Meaning Representations as Variational Quantum Circuits

Tilen G. Limbäck-Stokin, Tanishka A. Birdavade, Kin Ian Lo, Mehrnoosh Sadrzadeh
Quantum Learning Labs, University College London
London, United Kingdom
{tilen.limback-stokin.21, tanishka.birdavade.22, kin.lo.20, m.sadrzadeh}@ucl.ac.uk

Abstract

Large language and vision-language models (VLMs) struggle with a 'compositionality gap'. They treat language as a sequence of tokens lacking any structure and thus rely on a large number of parameters, making them computationally expensive. To address these issues, we propose **CCG-VQC**, a quantum framework that unifies statistical distributions with linguistic structure. Guided by Combinatory Categorial Grammar, our model maps syntactic rules into parametrised quantum circuits and models sentences as quantum states. We evaluate **CCG-VQC** on structural VLM benchmarks such as ARO and SVO-Swap. Our experiments show that **CCG-VQC** consistently outperforms a quantum bag-of-words model, as well as classical VLMs such as CLIP and OpenCLIP. **CCG-VQC** achieved 71.19% accuracy on ARO-Attribution, significantly outperforming the parameter-matched MicroCLIP, which struggled to surpass random chance with a maximum performance of 50.85%.

Keywords: Quantum Machine Learning, Syntax, Semantics, Vision-Language Models

1. Introduction

Large language models are thought to be one of the most successful learning algorithms of our age. They are applied to a wide range of domains, from weather forecasting (Li et al., 2025a) and medical modelling (Singhal et al., 2025) to theorem proving (Hubert et al., 2025) and scientific calculations in physics (Pan et al., 2025) and chemistry (Boiko et al., 2023). Although LLMs learn structural patterns present in language implicitly, they treat text as a series of tokens lacking explicit, controllable syntactic and semantic structure (Bender and Koller, 2020). Moreover, this is computed through self-attention, based on the statistics of co-occurrence, which is computationally intensive (Vaswani et al., 2023). This does not take full advantage of the generalisation that arises from the compositional nature of a grammatical rule set (Lake and Baroni, 2018).

Despite achieving high accuracies in generative tasks, they exhibit poor reliability and tend to hallucinate. Their success relies on a large number of parameters, making them energy inefficient and expensive. Moreover, it is not clearly known if LLMs are actually distilling their knowledge to generalise by learning fundamental rules governing the data, such as composition, or if they are just memorising at an immense scale, leading to a problem dubbed as the "compositionality gap" (Press et al., 2023; Li et al., 2025b; Ni et al., 2024).

The behaviour of LLMs is orthogonal to formal computational linguistic models that treat language as sequences of words generated according to the rules of grammar and abiding by structural, semantic and pragmatic constraints (Blackburn and Bos, 2005; Morrill, 2010; Steedman, 2000). The Achilles heel of these methods is their rigidity: they tend to only handle hand-picked examples and are inapplicable to large-scale, naturally occurring data. It is hard to fully describe, let alone formalise, all the rules of the grammar of a language. Formalising and reasoning about semantic and pragmatic constraints remains an open challenge.

A middle ground can be reached by developing meaning representations that unify statistical distributions of data with the linguistic structures embedded in it. Herein, semantic symbolic representations are assigned to syntactic structures via a homomorphic map. The symbolic representations are learnt in context, e.g. via labelled datasets or co-occurrence in corpora of text. Whereas an LLM uses statistical methods and learns vectors for tokens, the unified models learn higher-order linear maps informed by syntactic structure. For instance, the symbolic representations of a transitive sentence are a multilinear map modelling the predicative meaning of the verb, which then applies to the meaning representations of its subject and object, which are vectors.

In finite dimensions, learning a multilinear map is equivalent to learning a tensor. This amounts to training a multi-dimensional array, which suffers from exponential scaling. In quantum computation, tensors are treated as states of quantum systems and are efficiently modelled using variational quantum circuits (VQCs), represented by a handful of parametrised quantum operations known as "gates". This is particularly advantageous for grammar-based learning, where explicit tensor representations grow exponentially with sentence complexity. VQCs mitigate this parameter explosion by accurately approximating these large tensors using unitary rotations, each requiring only

Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 113–123
11 May 2026.

a single trainable parameter. Moreover, as quantum computers become larger and more noise-resistant, there is potential to run these models on actual hardware. This natively executes the required linear algebra—since qubits inherently reside in a tensor (Hilbert) space—thereby circumventing the memory bottlenecks of traditional classical computation. In this paper, we present **CCG-VQC**: a meaning representation for natural language using VQCs. Our representations are guided by the rules of Combinatory Categorial Grammar (CCG); we develop a homomorphic map that turns each rule into a series of quantum gates. The meaning of a sentence in this setting is a quantum state. Semantic similarity of two sentences is modelled by *fidelity*, a quantum information theoretic measure for the overlap between two quantum states.

The capabilities of **CCG-VQC** are showcased on a task from vision-language (VLMs). It has been shown that VLMs lack semantic understanding, as a result, struggle to align their embeddings in structural tasks (Koishigarina et al., 2025; Hendricks and Nematzadeh, 2021; Yuksekgonul et al., 2023; Lewis et al., 2024). A variety of datasets are developed to probe for this challenge, working with cases where the correct and the incorrect texts and images are structural renditions of each other. We evaluate our model on two such benchmarks, ARO (Yuksekgonul et al., 2023) and SVO-Swap (Lo et al., 2025), inspired by SVO-Probes (Hendricks and Nematzadeh, 2021).

The performance of our model is compared with a quantum counterpart of a bag-of-words model (QBoW). We also compare our results to CLIP, which is OpenAI's original VLM (Radford et al., 2021) and an open-sourced version of it called OpenCLIP (Ilharco et al., 2021). Our experiments show that firstly, **CCG-VQC** works better than QBoW, and secondly, **CCG-VQC** outperforms both CLIP and OpenCLIP.

