A Concept-Centric Approach to Multi-Modality Learning

Anonymous authors
Paper under double-blind review

Abstract

Humans possess a remarkable ability to acquire knowledge efficiently and apply it across diverse modalities through a coherent and shared understanding of the world. Inspired by this cognitive capability, we introduce a concept-centric multi-modality learning framework built around a modality-agnostic concept space that captures structured, abstract knowledge, alongside a set of modality-specific projection models that map raw inputs onto this shared space. The concept space is decoupled from any specific modality and serves as a repository of universally applicable knowledge. Once learned, the knowledge embedded in the concept space enables more efficient adaptation to new modalities, as projection models can align with existing conceptual representations rather than learning from scratch. This efficiency is empirically validated in our experiments, where the proposed framework exhibits faster convergence compared to baseline models. In addition, the framework's modular design supports seamless integration of new modalities, since projection models are trained independently yet produce unified outputs within the shared concept space.

We evaluate the framework on two representative downstream tasks. While the focus is not on task-specific optimization, the framework attains competitive results with a smaller training footprint, no task-specific fine-tuning, and inference performed entirely within a shared space of learned concepts that offers interpretability. These findings point toward a promising direction for developing learning systems that operate in a manner more consistent with human cognitive processes.

1 Introduction

Humans are capable of acquiring knowledge at remarkable speed even from a young age, which stands in stark contrast to most learning frameworks that require substantial resources to achieve human-like intelligence on specific tasks. Moreover, human cognition is grounded in a shared and coherent understanding of the world that spans across different modalities. For instance, when learning a new language, we do not build an entirely separate system of knowledge for it. Instead, we intuitively connect new linguistic elements to our existing understanding of the world, or in other words, to our common sense. We believe a concept-centric approach to multi-modality learning could be key to not only bridging the efficiency gap but also bringing us closer to a learning process that mirrors human cognition.

At the center of our framework is a concept space that carries universal knowledge applicable to diverse modalities, resembling the common sense embedded in the human mind. Recent inspiring works on Concept Learning often focus on linking concepts to specific neurons (Liu et al., 2023b) and encoded embedding vectors (Kalibhat et al., 2023; Wang et al., 2023) of a model, or injecting specific concepts as neurons into a model's structure (Sheth & Kahou, 2023; Koh et al., 2020). Compared to these works, our proposed framework takes a systematic approach by organizing modality-agnostic abstract concepts in an interpretable knowledge space and establishing connections to different modalities by projecting modality-specific inputs onto the same space.

While it is common in multi-modality learning to create a shared representation space for multiple modalities (Radford et al., 2021; Li et al., 2022; Ramesh et al., 2022) or even utilize projections to align features from different modalities (Liu et al., 2023a; Hsiung et al., 2022), our shared concept space differentiates itself by possessing abstract knowledge, which facilitates efficient learning and effortless incorporation of new

