^{K\times I}$ rather than learning an embedding directly, we can guide the model to achieve better performance. A commitment weight $\beta < 1$ is used to constraint the E close to z_k, which can be described as: $\beta(E(x)-\text{sg}[z_k])^2$. β is set to 0.25 following the same setup as (Van Den Oord et al., 2017) to preserve a behaviour consistent with their findings. The final objective function of T5VQVAE can be formalised as follows:

$$\mathcal{L}_{VQVAE} = \underbrace{P(x|D(z_k))}_{(1)reconstruction} + \underbrace{(\text{sg}[E(x)] - z_k)^2}_{(2)LatentSpace} + \underbrace{\beta\,(E(x) - \text{sg}[z_k])^2}_{(3)LatentSpaceConstraint}$$

Training the latent space. There are two possible strategies to update the latent space: *i.* k-means and *ii.* Gumbel softmax. Regarding k-means, for each token embedding w_i in a sentence, it selects the nearest latent token embedding, z_k, to its token embedding e^{w_i}. This process is equivalent to classifying e^{w_i} using k-means and then choosing the corresponding central point z_k as the input for $D(z_k)$. This can be expressed as follows:

$$z_{w_i} = z_k, \quad \text{where} \quad k = \text{argmin}_j \left\| e^{w_i} - z^j \right\|_2$$

To improve the stability of latent space training (term 2), we adapted the Exponential Moving Average (EMA) training scheme to update z (Roy et al., 2018). Figure 2 displays the training and testing loss curves of T5VQVAE with EMA or not. More details of EMA are provided in Appendix A. Instead of using k-means, which performs a soft selection of the index k, we can utilize the Gumbel softmax trick (Jang et al., 2016) for a hard sampling of the index k. This trick involves sampling a noise value g_k from the Gumbel distribution and then using the softmax function to normalize the output, resulting in a probability distribution. By selecting the index with the highest probability, we obtain a discrete choice. This entire process can be described as follows:

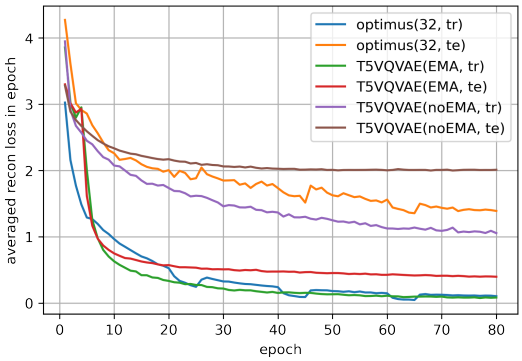

Figure 2: Loss curves of T5VQVAEs (base) with and without EMA and Optimus on the WorldTree corpus.

$$z_{w_i} = z_k, \text{ where}$$
$$k = \text{argmax}_k \frac{\exp(\log(t_k) + g_k)/\tau}{\sum_{k=1}^{K} \exp(\log(t_k) + g_k)/\tau}$$

In this context, t_k represents the probability of the k-th token, which can be obtained through a linear transformation before being fed into the Gumbel softmax. The parameter τ serves as a temperature hyper-parameter that controls the closeness of the new distribution to a discrete distribution. As τ approaches zero, the distribution becomes one-hot, while a non-zero value of τ leads to a more uniform distribution. In our experiments, we experienced convergence issues when using the Gumbel softmax scheme, and therefore decided to adopt the k-means mechanism which generally leads to better results.

Advantages of T5VQVAE. Compared with state-of-the-art Transformer VAEs such as Optimus (Li et al., 2020), our model has the following architectural advantages: (i) efficient and stable latent space compression. During the training of Optimus, in fact, the KL term in ELBO is regularized cyclically (Fu et al., 2019) to avoid KL vanishing and posterior collapse, which leads to an unstable training process (figure 2). In contrast, T5VQVAE avoid the KL regularization term since it becomes a constant value:

$$\text{KL}\left(q(z_k|x)||p(z_k)\right) = \sum_k q(z_k|x) \log \frac{q(z_k|x)}{p(z)}$$
$$= 1 \times \log \frac{1}{1/K} = \log K$$

where the prior $p(z) = 1/K$ is a uniform distribution. (ii) Better controllability. Hu et al. (2022)

revealed that in Optimus (Li et al., 2020), the latent representation is concatenated into key and value which is more likely to be ignored by most attention heads especially in lower layers where lexical-level semantics is captured. In contrast, the latent representations of T5VQVAE are designed to act on the attention heads directly.

3 Controllability Evaluation

Next, we put forward two metrics for quantitatively evaluating the controllability of the proposed model (T5VQVAE), which we refer to as *semantic disentanglement* and *interpolation smoothness*. The former evaluates the controllability from the perspective of disentanglement of semantic factors (e.g., arguments and associated semantic roles). The latter evaluates the smoothness and coherence of the latent space geometry during interpolation.

3.1 Semantic Disentanglement

Recent studies have attempted to adapt metrics from the image domain to evaluate the semantic disentanglement of sentences (Zhang et al., 2022; Carvalho et al., 2023). Semantic information in a sentence is more likely to be entangled, especially in the context of stacked multi-head self-attention models. As mentioned in (Zhang et al., 2022; Carvalho et al., 2023), conceptually dense sentences are clustered according to role-content combination over the VAE latent space. Each semantic role is jointly determined by multiple dimensions rather than one single dimension. Therefore, calculating the importance of one dimension to that semantic role as a disentanglement metric is unreliable. In this work, we quantitatively evaluate the disentanglement of the semantic roles by: (1) calculating the averaged Euclidean distance between different content under that role, such as the distance between *PRED-is* and *PRED-are*, and (2) counting the number of different indices of the same role-content after the vector quantisation. The smaller the distance or the less the number of indices, the more concentrated the distribution of this semantic role in the latent space, indicating better disentanglement.