To facilitate a fair comparison, we trained a compact transformer model, which we term Micro-CLIP, matching the parameter count of our **CCG-VQC**. Our experiments revealed that this reduced-scale transformer achieves only near-random performance on the ARO benchmarks, whereas **CCG-VQC**, which operates at that parameter scale by design, achieves significantly higher accuracy. This demonstrates that our structurally-informed model is substantially more parameter-efficient than standard attention-based architectures.

Existing Work and Novel Contributions Unified models of statistics and structure exist and are referred to as Compositional Distributional Semantics, sometimes dubbed as DisCoCat (Coecke et al., 2010, 2013; Maillard et al., 2014; Wijnholds et al., 2020; Grefenstette and Sadrzadeh, 2015). DisCoCat learns words very similar to our approach, but it is based on pregroup grammars (Lambek, 1999), which are not widely used in the community. There also exists a translation between DisCoCat and VQCs (Yeung and Kartsaklis, 2021; Lorenz et al., 2021; Wazni and Sadrzadeh, 2023; Kartsaklis et al., 2021; Wazni et al., 2024), but it relies on the theory of categories and has not been tested on large scale structured benchmarks or compared with state-of-the-art technology. A tensor network semantics was defined for CCG and evaluated on VLM tasks (Lo et al., 2025). VQCs have a significantly lower parameter count than tensor networks. Furthermore, tensor networks remain classical objects. Their output is a vector that can be inputted in the `InfoNCE` learning objective of CLIP-like architectures. The output of a VQC is a quantum state, and we design the new `QInfoNCE` objective function for contractive learning.

2. From Syntax to Quantum Semantics

The textual pipeline makes use of grammar as a structural blueprint for designing quantum circuits. Input sentences are processed according to the rules of a CCG. We have chosen to use CCG here as it is a type-driven logic with functional application rules, making it particularly compatible with tensor algebra. We can simply assign a tensor space to each atomic type and treat function applications as contractions. This will become evident in the following explanation from CCG to quantum circuits. Throughout this work we use the Bobcat parser to tag the sentences and generate the parse trees (Clark, 2021). A set of rules, shown in Figure 1, translates the below described CCG trees into quantum circuits, mapping atomic types to qubits, words to variational quantum circuits, and compositions to Bell measurement with post-selection to the Bell state $(|00\rangle + |11\rangle)/\sqrt{2}$. Specifically, each word in the sentence is represented as a parametrised quantum state, referred to as a variational quantum circuit, which is optimised during training to produce the final textual quantum state $|\psi_{\text{txt}}\rangle$.

2.1. Rules of Grammar

CCG consists of a basic and an advanced set of types and inference rules. The basic CCG has the set of atomic types $\{N, NP, S\}$ representing nouns, noun phrases, and sentences, and two function types: forward and backward application. The function type $A\backslash B$ outputs type A given B is to the left and A/B outputs A given B is on the right. The formal definitions of these rules are below.

$$\frac{X/Y \quad Y}{X}\,{<} \qquad \frac{Y \quad X\backslash Y}{X}\,{>}$$

Consider the following example sentence "*Alice likes Bob*". The syntactic structure of this sentence is derivable using both of the inference rules, see below.

$$\dfrac{\dfrac{\text{Alice}}{NP} \quad \dfrac{\dfrac{\text{likes}}{(S\backslash NP)/NP} \quad \dfrac{\text{Bob}}{NP}}{S\backslash NP}\,{<}}{S}\,{>}$$

Basic CCG has other rules such as type raising T and composition B, each with forward and backward variants. The formal definitions of these rules are given below.

$$\frac{X/Y \quad Y/Z}{X/Z}\,{<}B \qquad \frac{Y\backslash Z \quad X\backslash Y}{X\backslash Z}\,{>}B$$

$$\frac{X}{T/(T\backslash X)}\,{<}T \qquad \frac{X}{T\backslash(T/X)}\,{>}T$$

Composition is used for phrase formation to allow for non-standard constituents, such as those found in complex auxiliary structures. As an example, consider the derivation below.

$$\dfrac{\dfrac{\text{Alice}}{NP} \quad \dfrac{\dfrac{\text{might}}{(S\backslash NP)/VP} \quad \dfrac{\dfrac{\text{like}}{VP/NP} \quad \dfrac{\text{Bob}}{NP}}{(S\backslash NP)/NP}\,{>}\mathbf{B}}{S\backslash NP}\,{<}}{S}\,{>}$$

Type raising allows for non-local dependencies such as those seen in relative clause formation. For instance, see the derivation below.

$$\dfrac{\dfrac{\dfrac{\text{The}}{NP/N} \quad \dfrac{\text{food}}{N}}{NP}\,{>} \quad \dfrac{\text{which}}{(NP\backslash NP)/(S/NP)} \quad \dfrac{\dfrac{\dfrac{\text{Mary}}{NP}}{S/(S\backslash NP)}\,{>}\mathbf{T} \quad \dfrac{\text{bought}}{(S\backslash NP)/NP}}{S/NP}\,{>}\mathbf{B}}{NP\backslash NP}\,{>}}{NP}\,{<}$$

In the above, the subscripts $_>$ and $_<$ let us know that an inference rule is forward or backward applied, respectively. Basic CCG has an expressive power equivalent to context-free grammars (Steedman and Baldridge, 2011). CCG can be endowed with a set of additional rules to increase its expressivity. For example, there are rules for type raising T, coordination Φ, and cross composition $B_\times$. Coordination is used for conjunctives such as "and, or, because". Cross-composition is applied to complex situations such as the treatment of phrasal verbs. For further details we refer the reader to a CCG reference, e.g. see (Steedman, 2000).

2.2. Quantum Computing Foundations

While we provide a brief overview of quantum principles here to establish notation, we recognise that the following summary may be challenging for those new to the field. For a complete pedagogical introduction, we refer the reader to the standard text by Nielsen and Chuang (Nielsen and Chuang, 2000).