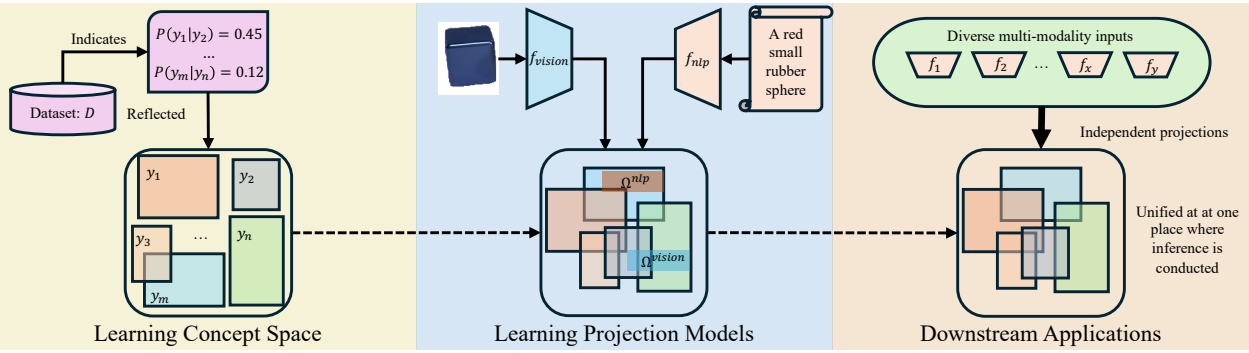

Figure 1: Overall structure of the proposed concept-centric multi-modality learning framework. A modality-agnostic concept space is trained to reflect the relations between the set of concepts $\mathcal{Y}$ as observed in a training dataset $\mathcal{D}$ (left). Modality-specific projection models are trained to create projections Ω for their inputs based on the inputs' associations with concepts (middle). The modular design of the framework offers great flexibility and adaptability to a wide range of downstream tasks (right).

modalities into the framework, as demonstrated in our experiments. We believe the proposed framework is a step closer to matching the capabilities of human learning, where we excel in creating a cohesive comprehension of concepts and seamlessly connecting multiple modalities, such as vision and language, to learned knowledge.

Specifically, as outlined in Fig. 1, the proposed multi-modality learning framework features an abstract concept space and a set of modality-specific projection models. The modality-agnostic concept space, inspired by prior works on structured embedding spaces (Vilnis et al., 2018; Li et al., 2018), optimally reflects real-world relations between concepts via entailment probabilities (Fig. 1, left). Probing this concept space can also be achieved through simple queries of concept pairs of interest, providing interpretability into the learned knowledge.

Complementing the concept space, modality-specific projection models process inputs from different modalities and map them into a shared domain, which we refer to as the knowledge space (Fig. 1, middle). This knowledge space hosts both the abstract knowledge encoded in the concept space and the specific information extracted from individual inputs. By decoupling the projection models from the concept space, the framework enables efficient and modular learning. Each projection model is only required to produce consistent outputs within the knowledge space, allowing flexibility in architecture and optimization for different modalities. Although the projection models operate independently, their outputs are unified in the knowledge space, where they can interact with each other and with the learned concept representations, resulting in a structure that supports probabilistic reasoning and cross-modality interactions.

The proposed design, characterized by a shared concept space with universally applicable knowledge and flexible projection mechanisms, naturally facilitates the reuse of learned knowledge across diverse modalities and task domains. Such a design enhances the generalizability of our framework and enables straightforward adaptation to various downstream tasks, with all inference processes conducted within the knowledge space (Fig. 1, right).

Contribution. Our contributions are three-fold. *First*, we propose a novel approach to multi-modality learning that centers around a concept space embedded with universally applicable knowledge. To our knowledge, this idea of a concept-focused learning scheme has rarely been explored in the field of multi-modality learning (Sec. 3). *Second*, we offer a clear motivation and justification for the proposed framework. Leveraging knowledge learned from the concept space, our framework demonstrates more efficient learning curves compared to traditional methods (Fig. 2). The effectiveness of the concept space is further validated through an ablation study (Sec. 5). *Third*, we evaluate our framework's performance on two downstream tasks. We show that the proposed framework, with a modest pretraining footprint, achieves comparable performance to benchmarks out-of-the-box without fine-tuning, while conducting all inference within a shared knowledge space containing interpretable concept representations (Sec. 4).

2 Related Work

Multi-Modality Learning. Vision and language modalities remain at the forefront of multi-modality learning research, with some works exploring alternative modalities like audio (Akbari et al., 2021; Shi et al., 2022) and biomedical data (Masood et al., 2025). Within the vision-language area, CLIP by Radford et al. (2021) employs two modality-specific encoders to learn a joint representation through image-text matching. Subsequent work by Ramesh et al. (2022) introduces a text-to-image generation framework, using a text encoder and an image decoder to generate high-quality images from textual descriptions. Transformer-based architectures (Vaswani et al., 2017) have been widely explored for cross-modality information exchange and learning (Singh et al., 2022a; Bao et al., 2022; Kim et al., 2021a).