3.2 Interpolation Smoothness

Interpolation is a standard process for evaluating the geometric properties of a latent space in both image and language domains (Li et al., 2020; Liu et al., 2021). It aims to generate a sequence of sentences following a spatial trajectory from source to target via latent arithmetics. For example, in the VAE latent space, the interpolation path can be described as $z_t = z_1 \cdot (1-t) + z_2 \cdot t$ with t increased from 0 to 1 by a step size of 0.1 where z_1 and z_2 represent latent vectors of source and target sentences, respectively. In this case, each intermediate output $D(z_t)$ should change fewer semantic concepts at each step if the latent space is smooth and regular. In this work, we employ a similar strategy, however follow the more granular token level within the VQVAE. We directly manipulate the interpolation within the latent token space. At each step t, we obtain the intermediate latent token embedding $z_t^{w_i}$ within a sentence by calculating the weighted minimal distance between its preceding token embedding $z_{t-0.1}^{w_i}$ and the target token embeddings $z_2^{w_i}$. This process can be described as follows:

$$
\begin{aligned}
&z_1^{w_i} = e^{k_1}, z_2^{w_i} = e^{k_2}, \text{where } \; i = [1, ..., L] \\
&z_t^{w_i} = z^k, \text{where} \\
&\quad k = \text{argmin}_j \; (1-t) \times \left\| z_{t-0.1}^{w_i} - z^j \right\|_2 \\
&\qquad + t \times \left\| z_2^{w_i} - z^j \right\|_2 \\
&\quad s_t = [z_t^{w_1}; \ldots; z_t^{w_L}]
\end{aligned}
$$

where s_t represents the sentence embeddings at step t. The final generated sentence can be decoded as $s_t = D(s_t)$. Once we have obtained the interpolation path, we introduce the interpolation smoothness (IS) metric to quantitatively evaluate its smoothness. This metric involves calculating the aligned semantic distance between the source and the target (referred to as the ideal semantic distance). Subsequently, we calculate the sum of the aligned semantic distances between each pair of adjacent sentences in the path (referred to as the actual semantic distance). Finally, by dividing the ideal semantic distance by the actual semantic distance, we obtain a measure of smoothness. If the result is 1, it indicates that the actual path aligns perfectly with the ideal path, suggesting better geometric properties. Conversely, it suggests a less coherent transformation path, indicating poorer geometric properties. The metric is defined as follows:

$$
\text{IS} = \mathbb{E}_{(s_0,\ldots,s_T)\sim P} \frac{\delta(\text{align}(s_0, s_T))}{\sum_{t=0}^{T} \delta(\text{align}(s_t, s_{t+0.1}))}
$$

where δ and align are sentence similarity and alignment functions, respectively. In this experiment, sentence similarity and alignment are performed

via Word Mover's Distance (Zhao et al., 2019) since it can softly perform the semantic alignment.

4 Experiments

4.1 AutoEncoding Task

Pre-training Data. In this work, we focus on the use of conceptually dense explanatory sentences (Dalvi et al., 2021) and mathematical latex expressions (Meadows et al., 2023b) to evaluate model performance. The rationale behind this choice is that (1) explanatory sentences provide a semantically challenging yet sufficiently well-scoped scenario to evaluate the syntactic and semantic organisation of the space (Thayaparan et al., 2020; Valentino et al., 2022a,b); (2) mathematical expressions follow a well-defined syntactic structure and set of symbolic rules that are notoriously difficult for neural models (Meadows et al., 2023a). Moreover, the set of rules applicable to a mathematical expression fully determines its semantics, allowing for an in-depth inspection and analysis of the precision and level of generalisation achieved by the models (Welleck et al., 2022; Valentino et al., 2023). Firstly, we conduct a pre-training phase, evaluating the performance of T5VQVAE in reconstructing scientific explanatory sentences from WorldTree (Jansen et al., 2018) and mathematical latex expressions from the dataset proposed by Meadows et al. (2023b).

Baselines. We consider both *small* and *base* versions of pretrained T5 to initialise the T5VQVAE, where the codebook size is 10000. The effect of different codebook sizes on its performance and the optimal point within the architecture (different hidden layers of the encoder) to learn the codebook are reported in Table 11. As for the large VAE model, we consider Optimus with random initial weights and pre-trained weights (Li et al., 2020) and Della (Hu et al., 2022). We chose two different latent dimension sizes (32 and 768) for both of them. Moreover, we also select several LSTM language autoencoders (AE), including denoising AE (Vincent et al. (2008), DAE), β-VAE (Higgins et al., 2016), adversarial AE (Makhzani et al. (2015), AAE), label adversarial AE (Rubenstein et al. (2018), LAAE), and denoising adversarial autoencoder (Shen et al. (2020), DAAE). Additional details on the training setup are provided in Appendix A. The full source code of the experimental pipeline is available at an anonymised link for reproducibility purposes.

Explanatory sentences					
Evaluation Metrics	BLEU	BLEURT	Cosine	Loss ↓	PPL ↓
DAE(768)	**0.74**	**0.03**	**0.91**	**1.63**	**5.10**
AAE(768)	0.35	-0.95	0.80	3.35	28.50
LAAE(768)	0.26	-1.07	0.78	3.71	40.85
DAAE(768)	0.22	-1.26	0.76	4.00	54.59
β-VAE(768)	0.06	-1.14	0.77	3.69	40.04
Optimus(32, rand)	0.54	0.14	0.92	1.08	2.94
Optimus(32, pre)	0.61	0.29	0.93	0.86	2.36
Optimus(768, rand)	0.49	-0.04	0.90	1.32	3.74
Optimus(768, pre)	0.68	0.48	0.95	0.65	1.91
DELLA(32, rand)	0.71	0.06	0.92	0.50	1.65
DELLA(768, rand)	0.72	0.21	0.95	**0.41**	**1.51**
T5VQVAE(small, soft)	0.81	**0.62**	**0.97**	0.46	1.58
T5VQVAE(base, soft)	**0.82**	**0.62**	**0.97**	0.75	2.11
Mathematical expressions					
Evaluation Datasets	EVAL	VAR	EASY	EQ	LEN
DAE(768)	**0.94**	**0.50**	**0.80**	**0.74**	**0.58**
AAE(768)	0.41	0.41	0.39	0.41	0.52
LAAE(768)	0.41	0.45	0.39	0.39	0.49
DAAE(768)	0.38	0.48	0.35	0.38	0.49
β-VAE(768)	0.39	0.48	0.37	0.39	0.50
Optimus(32, rand)	0.95	0.59	0.75	0.71	0.50
Optimus(768, rand)	0.96	0.61	0.79	0.75	0.54
DELLA(32, rand)	**1.00**	0.55	0.89	0.72	0.63
DELLA(768, rand)	**1.00**	0.55	0.93	0.79	0.64
T5VQVAE(small, soft)	0.97	**0.65**	**0.95**	**0.90**	**0.69**
T5VQVAE(base, soft)	0.98	0.62	**0.95**	0.85	0.68

Table 1: AutoEncoding task evaluation on the test set (soft: k-means). The highest scores of large VAE models and LSTM-based VAE models are highlighted in blue and in bold separately.