The fundamental unit of information is the qubit, a state vector $|\psi\rangle$ in a 2-dimensional complex Hilbert space $\mathbb{C}^2$. The states $|0\rangle$ and $|1\rangle$ form an orthonormal basis for this space, known as the computational basis. Unlike classical bits, a qubit can exist in a superposition state defined below:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \qquad \alpha, \beta \in \mathbb{C}$$

The following normalisation constraint holds for a qubit:

$$\langle\psi|\psi\rangle = |\alpha|^2 + |\beta|^2 = 1$$

For a system of n qubits, the corresponding state space is the following tensor product of dimension 2^n:

$$\mathcal{H} = \bigotimes_{i=1}^{n} \mathbb{C}^2.$$

The evolution of these states is governed by operators U that are referred to as *unitary*, which belong to the symmetry group $U \in SU(2^n)$. To extract information from the quantum system, we perform a *measurement*. In the computational basis, measuring a qubit collapses the state to $|0\rangle$ with probability $|\alpha|^2$ or $|1\rangle$ with probability $|\beta|^2$.

The above is the Hilbert Space formalism of quantum computing. In practice, it is advantageous to work with a circuit model, which provides a direct path for running the circuits on quantum computers. In the circuit model of computation, these operations are decomposed into a sequence of quantum gates. Single-qubit gates, such as the Hadamard (H) gate, are used to generate superposition:

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{1}$$

while rotation gates (R_x, R_z) shift the qubit's state by a specific angle θ:

$$R_z(\theta) = \exp\left(-i\frac{\theta}{2}Z\right) = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix} \tag{2}$$

Multi-qubit gates, such as the CNOT, are crucial for creating entanglement between qubits, which is necessary to explore the exponentially large state space, where the state of one qubit cannot be described independently of the others. This entanglement is what allows quantum circuits to model the interactions between words and phrases in a sentence.

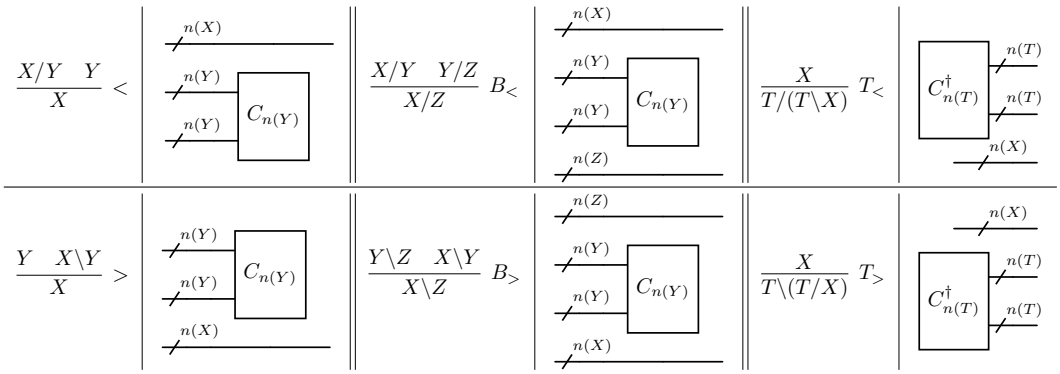

Figure 1: Mapping from the CCG rules to VQCs. The C gates, defined in Figure 2, are either performing Bell measurements mirroring predicate-application, or initialisation Bell states mirroring raising types.

Traditionally, quantum circuits are read from left to right. Each horizontal line represents a qubit (or a bundle of n qubits), and boxes represent gates. A vertical line connecting a filled circle ($\bullet$) to a plus symbol ($\oplus$) denotes a CNOT gate, where the bit of the target qubit is flipped only if the control qubit is in state $|1\rangle$. Triangles at the start of a wire ($|0\rangle$) indicate state preparation, while triangles at the end ($\langle 0|$) represent measurement and post-selection on the 0 outcome.

2.3. Variational Quantum Circuits for CCG Derivations

The conversion from a CCG derivation to a quantum circuit begins with a map n which assigns to each atomic type A a desired number of qubits $n(A)$. The choice of n for the atomic types is a hyperparameter of the model, which can be tuned to balance computational cost and model expressivity. The number of qubits assigned to a composite type X/Y or $X\backslash Y$ is recursively defined as

$$n(X/Y) = n(X\backslash Y) = n(X) + n(Y) \qquad (3)$$

The meaning of a word with type T is represented by an $n(T)$-qubit quantum state $|\psi\rangle$, prepared by a variational quantum circuit $U_{n(T)}(\Theta)$ acting on an initial $|0\rangle^{\otimes n}$ state. These circuits depend on a list of trainable parameters Θ, typically rotation angles. While an arbitrary n-qubit state requires $\mathcal{O}(2^n)$ parameters to specify, VQCs utilise an *ansatz* to explore the high-dimensional Hilbert space using only a polynomial number of parameters.

An *ansatz* is a parametrised circuit template designed to balance expressivity with trainability using a small set of rotation gates and entangling gates. The generalised version for the sentence 'Alice likes Bob' can be seen in Figure 2(a). In our diagrams, wires in parallel represent a tensor product of qubit spaces, while gates spanning multiple wires generate the entanglement necessary to model multilinear mappings. In Figure 2(c) and Figure 2(d), we demonstrate an *ansatz* schema being applied to approximate the space of the learnable state of the word. Application rules correspond to a quantum operation that contract the qubits of two adjacent words or phrases; which is the Bell measurement followed by post-selection (Lorenz et al., 2021), Figure 2(b1) and Figure 2(b2). In this paper, we use a variant of the Sim14 (Sim et al., 2019) ansatz which we call **SAP**, defined by a layer of R_y rotation gates followed by a ring of controlled R_x rotations, then another layer of R_y rotation gates and a ladder of controlled R_x rotations in the opposite direction. A complete example is shown in Figure 3.