Beyond combining and relating modalities, research has delved into diverse areas such as multi-modality few-shot learning (Alayrac et al., 2022; Li et al., 2021) and visual-textual pattern mining (He & Peng, 2020). Some studies propose generalized learning frameworks applicable across various modalities (Jaegle et al., 2021; Baevski et al., 2022a;b). While these frameworks showcase strong capabilities in tasks like text-to-image generation and visual-language few-shot learning, our work addresses a distinct and important issue: creating a universally applicable concept space with abstract knowledge reflecting real-world observations. Baevski et al. (2022b) present a versatile representation learning framework, yet it isolates modalities, impeding cross-modality interactions. In contrast, our proposed method directly combines modalities by projecting modality-specific inputs onto a unified concept space, eliminating the information barrier between them.

Concept Learning. Early approaches to Concept Learning utilized Boolean logic for defining concepts based on relationships with other concepts (Angluin, 1988) and their associated attributes (Mitchell, 1997). Lake et al. (2015) propose a Bayesian Program Learning framework that represents concepts as probabilistic programs. Nowadays, a prevalent method involves placing concepts within a structured embedding space. Concept learning frameworks such as those proposed by Mao et al. (2019) and Li et al. (2020b) construct embedding spaces that align concept representations with corresponding visual feature vectors. Lee et al. (2024) propose a framework that learns concept embeddings via distillation from pre-trained vision-language models. Methods from Vilnis et al. (2018) and Mei et al. (2022) emphasize entailment relationships between concepts in learned embedding spaces, while the work from Sinha et al. (2024) captures hierarchical information.

In a departure from structured concept embedding spaces, the Concept Bottleneck Model (CBM) (Koh et al., 2020) has become a popular framework that represents concepts as intermediate neural network outputs. CBM first predicts a set of pre-defined concepts aligned with human annotations and then produces a classification decision based on those concept predictions. Liu et al. (2023b) propose a method for identifying a small subset of model parameters responsible for generating specific concepts in a diffusion model. Kong et al. (2024) propose a theoretical view of concept learning as an identification problem of a discrete latent hierarchical model.

While we acknowledge that some motivating works adopt a similar strategy involving a concept embedding space, our approach stands out for several reasons. The primary distinction lies in the organization of our concept space, which reflects real-world knowledge by providing meaningful numerical entailment probabilities that mirror relationships among actual concepts. Furthermore, no barrier in our concept space prevents concepts belonging to different groups, such as *red* in `color` and *cube* in `shape`, from interacting with each other. More importantly, instead of being fitted to a specific modality, our concept space is designed to be abstract and modality-agnostic, thus allowing interactions between inputs from different modalities.

3 Method

Our proposed multi-modality learning framework consists of a modality-agnostic concept embedding space that models underlying relationships between concepts via entailment probabilities and a set of modality-specific projection models that extract representation from single-modality inputs and project them onto the domain where the concept space resides, i.e., the knowledge space.

Learning abstract knowledge in the concept space ensures generality, which makes its domain a good landing place for extracted representations from different modalities. Decoupled from the concept space and each other, modality-specific projection models can be tailored for adaptation to their unique inputs, while modality-specific knowledge remains connected after the projection.

We describe the design of the concept space in Sec. 3.1 and projection models in Sec. 3.2. Further implementation details can be found in Sec. 4.1.

3.1 Learning Concept Space

Davis et al. (1993) describe a knowledge representation as a surrogate that both carries the *thing* existing in the real world and serves as a medium for pragmatically efficient computation. Building upon their definition of a knowledge representation, we adopt an embedding space proposed by Li et al. (2018) to organize learned representations of abstract concepts. Like mental entities of specific knowledge in our brains, where we can relate concepts to each other, abstract entities in this concept space should be capable of interacting with each other, allowing reasoning inferences. In the proposed framework, we focus on entailment relations between concepts depicted by entailment probabilities to allow interactions between concepts. Contrary to latent spaces or learned ML model parameters, probing into the learned knowledge of this concept space can be easily achieved by querying the entailment probabilities of concept pairs of interest. Furthermore, our experiments demonstrate the efficiency of learning and referencing this concept space, facilitated by its compact parameter size, which qualifies it as *a medium for pragmatically efficient computation*.