Quantitative Evaluation. As for modelling explanatory sentences, we quantitatively evaluate the performance of the models using five metrics, including BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), cosine similarity from pretrained sentence T5 (Ni et al., 2022), cross-entropy (Loss), and perplexity (PPL). As for modelling mathematical expressions, we use BLEU to evaluate the robustness of models on the 5 test sets proposed by Meadows et al. (2023b), one designed to assess in-distribution performance, and four designed to assess out-of-distribution generalisation. Here we provide a full characterisation of the test sets: (1) EVAL: contains mathematical statements following the same distribution of the training set (like $U + cos(n)$), including expressions with similar lengths and set of symbols (2) VAR: full mathematical statements with variable perturbations (like $U + cos(beta)$), designed to test the robustness of the models when dealing with expressions containing variables never seen during training; (3) EASY: simpler mathematical expressions with a lower number of variables, designed to test length generalisation (like $cos(n)$), (4) EQ: full mathematical statements with equality insertions (like $E = U + cos(n)$), designed to test the behaviour of

Role-content	NUM centers	AVG dis	MAX dis	MIN dis
ARG0-animal	3	0.28	0.52	0.35
ARG1-animal	3	0.28	0.52	0.35
ARG2-animal	4	0.33	0.55	0.35
PRED-is	24	0.60	1.08	0.22
PRED-are	6	0.31	0.64	0.21
MOD-can	5	0.40	0.82	0.28
NEG-not	2	0.25	0.51	0.51

Table 2: Semantic role disentanglement.

the model on equivalent mathematical expressions with minimal perturbations (5) LEN: mathematical statements with a higher number of variables (like $U + cos(n)) + A + B$), designed to test generalisation on more complex expressions.

As shown in Table 1, the highest scores for large VAE models and LSTM-based VAE models are highlighted in blue and bold, respectively. Among them, T5VQVAEs with the k-means scheme outperforms Optimus and LSTM-based VAEs in both corpora and compared with Della, it can deliver better generation and generalization. We provide examples with low BLEURT scores in Appendix C

Next, we quantitatively evaluate the disentanglement of T5VQVAE following the semantic disentanglement reference metric 3.1. As displayed in Table 2, the number of central points for *PRED* is higher than the remaining role-content, being 24 in *PRED-is* and 6 in *PRED-are*. This indicates that the semantic information of *PRED* is more widely distributed in the latent space when compared to other roles. This behaviour might be attributed to the fact that the aforementioned predicates are widely used across sentences in the corpus. The full visualisation of the semantic disentanglement achieved by T5VQVAE is provided in Figure 3.

4.2 Text Transfer Task

Next, we investigate the controllability of T5VQVAE by manipulating the latent space via geometric transformations. This is referred to as the Text Transfer task. We compare the performance of T5VQVAE (base, soft) and Optimus (32, pretrain) - both trained in the AutoEncoding task - as baselines. We evaluate the latent space using latent traversal, interpolation, and vector arithmetics.

Latent Traversal. The traversal is inspired by the image domain, only changing the feature interpretation (Higgins et al., 2017; Kim and Mnih, 2018). Specifically, if the vector projection within the latent space can be modified when traversing (re-sampling) one dimension, the output should only change well-defined semantic features corresponding to that dimension. In this experiment, the traversal is set up from a starting sentence. As illustrated in Table 3, the T5VQVAE can provide localised semantic control by operating the discrete latent space. Different dimensions in the discrete sentence space can control different parts of the sentence. The traversal for Optimus is provided in Appendix D.

Latent Interpolation. As described in section 3.2, interpolation aims to generate a sequence of sentences from source to target via latent vector arithmetic. An ideal interpolation should lead to reasonable semantic controls at each step. In Table 4, we can observe that compared with Optimus's interpolation (bottom) where the semantics are changed redundantly, e.g., from *some birds* to *some species mammals* to *most birds* and from *have* to *don't have* to *have*, T5VQVAE (top) leads to a more reasonable (coherent/smoother) pathway. E.g., from *speckled brown color* to *speckled brown feathers* to *speckled wings* to *wings*. Additional examples are provided in Appendix D.

More importantly, we quantitatively evaluate the interpolation behaviour via the IS metric. We randomly select 100 (source, target) pairs and interpolate the path between them. Then, we calculate the averaged, maximal, and minimal ISs. As shown in Table 5, T5VQVAE outperforms Optimus by over 43% in average, which indicates that T5QVAE induces a latent space which can better separate the syntactic and semantic factors when contrasted to Optimus.

Latent Vector Arithmetics. Inspired by word embedding arithmetics, e.g., $king - man + woman = queen$, we explore the compositional semantics via latent arithmetic with the target of sentence-level semantic control. After adding two latent vectors corresponding to two sentences $s_c = s_A + s_B$, we expect the resulting sentence to express the semantic information of both sentences. From Table 6, we can observe that T5VQVAE can generate the outputs containing both inputs' semantic information. E.g., the output contains *are likely to* and *their environment* from s_A and *to survive* and / from s_B. In contrast, Optimus is not able to preserve to support this behaviour. Additional examples are provided in Appendix D (Table 16).

an animal requires warmth in cold environments

dim0: an animal requires warmth in cold environments
dim0: a animal requires warmth in cold environments
dim0: the animal requires warmth in cold environments

dim1: an organism requires warmth in cold environments
dim1: an animal requires warmth in cold environments
dim1: an object requires warmth in cold environments

dim2: an animal needs warmth in cold environments
dim2: an animal must find warmth in cold environments
dim2: an animal brings warmth in cold environments
dim2: an animal wants warmth in cold environments

dim4: an animal requires warmth during cold temperatures
dim4: an animal requires warmth in cold environments
dim4: an animal requires warmth to cold environments

dim5: an animal requires warmth in temperatures
dim5: an animal requires warmth in warm environments
dim5: an animal requires warmth in a warm environment

dim6: an animal requires warmth in cold temperatures
dim6: an animal requires warmth in cold climates
dim6: an animal requires warmth in cold systems

Table 3: T5VQVAE(base): traversals showing controlled semantic concepts in explanations. We also provide the traversal of Optimus latent space for comparison in Table 13.