3. A Vision-Language Challenge

Vision-language understanding is a key challenge in AI, with applications to image captioning and multimodal retrieval. Models such as OpenAI's CLIP (Radford et al., 2021) have shown that large-scale joint embeddings can effectively connect visual and textual data. However, these models use transformer architectures with dense attention, which often overlook linguistic structure. As a result, there has been an increasing interest in evaluating vision-language models (VLMs) against syntactic and semantic structures such as predicate-argument meaning and word order. Various datasets have been developed for this purpose. In this paper, we consider the Attribution, Relation, and Order (ARO) (Yuksekgonul et al., 2023) and the SVO-Swap datasets. We evaluate our VQCs on these two datasets to determine whether they capture the structural relationships between predicates and nouns and whether they can correctly align text descriptions with images.

VLM architectures consist of two parallel pipelines: one for learning text and another for learning image representations in separate semantic spaces. The two are combined with an objective that aligns them in a shared output space. In our framework, the text pipeline learns quantum mean-

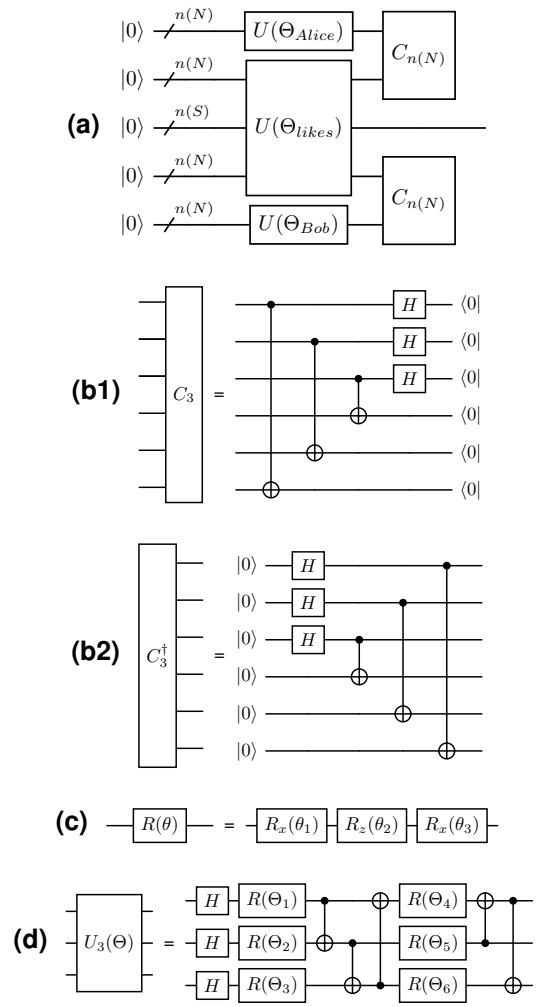

Figure 2: High level VQC for the sentence *Alice likes Bob*. Operator U depends on parameters Θ. C_n is the contractor for n number of qubits.

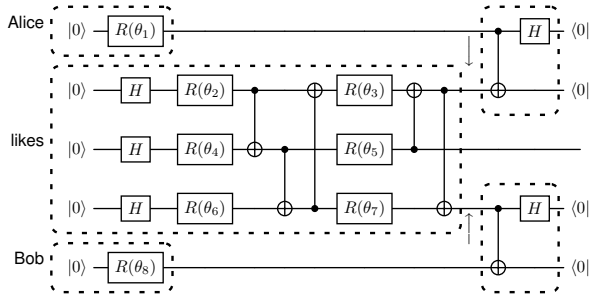

Figure 3: The quantum circuit for the sentence "Alice likes Bob" with $n(N) = n(NP) = n(S) = 1$.

ing representations in a Hilbert space. To align these representations to images, we turn the images into quantum representations using a method known as amplitude encoding. Amplitude encoding turns an m-dimensional image embedding $\overrightarrow{v}_{\text{img}}$ into the amplitudes of a quantum state ψ_{img} that consists of k qubits such that $2^k = m$. The amplitude encoding of $\overrightarrow{v}_{\text{img}} := \Sigma_i a_i \overrightarrow{e}_i$ is

$$|\psi_{\text{img}}\rangle = \frac{1}{\|v_{\text{img}}\|_2} \sum_{i=0}^{m-1} a_i |i\rangle \tag{4}$$

Where $\|v_{\text{img}}\|_2$ is the standard L_2 norm. Images are encoded using CLIP's pre-trained ViT-B/32 image encoder (Radford et al., 2021), which is kept frozen throughout training. It produces a 512-dimensional feature vector, giving $k = 9$ qubits since $2^9 = 512$. For cases where 2^k is not equal to m, the closest exponent is chosen.

The image and text representations are aligned in a unified space by maximising the similarity between matching image-caption pairs and minimising that between mismatched pairs. The objective function used in CLIP is the `InfoNCE` loss (van den Oord et al., 2019). Given a batch of N image-caption pairs $\{I_i, T_i\}_{i=1}^N$, where I_i and T_i are 512-dimensional vectors, t is a temperature parameter, and $s(\cdot, \cdot)$ is a similarity function between an image embedding I (produced by CLIP's encoder) and a caption embedding T (produced by our quantum model), the loss is:

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp(s(I_i, T_i)/t)}{\sum_{j=1}^{N} \exp(s(I_i, T_j)/t)} \tag{5}$$

In the above, s is a similarity measure. For vectors in a vector space, it is the cosine function. In our setting, we need to compute the overlap of two quantum states and develop a quantum version of `InfoNCE`, which we call `QInfoNCE`, where the overlap is defined below

$$s(|\psi_{\text{txt}}\rangle, |\psi_{\text{img}}\rangle) = |\langle \psi_{\text{txt}} \mid \psi_{\text{img}} \rangle|^2 \tag{6}$$

This is the inner product between two quantum states, living in the complex Hilbert space, known as *fidelity*. However, using fidelity directly leads to a sharp loss landscape and narrow neighbourhoods of high gradient (Cerezo et al., 2021). To alleviate this obstacle, we use a slightly modified smooth version of it, based on the Fubini-Study metric (Hai and Ho, 2023; Stokes et al., 2020; Haddou and Bennai, 2025), which can be used as a measure of similarity in the form

$$s(|\psi_{\text{txt}}\rangle, |\psi_{\text{img}}\rangle) = \arcsin(|\langle \psi_{\text{txt}} \mid \psi_{\text{img}} \rangle|) \tag{7}$$

This is essentially the quantum equivalent of the geometric distance rather than the plain Euclidean distance. It considers the curved path along a manifold between two points lying on it, instead of cutting straight through it with a straight line.