Defining Concept Space. We first define a knowledge space $\mathcal{K} \subset \mathbb{R}^d$ as a d-dimensional embedding space. Let $\mathcal{Y}$ be a set for modality-agnostic concepts. Each concept $y \in \mathcal{Y}$ is represented in $\mathcal{K}$ by a box embedding (the *surrogate*), defined by a pair of vectors $\Omega_y = (\omega_{\min,y}, \omega_{\max,y})$, where $\omega_{\min,y}, \omega_{\max,y} \in \mathcal{K}$ correspond to the minimum and maximum boundaries of the box in $\mathcal{K}$. We use $\mathcal{C} = \{\Omega_y \mid y \ \in \mathcal{Y}\} \subset \mathcal{K}$ to denote a set of box embeddings for every concepts in $\mathcal{Y}$ and we call $\mathcal{C}$ the concept space whose parameters are optimized to reflect real-world knowledge.

A smoothing function $m^i_{\text{soft}}(\omega) = \frac{\text{softplus}(\omega^i)}{\text{softplus}(G^i_{\max} - G^i_{\min})}$ is introduced on each dimension i of $\mathcal{K}$ so a joint probability between two disjoint concepts can still be obtained. $G^i_{\max}, G^i_{\min}$ terms are the global maximum and minimum values at the i dimension among all Ω_ys in $\mathcal{C}$. More details of m^i_{soft} can be found in Appendix A.1. The probability of a single concept y is calculated as $P(y) = P(\Omega_y) = \prod_{i=1}^{d} m^i_{\text{soft}}(\omega_{\max,y} - \omega_{\min,y})$. The joint probability between two concepts y_1 and y_2 is calculated as

$$P(y_1 \cap y_2) = P(\Omega_{y_1} \cap \Omega_{y_2}) = \prod_{i=1}^{d} m^i_{\text{soft}}(\min(\omega_{\max,y_1}, \omega_{\max,y_2}) - \max(\omega_{\min,y_1}, \omega_{\min,y_2}))$$

Embedding Knowledge. Let $\mathcal{X}_*$ denote a sample space of an unspecified modality marked by *, where each sample can be associated by a subset of modality-agnostic concepts in $\mathcal{Y}$. A training dataset is given as $\mathcal{D}_* = \{(x^*_i, \boldsymbol{y}_i)\}_{i=1}^N$, where $x^*_i \in \mathcal{X}_*$ and $\boldsymbol{y}_i = \{y_j \mid y_j \in \mathcal{Y} \text{ and } y_j \text{ describes } x^*_i\}$. This set of concepts that describe x^*_i can include both attribute concepts, like *fluffy* and *blue*, as well as category concepts, like *dog* and *sky*.

Modality-agnostic abstract knowledge can be extracted from $\mathcal{D}_*$ by examining entailment probabilities between concepts indicated by $\{\boldsymbol{y}_i\}_{i=1}^N$. Specifically, the ground-truth probability of a single concept and the entailment probability of a concept pair (y_1, y_2) are calculated by $P(y) = \frac{\text{count}(y)}{\sum_{y' \in \mathcal{Y}} \text{count}(y')}$ and $P(y_1 \mid y_2) = \frac{\text{count}((y_1 \cap y_2))}{\text{count}(y_2)}$ as they appear in $\mathcal{D}_*$.

To drive the concept space to reflect real-world relationships between concepts via entailment probabilities, the objective for pretraining $\mathcal{C}$ is naturally defined as minimizing the KL divergence between predicted probabilities obtained from $\mathcal{C}$ and true probabilities observed in $\mathcal{D}_*$. In addition to true concepts in $\boldsymbol{y}_i$ for each data point, a set of negative concepts is sampled and added to $\boldsymbol{y}_i$. A well-organized concept space should also reflect these negative concepts' true entailment probabilities with the original concepts. Details

of this negative sampling procedure vary by specific datasets and further information is provided in Sec. 4. For each sample, we calculate an entailment probability $Q(y_1 \mid y_2)$ indicated by the concept space for every possible combination of concept pairs (y_1, y_2) in $\boldsymbol{y}_i$ and compare them to the true entailment probabilities $P(y_1 \mid y_2)$. We refer readers to Appendix A for further details regarding the concept space.