Source: some birds have a speckled brown color

1. some birds have a speckled brown color
2. some birds do not have speckled brown feathers
3. some species mammals do not have speckled wings
4. most species mammals do not have wings

1. some birds have scales
2. some birds have a speckled brown color
3. some species mammals have wings
4. most birds don't have wings
5. most insects have wings
6. most species mammals don't have wings

Target: most species mammals do not have wings

Table 4: Interpolation for T5VQVAE (top) and Optimus (bottom) where blue, underline, and orange represent subject, verb, and object, respectively. Only unique sentences are shown.

Evaluation Metrics	avg IS	max IS	min IS
Optimus(32, pretrain)	0.22	0.53	0.13
Optimus(768, pretrain)	0.21	0.50	0.10
T5VQVAE(base, soft)	**0.65**	**1.00**	**0.18**

Table 5: Interpolation smoothness.

4.3 Inference Task

Lastly, we move to downstream inference tasks, in which we aim to explore the controllability of T5VQVAE for reasoning with natural and symbolic languages. Specifically, we focus on two tasks including syllogistic-deductive natural language inference in EntailmentBank (Dalvi et al., 2021), where a natural language conclusion has to be inferred from two premises, and mathematical expression derivation (Meadows et al., 2023b), where the goal is to predict the result of applying a mathematical operation to a given premise expression (written in latex).

s_A: animals are likely to have the same color as their environment
s_B: animals require respiration to survive / use energy

T5VQVAE: animals are likely to survive / to survive in their environment
Optimus: animals have evolved from animals with traits that have an animal instinct

Table 6: Latent arithmetic $s_A + s_B$ for T5VQVAE(base) and Optimus(32). blue, orange, and shallow blue indicate the semantic information from both s_A and s_B, from s_A only, from s_B only, respectively.

Quantitative Evaluation. We quantitatively evaluate several baselines following the same procedure as the AutoEncoding task. Table 7 shows that T5VQVAE outperforms all VAE models on both benchmarks.

Qualitative Evaluation. Next, we focus on the NLI task to explore the controllability of T5VQVAE for sentence-level inference traversing the latent space. As illustrated in Table 8, traversing the dimension corresponding to an individual word (e.g., *object* from premise 1 (P1)) cannot preserve the target word during the traversal along with the semantic coherence of the transitions, indicating that the inference is done entirely in the Encoder. Therefore, we next explore how to manipulate the latent representation to deliver a more controllable

Natural Language Inference (EntailmentBank)					
Evaluation Metrics	BLEU	Cosine	BLEURT	Loss ↓	PPL ↓
T5(small)	0.54	0.96	0.22	0.69	1.99
T5(base)	**0.57**	**0.96**	**0.33**	**0.61**	**1.84**
Bart(base)	0.54	0.96	0.17	0.63	1.87
FlanT5(small)	0.22	0.89	-1.33	0.99	2.69
FlanT5(base)	0.32	0.89	-0.31	0.95	2.58
T5bottleneck(base)	0.35	0.91	-0.20	1.24	3.45
Optimus(32)	0.07	0.74	-1.20	1.13	2.31
Optimus(768)	0.08	0.74	-1.21	0.82	2.27
DELLA(32)	0.08	0.85	-1.23	1.69	5.41
DELLA(768)	0.09	0.87	-1.09	1.54	4.66
T5VQVAE(small)	0.11	0.73	-1.23	0.85	2.33
T5VQVAE(base)	**0.46**	**0.94**	**0.10**	**0.84**	**2.31**
Mathematical Expression Derivation					
Evaluation Datasets	EVAL	SWAP	EASY	EQ	LEN
T5(small)	0.69	0.48	0.57	0.60	0.63
T5(base)	0.97	0.65	0.90	0.72	0.81
Optimus(32)	0.72	0.50	0.59	0.23	0.40
Optimus(768)	0.79	0.56	0.63	0.29	0.44
DELLA(32)	0.12	0.16	0.13	0.13	0.13
DELLA(768)	0.13	0.18	0.12	0.13	0.14
T5VQVAE(small)	0.75	**0.57**	0.77	**0.48**	**0.50**
T5VQVAE(base)	**0.76**	0.56	**0.78**	0.47	**0.50**

Table 7: Quantitative evaluation on inference tasks.

P1: a human is a kind of object
P2: a child is a kind of young human
C: a child is a kind of object

dim6: a young object is a kind of child
dim6: a boy is a kind of young object
dim6: a little boy is a kind of young human

Table 8: T5VQVAE (base): traversed conclusions.

inference behaviour.

Recent work (Zhang et al., 2023c) has provided a granular annotated dataset of step-wise explanatory inference types, which reflect symbolic (syllogistic-style) operations between premises and conclusions, including *argument/verb substitution*, *further specification*, and *conjunction*. We leverage this annotation to input two premises into the Encoder to derive the latent token embeddings of individual arguments and guide the generation of the conclusion via the Decoder. For example, for *argument substitution* and *verb substitution*, which refers to the process of obtaining a conclusion by substituting one argument/verb from the first premise to an argument/verb of the second premise, we substitute the respective token embeddings in the latent space and feed the resulting representation to the decoder. Table 9 shows that by substituting the embeddings of the arguments, we can control the behaviour of the model and elicit a systematic inference behaviour. We provide *further specification* and *conjunction* in Table 18. These results show that the latent embeddings can be manipulated to deliver a syllogistic-style inference behaviour. In particular, we demonstrate that the distributed semantic information in the latent space contains information about co-occurring tokens within the sentence that can be systematically localised (within specific arguments, predicates or clauses) and manipulated to generate a sound conclusion. This behaviour can be potentially leveraged as a foundation to build an interpretable and multi-step natural language inference model. More examples are reported in the Appendix E.