Once the specific similarity score function is chosen, we define how the model decides between a correct and incorrect caption given an image, making the corresponding classification using the

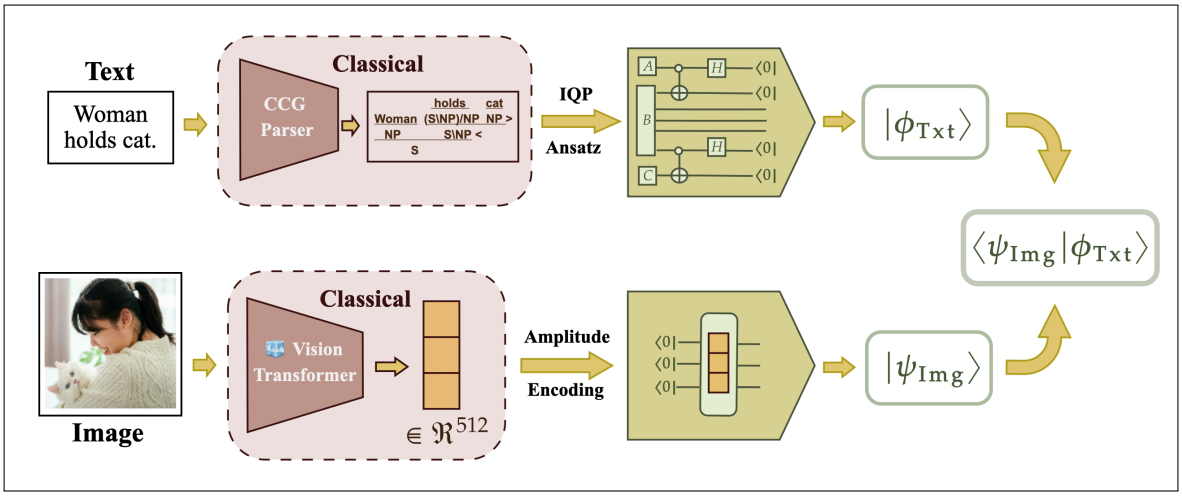

Figure 4: CCG-VQC Vision-Language Pipeline. The architecture maps multimodal inputs into quantum state space. A CCG parser and a quantum circuit are used to encode the input text into the state $|\phi_{\text{Txt}}\rangle$, while a frozen vision transformer and amplitude encoding map the input image into the state $|\psi_{\text{Img}}\rangle$. The similarity between the two modalities is computed via their inner product $\langle\psi_{\text{Img}}|\phi_{\text{Txt}}\rangle$.

'margin'. The margin $\mathcal{M}$ for an image p against a positive text label q^+ and a negative text label q^- is defined by Equation 8, where s is the comparison metric and f and g represent the classical image and quantum text encoders, respectively.

$$\mathcal{M}(p, q^+, q^-) = s(f(p), g(q^+)) - s(f(p), g(q^-)) \quad (8)$$

Therefore, if $\mathcal{M} > 0$ the correct positive label is predicted and if $\mathcal{M} < 0$ the negative incorrect label is selected instead. A higher margin magnitude reflects greater model confidence.

4. Experiments and Results

We evaluated our model on two datasets. The first is the ARO benchmark (Yuksekgonul et al., 2023), which probes noun-attribute and predicate-argument understanding. The noun-attribute subset of the dataset is referred to as **ARO-Attribution** and its predicate-argument subset as **ARO-Relation**. An example of an image entry from ARO-Attribution is the image of a silver fork on a plate with a piece of cake which has dark brown icing. The task is to choose between one of the two pieces of texts: one that describes it as "silver fork and dark brown icing" and another that says "dark brown fork and silver icing". Notably, one piece of text is obtained from the other by swapping the attributes of the nouns. In this case, the nouns are "fork" and "icing", whereas the attributes are "silver" versus "dark brown". Similarly, each entry of ARO-Relation has an image and two pieces of text describing it, where the verb is changed from one text to the other. An example image here is that of a red bus which is to the right of a big building, and the following two pieces of text: "the red bus is to the right of the big building" versus "the big building is to the right of the red bus".

We also evaluate our model on **SVO-Swap** (Lo et al., 2025), a dataset similar to ARO-Relation but inspired by the SVO-Probes benchmark (Hendricks and Nematzadeh, 2021). It works with diverse verb types, whereas the only verb used in ARO is the verb "to be". An example image is that of a woman holding a cat, described by the two captions "A woman holds a cat" versus "A cat holds a woman". SVO-Swap is a small pilot dataset of 95 evaluation pairs, created from SVO-Probes by swapping subjects and objects in its captions.

We trained the model using the `Adam` optimiser (Kingma and Ba, 2015) along with the ReduceLROnPlateau scheduler from the PyTorch python library. We use a batch size of 256 for SVO-Swap and 512 for ARO. We also tested a variety of combinations of qubits and layers for our ansätze. We denote these varieties by pairs (n_q, n_l), where n_q is the number of qubits and n_l is the number of layers. We tested with the following varieties: (1,2), (2,2), (3,2), (4,3), and (5,3). Additionally, we test our CCG-VQC with two different ansatz.