P1: a shark is a kind of fish
P2: a fish is a kind of aquatic animal
Pred: a shark is a kind of aquatic animal

P1: to move something can mean to transfer something
P2: flowing is a kind of movement for energy
Pred: flowing is a kind of transfer of energy

Table 9: T5VQVAE(base): quasi-symbolic inference examination in AutoEncoder (Top: argument substitution, Bottom: Verb substitution).

5 Related work

Semantic Control via Latent Spaces. Zhang et al. (2022, 2023a) investigated the semantic control of latent sentence spaces, demonstrating the basic geometric-semantic properties of VAE-based models. Mercatali and Freitas (2021) defined disentangled latent spaces focusing on the separation between content and syntactic generative factors. Moreover, some works focused on defining two separate latent spaces to control natural language generation on specific downstream tasks, such as style-transfer and paraphrasing (Bao et al., 2019a; John et al., 2019a). Comparatively, this work explores more granular control and a broader spectrum of tasks: from syllogistic to symbolic inference.

Language VAEs. Instead of Optimus (Li et al., 2020) and its variation (Fang et al., 2022; Hu et al., 2022) where the encoder and decoder are BERT and GPT2, respectively, most of the language VAE literature are based on LSTM architectures instantiated on different text generation tasks, including story generation (Fang et al., 2021), dialogue generation (Zhao et al., 2017), text style transfer

(John et al., 2019a; Shen et al., 2020), text paraphrasing (Bao et al., 2019a), among others. Some works also investigated different latent spaces or priors to improve representation capabilities (Dai et al., 2021; Ding and Gimpel, 2021; Fang et al., 2022). Comparatively, this work contributes by focusing on the close integration between language models and vector-quantized VAE-driven granular control, instantiating it in the context of a state-of-the-art, accessible, and cross-task performing language model (T5).

6 Conclusion and Future Works

In this work, we build a model for improving the semantic and inference control for VAE-enabled language model (autoencoding) architectures. We propose a new model (i.e., T5VQVAE) which is based on the close integration of a vector-quantized VAE and a consistently accessible and high-performing language model (T5). The proposed model was extensively evaluated with regard to its syntactic, semantic and inference controls using three downstream tasks (autoencoding, text transfer, and inference task). Our experimental results indicate that the T5VQVAE can outperform the canonical state-of-the-art models in those tasks and can deliver a quasi-symbolic behaviour in the inference task (via the direct manipulation of the latent space).

As future work, we plan to further explore applications on symbolic natural language inference via the direct manipulation of the latent space, and to investigate the controllability of recent large language models through the VQVAE architecture. Moreover, additional research directions could be informed by the current work:

Word-level Disentanglement. Our architecture provides a foundation to explore token/word-level disentanglement for more general sentence and inference representation tasks. While sentence-level disentanglement is widely explored in the NLP domain, such as sentiment-content (John et al., 2019b; Hu and Li, 2021), semantic-syntax (Bao et al., 2019b; Zhang et al., 2023d), and negation-uncertainty (Vasilakes et al., 2022), or syntactic-level disentanglement (Felhi et al., 2022), this mechanism is still under-explored in other NLP tasks (Liao et al., 2020).

Interpretability. Discrete properties derived from vector quantization can enable the further probing and interpretability of neural networks by discretizing continuous neural latent spaces, where symbolic concepts are emerging in both images (Deng et al., 2021; Li and Zhang, 2023) and natural language (Tamkin et al., 2023) domains.

Limitations

While T5VQVAE can improve inference performance and deliver inference control on syllogistic-deductive style explanations, the application on more complex reasoning tasks (e.g. involving quantifiers and multi-hop inference) is not fully explored. Besides, we still observe limitations in out-of-distribution generalisation in the mathematical expressions corpus despite the improvement over existing VAE models in terms of robustness. This, in particular, is highlighted by the decrease in performance obtained on the length generalisation split (LEN) for both autoencoding and expression derivation tasks.

Acknowledgements

We appreciate the reviewers for their insightful comments and suggestions. This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617), by the EPSRC grant EP/T026995/1 entitled "EnnCore: End-to-End Conceptual Guarding of Neural Architectures" under Security for all in an AI enabled society, by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.

References

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

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

Danilo S Carvalho, Giangiacomo Mercatali, Yingji Zhang, 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)*.

Shuyang Dai, Zhe Gan, Yu Cheng, Chenyang Tao, Lawrence Carin, and Jingjing Liu. 2021. APo-VAE: Text generation in hyperbolic space. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 416–431, Online. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. 2021. Discovering and explaining the representation bottleneck of dnns. *arXiv preprint arXiv:2111.06236*.

Xiaoan Ding and Kevin Gimpel. 2021. FlowPrior: Learning expressive priors for latent variable sentence models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3242–3258, Online. Association for Computational Linguistics.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *arXiv preprint arXiv:2101.00828*.

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. 2022. Towards unsupervised content disentanglement in sentence representations via syntactic roles. *arXiv preprint arXiv:2206.11184*.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. 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 240–250, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson H S Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. A deep semantic natural language processing platform.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

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*.

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.

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

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

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

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

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019b. 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.

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*.

Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3603–3614, Hong Kong, China. Association for Computational Linguistics.

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.

Mingjie Li and Quanshi Zhang. 2023. Does a neural network really encode symbolic concept? *arXiv preprint arXiv:2302.13080*.

Keng-Te Liao, Cheng-Syuan Lee, Zhong-Yu Huang, and Shou-de Lin. 2020. Explaining word embeddings via disentangled representation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 720–725, Suzhou, China. Association for Computational Linguistics.

Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. 2021. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10785–10794.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Jordan Meadows, Marco Valentino, and Andre Freitas. 2023a. Generating mathematical derivations with large language models. *arXiv preprint arXiv:2307.09998*.

Jordan Meadows, Marco Valentino, Damien Teney, and Andre Freitas. 2023b. 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. *arXiv preprint arXiv:2109.07169*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. 2018. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*.

Paul K Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. 2018. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Alex Tamkin, Mohammad Taufeeque, and Noah D Goodman. 2023. Codebook features: Sparse and discrete interpretability for neural networks. *arXiv preprint arXiv:2310.17230*.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Marco Valentino, Jordan Meadows, Lan Zhang, and André Freitas. 2023. Multi-operational mathematical derivations in latent space. *arXiv preprint arXiv:2311.01230*.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022a. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2022b. Case-based abductive natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1556–1568, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA. Association for Computing Machinery.

Sean Welleck, Peter West, Jize Cao, and Yejin Choi. 2022. Symbolic brittleness in sequence models: on systematic generalization in symbolic mathematics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8629–8637.

Yingji Zhang, Danilo S Carvalho, Ian Pratt-Hartmann, and André Freitas. 2022. Quasi-symbolic explanatory nli via disentanglement: A geometrical examination. *arXiv preprint arXiv:2210.06230*.

Yingji Zhang, Danilo S Carvalho, Ian Pratt-Hartmann, and André Freitas. 2023a. Learning disentangled semantic spaces of explanations via invertible neural networks. *arXiv preprint arXiv:2305.01713*.

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

Yingji Zhang, Danilo S Carvalho, Ian Pratt-Hartmann, and Andre Freitas. 2023c. Towards controllable natural language inference through lexical inference types. *arXiv preprint arXiv:2308.03581*.

Yingji Zhang, Marco Valentino, Danilo S Carvalho, Ian Pratt-Hartmann, and André Freitas. 2023d. Graph-induced syntactic-semantic spaces in transformer-based variational autoencoders. *arXiv preprint arXiv:2311.08579*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Training setup

Datasets Table 10 displays the statistical information of the datasets used in the experiment. As for the AutoEncoder setup, we use the non-repetitive explanations selected from both datasets as the experimental data. As for the Inference task, we use the data from EntailmentBank and Math Symbol Inference. The semantic roles of our data are annotated by automatic semantic role labelling tool (Gardner et al., 2017).

Corpus	Num data.	Avg. length
WorldTree	11430	8.65
EntailmentBank	5134	10.35
Math Symbol	32000	6.84
Math Symbol Inference	32000	51.84

Table 10: Statistics from datasets.

T5VQVAE training We use T5VQVAE(small) to choose the most appropriate codebook size between 2000 and 22000. In the experiment, the maximal epoch is 100. The learning rate is 5e-5. We use exponential moving averages (EMA) to update the codebook. Besides, we also investigated the optimal point within the architecture to learn the codebook. As shown in Table 11, T5VQVAE performs better when the codebook is learned at the end of the Encoder. This observation suggests that cross-attention is crucial in vector quantisation (VQ) learning.

Metrics	BLEU	BLEURT	cosine	Loss ↓	PPL ↓
02000	0.73	0.21	0.93	0.79	2.20
06000	0.79	0.45	0.95	0.61	1.84
10000	0.81	0.62	0.97	0.46	1.58
14000	0.82	0.62	0.96	0.42	1.52
18000	0.83	0.64	0.96	0.38	1.46
22000	0.83	0.67	0.96	0.34	1.40
T5VQVAE(small) with different depth L in Encoder					
T5VQVAE(L=05)	0.47	-0.80	0.80	0.91	2.48
T5VQVAE(L=04)	0.59	-0.56	0.84	0.76	2.13
T5VQVAE(L=03)	0.65	-0.42	0.85	0.68	1.97
T5VQVAE(L=02)	0.70	-0.21	0.88	0.65	1.91

Table 11: T5VQVAE(small): Different sizes of codebook and optimal point.

Expotential Moving Average (EMA) Let $\{E(x_{k,1}), ..., E(x_{k,n_k})\}$ be the set of word embedding $x_{k,i}$ belonging to the z_k. The optimal value for z_k is the average of elements in this set, which can be described as:

$$z_k = \frac{1}{n_k} \sum_{i}^{n_k} E(x_i)$$

However, we cannot use this to update z_k since we usually work on mini-batches. Instead, we can use EMA to update z_k.

$$N_k^{(t)} := N_k^{(t-1)} \times \lambda + n_k^{(t)}(1-\lambda)$$
$$m_k^{(t)} := m_k^{(t-1)} \times \lambda + \sum_i E(x_{k,i})$$
$$z_k := \frac{m_k^{(t)}}{N_k^{(t)}}$$

Where λ is 0.99 following the setup of (Van Den Oord et al., 2017).

Optimus and DELLA training setup Both of them can be trained via the evidence lower bound (ELBO) on the log-likelihood of the data x (Kingma and Welling, 2013). To avoid KL vanishing issue, which refers to the Kullback-Leibler (KL) divergence term in the ELBO becomes very small or approaches zero, we select the cyclical schedule to increase weights of KL β from 0 to 1 (Fu et al., 2019) and KL thresholding scheme (Li et al., 2019) that chooses the maximal between KL and threshold λ. The final objective function can be described as follows:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] - \beta \max\left[\lambda, \text{KL} q_\phi(z|x)||p(z)\right]$$

B Visualization

In Figure 3, we visualise the latent space of T5VQVAE via t-distributed Stochastic Neighbor Embedding (T-SNE) (Van der Maaten and Hinton, 2008) to analyse the organization of key semantic clusters. Specifically, we visualize the clusters of token embeddings with the same role-content, different roles, and the same content with different roles, respectively. We can observe that under the same role-content (left), the latent token embeddings are widely distributed in the latent space as the representation of the role-content is affected by the context, which indicates poor disentanglement. For different roles (middle), there are big overlaps between different semantic roles, which indicates poor disentanglement of semantic role structure. For the same content with different roles (right), it can be observed that different semantic role clusters are fully overlapped. Those visualizations indicate that the semantic information is naturally entangled after an attention-based Encoder.