To ensure a fair comparison, we evaluate our approach against two baselines: a VQC made from a bag-of-words model, called *QBoW* and a family of text transformer models sharing the same architecture as the ones used in CLIP but with greatly reduced parameters, termed *MicroCLIP*. In the QBoW each word is translated into an independent ansatz. Words are combined with each other by taking their Frobenius multiplications, their pointwise multiplication. This is a commutative operation so QBoW does not even preserve word order let alone grammatical structure. To ensure

Attribution (Sim: 0.6578, Margin: -0.0198)
C: The tall tree and the green grass.
S: The green tree and the tall grass.

Relation (Sim: 1.0223, Margin: -0.0075)
C: The man is to the left of the person.
S: The person is to the left of the man.

Figure 5: Here 'sim' is the similarity score between the two captions and the 'margin' is the difference between the positive and negative similarities as defined in Equation 8. Here we look at ambiguous cases (high 'sim') from the ARO attribution and relation datasets that the model failed to correctly caption (negative 'margin'). Captions are labelled as Correct (**C**) and Swapped (**S**).

our architectural contributions are evaluated independently of model scale and training volume, we introduce MicroCLIP, a family of models trained from scratch using the same data and protocol as our proposed models. To match the low parameter count of our structured VQCs (10K-100K) and achieve a fair comparison, we drastically shrink the standard CLIP text transformer encoder by reducing the number of layers from 12 to 2, attention heads from 8 to 2, and embedding dimensions from 512 to 8, 16, or 32 (yielding MicroCLIP-8, MicroCLIP-16, and MicroCLIP-32).

Lastly, we also display previous results based on the tensor networks variant of our model before conversion to VQC (Lo et al., 2025).

4.1. Results

Table 1 summarises our performance on the SVO-Swap and ARO datasets. Our fully structured VQC outperforms all others on the ARO-Attribution and ARO-Relation, as well as on the newly developed dataset SVO-Swap, demonstrating the benefit of encoding linguistic structure without training on hard negatives. The bag-of-words version of VQCs performs as expected, reaching only an accuracy of 50%, since it is commutative and ignores the word order. In this model, the quantum circuits of both texts of the entries of each dataset are equivalent up to the parameters.

Despite using 63M parameters, CLIP and its variant OpenCLIP achieve lower accuracy than the **CCG-VQC** across all benchmarks. CLIP achieves 57.89% on SVO-Swap versus 83.16% for **CCG-VQC**; 61.00% versus 71.19% on ARO-Attribution; and 51.53% versus 57.33% on ARO-Relation. OpenCLIP improves over CLIP on SVO-Swap (63.16%), but still falls short of **CCG-VQC** (83.16%), and shows no improvement on ARO.

In Figure 6(a), we show that CCG-VQC outperforms MicroCLIP while using an order of magnitude fewer parameters. As seen in Figure 6(b), the model performed best in SVO-Swap, which shows a positive mean margin and an accuracy of 83.16%. This plot shows that the majority of data had a positive margin, meaning the majority of captions were correctly aligned with images. The outliers below the decision boundary $(\mathcal{M} < 0)$ expose occasional misclassification; this was in cases where the incorrect caption was aligned with an image. A deeper look at the results reveals that in these cases the visual pairing of the subject and the object was ambiguous, i.e. the subject and object looked alike in the image or the captions were very similar, Figure 5. An interesting challenge to note is that as the model achieved 90.1% accuracy

Table 1: Results on SVO-Swap and ARO.

	SVO-Swap	ARO	
		Attribution	Relation
QBoW	50.00	50.00	50.00
CCG-VQC$_{SAP}$	**83.16**	**71.19**	**57.33**
MicroCLIP-8	68.42	50.33	50.11
MicroCLIP-16	69.99	50.27	49.84
MicroCLIP-32	69.68	50.85	51.05
CLIP	57.89	61.00	51.53
OpenCLIP	63.16	59.13	50.71

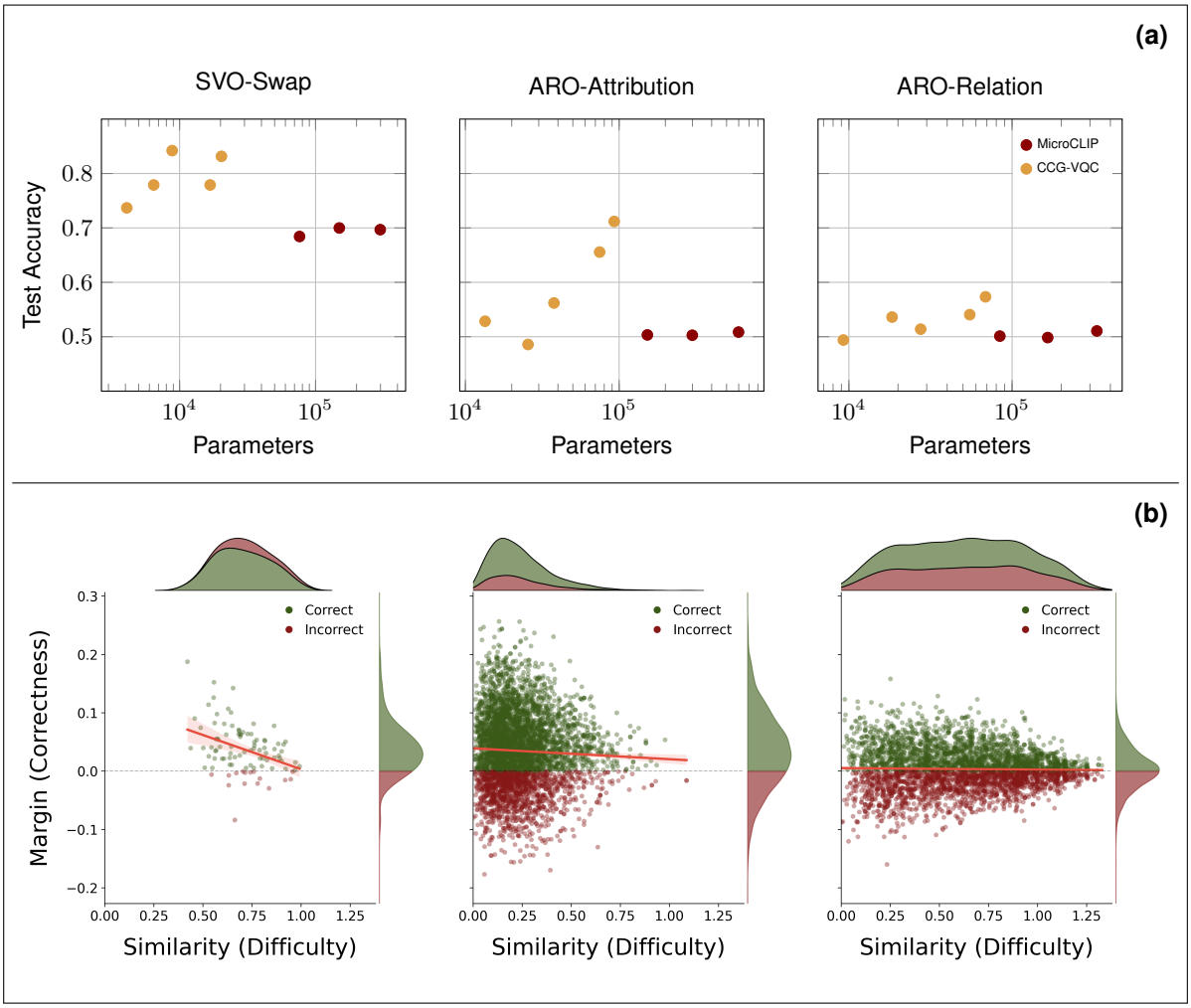

Figure 6: (a) Test accuracy vs. trainable parameters. CCG-VQC achieves competitive or superior accuracy while requiring an order of magnitude fewer parameters than MicroCLIP. (b) Margin of correctness (Equation 8) against caption similarity. Regression lines (red) illustrate performance degradation as lexical similarity (difficulty) increases. Marginal distributions on each axis indicate the density of correct (green) and incorrect (red) classifications relative to the decision boundary (dotted line).

during training on the ARO-Attribution dataset this visually contrasts with the test accuracy of 71.19%. This suggests a generalisation gap when applying learned adjective-noun structures to unseen data. Moreover, for the ARO-Relation dataset both the training accuracy and test accuracy remained near the 50% mark. This suggests a structural collapse when trying to capture spatial relations.

5. Discussion and Analysis

The overall performance of our model across all datasets demonstrates that the use of VQCs and quantum states, in most cases, enabled us to successfully create structurally aware meaning representations capable of capturing the asymmetric behaviour of verbs and other predicates, such as prepositions and adjectives. Here, our **SAP** ansatz did the best and CLIP and OpenCLIP fell short due to their transformer architecture having a lack of explicit structural inductive bias.

This is most evidently seen in the SVO-Swap dataset, where our **CCG-VQC** model achieves an accuracy of 83.16%. SVO-Swap has a primitive linguistic structure; its sentences consist of only three roles: subject, verb, and object. This result demonstrates that when parse complexity is minimised, the model learns it well. This extends to ARO-Attribution, achieving an accuracy of 71.19%, further showing that our model can effectively assign attributes to corresponding objects. On ARO-Relation, which is the hardest of our benchmarks, the model achieves an accuracy of only 57.33%, while low in isolation, many classical models only manage to reach an accuracy of around 50%. This indicates that our model can capture some, albeit minimal, relational dependencies.

This performance shift highlights limitations of

the CCG framework for complex linguistic tasks. Hence after conducting error analysis, in particular investigating the similarity between captions, we noticed the model's lower accuracy seems to arise from the text encoder failing to distinguish between very similar captions. In our failure cases, ARO-Relation text circuits showed high similarity, contrasting with the clear separation observed in ARO-Attribution (see Figure 5). ARO-Relation underperforms compared to SVO-Swap for two main reasons: its verbs lack strong selection preferences, and it contains more complex sentences largely consisting of prepositional phrases. The CCG trees for these phrases are highly nested. This leads to high-rank tensors that disperse semantic information and saturate the representation space, which in turn forces the meaning representations to collapse into a very concentrated region of the Hilbert space. This heavily affects our model's discriminative power. It would be fruitful to explore the relationship between the parse complexity of such models and their resulting performance in terms of accuracy.

There are also structural asymmetries between the text and image modalities. Unlike text, we use a very basic VQC for images, which does not reflect its semantic structure. Further, this encoding is built on the frozen classical vision transformer, which treats the images as a flat grid of patches. As a result, the quantum text encoder is forced to do all the structural heavy lifting, trying to align a highly structured linguistic state to a static and structurally flat visual state, potentially causing the representation to lose structural awareness.

The integration of these modalities introduces further architectural challenges and topological mismatches. Similar to CLIP, our approach relies on late fusion, where the text and image are processed in complete isolation and only interact at the end via similarity. This prevents the text's relational structure from explicitly guiding the visual embeddings or vice versa. Indeed, it has been shown theoretically that no joint embedding space can simultaneously represent concept categorisation, attribute binding, and spatial relationships when cross-modal interaction is reduced to a single scalar similarity score (Kang et al., 2025). Additionally, there is a fundamental topological mismatch between the learnt text embeddings in the complex Hilbert space and the frozen image embeddings originally learnt in Euclidean space. Enforcing alignment on a static Euclidean-informed image manifold crudely projected onto the complex Hilbert space risks stripping the quantum state vectors of their delicate structural representations, further contributing to a potential alignment collapse. Overall, an interesting future direction would be to explore where specifically these bottlenecks lie. If a learnable image component was used with more sophisticated fusion would the text encoder actually suffer from the aforementioned issues or would it perform well in spite of them.

6. Summary and Outlook

In this paper, we introduced **CCG-VQC**, a meaning representation for natural language that uses Variational Quantum Circuits (VQCs) guided by the rules of Combinatory Categorial Grammar (CCG). By mapping syntactic CCG derivation trees into quantum circuits, our model represents words as trainable quantum states and linguistic compositions as entangling operations, encoding sentence structure directly in the circuit topology. Our experiments show that this structural inductive bias leads to consistent gains on compositional benchmarks.