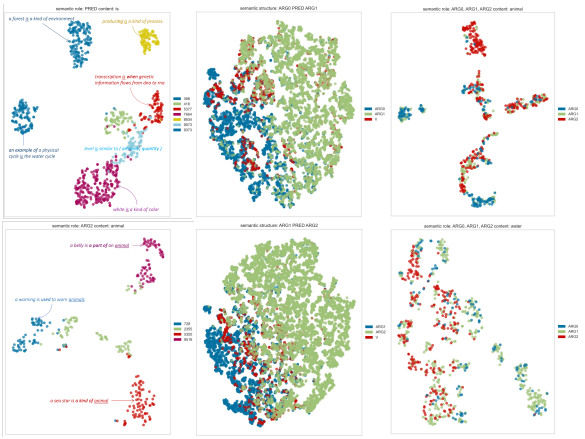

Figure 3: t-SNE plot of the T5VQVAE latent space. Left: same role-content(*PRED-is*, *ARG2-animal*). Middle: different role-content(*ARG0-PRED-ARG1*, *ARG1-PRED-ARG2*). Right: different roles with same content (*ARG0, 1, 2 - animal*, *ARG0, 1, 2 - water*).

C AutoEncoding Task

We provide more reconstructed explanations with low BLEURT scores in Table 12. we manually evaluate its performance and show the common issues in the AutoEncoding setup. (1) repetition: some explanations that describe the synonym are suffered from information loss. E.g., the prediction is *the grand canyon is a kind of canyon* where the golden is *the grand canyon is a kind of place*. (2) wrong numerical token: the model cannot precisely reconstruct the numerical token. E.g., *the speed of the boat can be calculated by dividing the length of a boat* compared with the golden: *the speed of the sailboat can be calculated by dividing 35 by 5*.

D Text Transfer Task

We provide more traversal, interpolation, and arithmetic examples in Tables 13,14, 15, and 16.

E Inference Task

We provide more examples in Tables 17 and 18.

Golden Explanations	Predicted Explanations	BLEURT	BLEU
the grand canyon is a kind of place	the grand canyon is a kind of canyon	0.26	0.87
a blood thinner can be used to treat people with heart attacks and strokes	a heart thinner can be used to treat people with blood and heart	-0.05	0.44
the plant offspring has yellow flowers	offspring means offspring	-1.30	0.12
lack is similar to (low ; little)	little means (little ; little) in quality	-1.18	0.44
preserved means (from the past ; from long ago)	preserved means used to be (preserved ; preserved) from a long time	-0.01	0.50
the plant offspring has yellow flowers	offspring means offspring	-1.30	0.12
electricity causes less pollution than gasoline	gasoline causes less gasoline than gasoline	-0.22	0.66
insulin is a kind of hormone	insulin is made of insulin	-0.31	0.49
living things all require a producers for survival	living things all require a living thing for survival	0.03	0.77
gravity causes nebulas to collapse	gravity causes a sleef of an artery to collapse	-1.30	0.44
out is synonymous with outside	outward is synonymous with out	-0.36	0.80
to prevent means to make it not happen	to make means to not happen	-0.74	0.71
a branch is a kind of object	a branch is a kind of branch	-0.03	0.85
force requires energy	force means amount	-0.40	0.33
spot means location	place means kind of place	-0.14	0.20
gritty is similar to rough	grease is similar to grease	-0.80	0.60
sidewalk means pavement	bike means bike	-0.62	0.33
a gravel pit is a kind of environment	a gravel pit is a kind of gravel	0.03	0.87
a electron has a negative (-1) electric charge	a electron has a negative (electric charge ; negative charge)	0.23	0.75
fish is a kind of meat	fish are a kind of fish	-0.29	0.66
jogging is similar to running	running is a kind of running	-0.23	0.33
the speed of the sailboat can be calculated by dividing 35 by 5	the speed of the boat can be calculated by dividing the length of a boat	0.20	0.60
if an object has 0 mechanical energy then the object will stop moving	if an object has a mechanical energy then the object has to move to 0	0.09	0.66

Table 12: T5VQVAE(base): more examples with low BLEURT score.

Traversal

an animal requires warmth in cold environments

dim0: animals usually maintain a safe distance from predators during the hibernation process
dim0: animals usually require warmth in cold temperatures for survival
dim0: animals must sense prey to survive / find food
dim0: animals must sense food to survive in the cold environment

dim1: animals must protect themselves (against predators ; from predators)
dim1: animals with pacemakers must sense danger in order to eat prey
dim1: animals with sensory organs provided shelter in cold environments
dim1: animals with diabetes should be protected from predators in the water

dim2: animals must sense (predators ; food) to survive
dim2: animals must sense other animals for food / shelter
dim2: animals must sense other animals for survival in cold environments
dim2: animals with circulatory system have a positive impact on themselves by breathing air

dim4: animals with cold cardiovascular systems can survive in cold environments by breathing
dim4: animals must sense prey to survive in cold environments
dim4: animals must sense other animals for survival while they are at sea; in an environment
dim4: animals usually nurse their offspring through the winter

dim5: animals must sense prey to survive and reproduce
dim5: animals must sense food to find food
dim5: animals must sense prey in order to survive survival in the cold environment
dim5: animals require warmth in cold environments to (survive ; find food)

dim6: animals must sense food in order to survive in cold environments
dim6: animals must sense prey in order to survive / find food
dim6: animals with heat - circulatory system must cool themselves in cold environments
dim6: animals must sense prey to survive in cold environments

Table 13: Traversal for Optimus latent space.

Traversal

an astronaut requires the oxygen in a spacesuit backpack to breathe

dim1: an astronaut requires the oxygen in a spacesuit backpack to breathe
dim1: an organism requires the oxygen in a spacesuit backpack to breathe
dim1: an animal requires the oxygen in a spacesuit backpack to breathe
dim1: an student requires the oxygen in a spacesuit backpack to breathe

dim2: an astronaut requires the oxygen in a spacesuit backpack to breathe
dim2: an astronaut can wear the oxygen in a spacesuit backpack to breathe
dim2: an astronaut requires the oxygen in a spacesuit backpack to breathe
dim2: an astronaut requires the oxygen in a spacesuit backpack to breathe

dim1: astronauts wear spacesuits in the space station to avoid the issue of heat loss after a space probe
dim1: astronauts wear spacesuits in the space environment to protect the astronaut from harmful chemical reactions
dim1: astronauts wear spacesuits in the space station to keep the body warm
dim1: astronauts wear spacesuits in the spacesuit worn by the astronauts to take in oxygen

dim2: astronauts wear spacesuits in the space station in space
dim2: astronauts conducting the orbit of the moon in space during the last stage of a lunar cell might cause direct sunlight to lands on the moon
dim2: astronauts wear on the body the oxygen in a spacesuit backpack after the spacecraft escapes the atmosphere
dim2: astronauts wear spacesuits in the space station to protect the body of an astronaut

Table 14: Traversal comparison (left: T5VQVAE(base), right: Optimus).