The unstructured quantum bag-of-words baseline (QBoW) collapses to 50% accuracy on every dataset, while **CCG-VQC** reaches 83.16% on SVO-Swap, 71.19% on ARO-Attribution, and 57.33% on ARO-Relation, surpassing both CLIP and OpenCLIP despite using two orders of magnitude fewer parameters. A parameter-matched MicroCLIP baseline confirms that the gains stem from architectural structure rather than scale alone.

Several directions remain open for future work. An interesting area to explore is how different grammar-based topologies handle deep nesting in complex linguistic structures and whether this helps reduce parse dilution. One interesting alternative that could be explored is Universal Dependency (UD) grammars. In multimodality, a two-fold improvement would involve an early-fusion or unified space with a learnable image encoder, allowing structurally aware text embeddings to inform image embeddings and enabling bidirectional alignment that better preserves embedding topology.

Pre-training on larger datasets such as MSCOCO (Lin et al., 2014) and ConceptualCaptions (Sharma et al., 2018) should improve generalisation, particularly for ARO-Relation.

Evaluating **CCG-VQC** on structural NLP benchmarks beyond VLM tasks is another natural future work, as is assessing its robustness to gate noise and running it on near-term quantum hardware.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Patrick Blackburn and Johan Bos. 2005. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Center for the Study of Language and Information.

D. A. Boiko, R. MacKnight, B. Kline, et al. 2023. Autonomous chemical research with large language models. *Nature*, 624:570–578.

M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. 2021. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1).

Stephen Clark. 2021. Something old, something new: Grammar-based ccg parsing with transformer models.

Bob Coecke, Edward Grefenstette, and Mehrnoosh Sadrzadeh. 2013. Lambek vs Lambek: Functorial vector space semantics and string diagrams for Lambek calculus. *Annals of Pure and Applied Logic*, 164(11):1079–1100.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2015. Concrete models and empirical evaluations for the categorical compositional distributional model of meaning. *Computational Linguistics*, 41(1):71–118.

Marwan Ait Haddou and Mohamed Bennai. 2025. Sculpting quantum landscapes: Fubini-study metric conditioning for geometry aware learning in parameterized quantum circuits.

Vu Tuan Hai and Le Bin Ho. 2023. Universal compilation for quantum state tomography. *Scientific Reports*, 13(1):3750.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image–language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

T. Hubert, R. Mehta, L. Sartran, et al. 2025. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 632:290–297.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. 2025. Is clip ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22436–22446.

Dimitri Kartsaklis, Ian Fan, Richie Yeung, Anna Pearson, Robin Lorenz, Alexis Toumi, Giovanni de Felice, Konstantinos Meichanetzidis, Stephen Clark, and Bob Coecke. 2021. lambeq: An Efficient High-Level Python Library for Quantum NLP. *arXiv preprint arXiv:2110.04236*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. 2025. Clip behaves like a bag-of-words model cross-modally but not uni-modally.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks.

J. Lambek. 1999. Type grammar revisited. In *Logical Aspects of Computational Linguistics*, pages 1–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. 2024. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500.

Haobo Li, Zhaowei Wang, Jiachen Wang, Yueya Wang, Alexis Kai Hon Lau, and Huamin Qu. 2025a. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Tianle Li, Jihai Zhang, Yongming Rao, and Yu Cheng. 2025b. Unveiling the compositional ability gap in vision-language reasoning model.

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 (ECCV)*, pages 740–755. Springer.

Kin Ian Lo, Hala Hawashin, Mina Abbaszadeh, Tilen Gaetano Limbäck-Stokin, Hadi Wazni, and Mehrnoosh Sadrzadeh. 2025. DisCoCLIP: A distributional compositional tensor network encoder

for vision-language understanding. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 316–327, Suzhou, China. Association for Computational Linguistics.

Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. 2021. Qnlp in practice: Running compositional models of meaning on a quantum computer.

Jean Maillard, Stephen Clark, and Edward Grefenstette. 2014. A type-driven tensor-based semantics for CCG. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 46–54.

Glyn Morrill. 2010. *Categorial Grammar: Logical Syntax, Semantics, and Processing*. Oxford University Press.

Ruikang Ni, Da Xiao, Qingye Meng, Xiangyu Li, Shihui Zheng, and Hongliang Liang. 2024. Benchmarking and understanding compositional relational reasoning of llms.

Michael A. Nielsen and Isaac L. Chuang. 2000. *Quantum Computation and Quantum Information*. Cambridge University Press.

H. Pan, N. Mudur, W. Taranto, et al. 2025. Quantum many-body physics calculations with large language models. *Communications Physics*, 8:123–130.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565.

Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. 2019. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12).

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Mark Steedman and Jason Baldridge. 2011. *Combinatory Categorial Grammar*, chapter 5. John Wiley & Sons, Ltd.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. 2020. Quantum natural gradient. *Quantum*, 4:269.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Hadi Wazni, Kin Ian Lo, Lachlan McPheat, and Mehrnoosh Sadrzadeh. 2024. Large scale structure-aware pronoun resolution using quantum natural language processing. *Quantum Machine Intelligence*, 6(2):60.

Hadi Wazni and Mehrnoosh Sadrzadeh. 2023. Towards transparency in coreference resolution: A quantum-inspired approach. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 15–27, Singapore. Association for Computational Linguistics.

Gijs Wijnholds, Mehrnoosh Sadrzadeh, and Stephen Clark. 2020. Representation learning for type-driven composition. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 313–324.

Richie Yeung and Dimitri Kartsaklis. 2021. A CCG-based version of the DisCoCat framework. In *Proceedings of the 2021 Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science (SemSpace)*, pages 20–31, Groningen, The Netherlands. Association for Computational Linguistics.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In *International Conference on Learning Representations (ICLR)*.