Traversal
pedals are a kind of object
dim0: pedals are a kind of pedal
dim0: pedaling is a kind of object
dim0: a pedal is a kind of object
dim0: leather is a kind of object
dim1: a pedal is a kind of object
dim1: pedals are a kind of object
dim1: pedals are a kind of object
dim1: a pedal is a kind of object
dim0: objects are a kind of kind of nonliving thing
dim0: rust is a kind of object
dim0: objects are a kind of kind of heavy object
dim0: rust is a kind of object
dim1: objects are a kind of kind of nonliving thing
dim1: rust is a kind of object
dim1: bones are a kind of object
dim1: objects are a kind of kind of small particle
travel means to move
dim2: travel means move
dim2: travel is similar to move
dim2: travel is used to move
dim2: travel is a kind of movement
dim3: travel means to move
dim3: travel means stay
dim3: travel means to withstand travel
dim3: travel means to be transported
dim2: to move means to move
dim2: to pedal means to move something faster
dim2: to move means to move
dim2: to move means to move
dim3: to raise means to move something
dim3: to pedal means to move faster
dim3: to move means to move
dim3: to pedal means to move quickly

Table 15: Traversal comparison (top: T5VQVAE(base), bottom: Optimus). We can observe that T5VQVAE can provide better semantic control than Optimus.

Arithmetic
x_A**: a forest is a kind of land**
x_B**: a tornado is narrow in width**
T5VQVAE: a tornado is small in land
Optimus: plants are a kind of resource
x_A**: a rabbit is a kind of animal that may live in a meadow** x_B**: december is during the winter in the northern hemisphere**
T5VQVAE: december is a kind of animal that may be in a winter
Optimus: a animal can usually find something to eat
x_A**: fossil fuels are formed from dead prehistoric organisms** x_B**: orange is a kind of color**
T5VQVAE: orange fossil fuels are formed from dead prey
Optimus: prehistoric organisms developed defenses against disease by compacting and burying large amounts of remains
x_A**: waves travel outward from the source** x_B**: water is made of matter**
T5VQVAE: water points away from the source
Optimus: transverse waves cause the person to move perpendicular to the direction of the wave
x_A**: rotation is a kind of motion** x_B**: Leo is a kind of constellation**
T5VQVAE: Leo is a kind of motion
Optimus: friction occurs when two object colliding causes the speed of their movement to increase
x_A**: the milky way is a kind of galaxy** x_B**: a rock is usually a solid**
T5VQVAE: the milky way is usually a solid
Optimus: x -sex cells are inherited characteristics

Table 16: Addition Arithmetic comparison. We can observe that the T5VQVAE can hold the semantic information of both sentences after addition arithmetic.

Argument and Verb substitution

P1: heat is a kind of energy
P2: flowing can be a kind of transfer of energy
Pred: flowing can be a kind of transfer of heat

P1: aluminum is always nonmagnetic
P2: aluminum cans are made of aluminums
Pred: aluminum cans are nonmagnetic

P1: ground water is a kind of water
P2: a desert environment is low in availability of water / availability of food
Pred: a desert environment is low in availability of ground water / availability of food

P1: marine fossils are fossils of water animals
P2: marine fossils are found in mountains
Pred: marine fossils are fossils of marine animals
Gold: fossils of water animals are found in mountains

P1: in order to breath air you must take in air
P2: a lung is a kind of organ for breath air
Pred: a lung is a kind of organ for taking in air

P1: running is a kind of movement
P2: an animal requires energy to move
Pred: an animal requires energy to run

P1: eating is a kind of method for consuming
P2: decomposers obtain nutrients by eating waste
Pred: decomposers obtain nutrients by consuming waste

P1: rolling on a surface is a kind of moving on a surface
P2: the rubber ball is rolling across a surface
Pred: the rubber ball is moving across a surface

Table 17: T5VQVAE(base): quasi-symbolic inference examination in AutoEncoder (Top: argument substitution, Bottom: Verb substitution).

Further specification and Conjunction

P1: a plant requires energy to make food
P2: plants get energy from sunlight
Pred: plants get energy from sunlight to make food

P1: a mountain is made of rocks
P2: a mountain range contains mountains
Pred: a mountain range contains mountains made of rocks

P1: a dinosaur is a kind of animal
P2: dinosaurs are cold-blooded
Pred: dinosaurs are cold - blooded animals

P1: gravity causes orbits
P2: planets in the solar system orbit the sun
Pred: gravity causes planets in the solar system to orbit the sun

P1: scattering light decreases visibility
P2: water droplets scatter light
Pred: water droplets scattering light decreases air droplets
Gold: water droplets scattering light decreases the visibility

---

P1: sound travels slowest through gas
P2: sound travels fastest through solid
Pred: sound waves travel fastest through gas and gas
Gold: sound travels fastest through solid and slowest through gas

P1: matter in the solid phase has definite volume
P2: matter in the solid phase has definite shape
Pred: matter in the solid phase has definite shape and volume

P1: a plant requires nutrients from soil to grow
P2: a plant requires water from soil to grow
Pred: a plant requires water and nutrients from soil to grow

P1: a hurricane has high wind speed
P2: a hurricane has large amount of rain
Pred: a hurricane has high wind speed and large amount of rain

P1: fungi can be multicellular
P2: fungi have no chlorophyll
Pred: fungi have no chlorophyll and can be multicellular

Table 18: T5VQVAE(base): quasi-symbolic inference examination in AutoEncoder (Top: further specification, Bottom: conjunction).