3.2 Learning Projection Models

Defining Projection Models. Decoupled from the abstract concept space, each modality-specific projection model can be viewed as a mapping function $f_* : \mathcal{X}_* \to \mathcal{K}$ that generates a box representation in $\mathcal{K}$ for each input from its modality-specific sample space $\mathcal{X}_*$ of an unspecified modality denoted by *. This projection onto $\mathcal{K}$ allows interactions between specific objects from $\mathcal{X}_*$ and abstract concepts in $\mathcal{C}$. Specifically, given a modality-specific input $x_i^* \in \mathcal{X}_*$, its representation in $\mathcal{K}$ can be obtained by $f_*(x_i^*; \theta) = \Omega_i^*$ where $\Omega_i^* \subset \mathcal{K}$ follows the same definition of $\Omega_y \subset \mathcal{C}$. With this representation made available, the probability that an object is associated with a concept y can be naturally described by an entailment probability of $P(y \mid x_i^*) = P(\Omega_y \mid \Omega_i^*)$.

Adapting to the Concept Space. Given the training set $\mathcal{D}_*$ corresponding to a modality marked by $*$, the projection produced for an input x_i^* should entail not only a single concept y but also **all** other concepts associated with x_i^*. In other words, the projection Ω_i^* for x_i^* should lie at the **intersection** of the set of concepts describing x_i^*. Thus, the optimal projection for x_i^* should maximize the entailment probability $P(\bigcap_{y_j \in \boldsymbol{y}_i} y_j \mid x_i^*)$.

To drive projection models to produce this most optimal projection, we use a combination of a binary cross-entropy loss on attribute concepts $\mathcal{Y}^{\text{attr}} \subset \mathcal{Y}$:

$$\begin{aligned} \ell_{\text{attr}}(\boldsymbol{y}, \Omega_*) = \frac{1}{|\mathcal{Y}^{\text{attr}}|} \sum_{y \in \mathcal{Y}^{\text{attr}}} & \mathbb{I}(y \in \boldsymbol{y})[-w \cdot \log P(\Omega_y \mid \Omega_*)] \\ & + \mathbb{I}(y \notin \boldsymbol{y})[\log(1 - P(\Omega_y \mid \Omega_*)] \end{aligned} \tag{1}$$

(where w is a weight assigned to positive attribute concepts)

and a multi-class cross-entropy loss with SoftMax on category concepts $\mathcal{Y}^{\text{cat}} \subset \mathcal{Y}$:

$$\ell_{\text{cat}}(\boldsymbol{y}, \Omega_*) = -\log \frac{\exp P(\Omega_{y^{\text{cat}}} \mid \Omega_*)}{\sum_{y' \in \mathcal{Y}^{\text{cat}}} \exp P(\Omega_{y'} \mid \Omega_*)} \tag{2}$$

(where $y^{\text{cat}} \in \boldsymbol{y}$)

Now, given a specific modality denoted by A and its training dataset $\mathcal{D}_A$. The training objective for f_A is formally described as minimizing:

$$\mathcal{L}_A(\theta_A; \mathcal{D}_A) = \frac{1}{|\mathcal{D}_A|} \sum_{(x, \boldsymbol{y}) \in \mathcal{D}_A} \ell_{\text{attr}}(\boldsymbol{y}, f_A(x; \theta_A)) + \ell_{\text{cat}}(\boldsymbol{y}, f_A(x; \theta_A)) \tag{3}$$

While the training objective and projection outputs remain consistent across different modalities, projection models can be customized to accommodate unique modality-specific inputs, such as images or sequences of texts, bringing flexibility and versatility to the proposed framework.

3.3 Cross Modality Joint Training

To allow probabilistic analysis for cross-modality tasks, we introduce a joint training stage that encourages different projection models to produce projections that overlap with each other's for the same object. This

joint training stage is lightweight since modality-specific projection models have already been trained and adapted to a unified concept space. It requires very modest resources, with convergence occurring within a few hundred training steps, as indicated in Fig. 5 of Appendix. Subsequently, this design with demonstrated efficiency allows the effortless incorporation of new projection models into our proposed framework, mirroring humans' ability to learn and link knowledge across modalities in a fast and efficient manner. Specifically, consider a system with two modalities, A and B, as an example. The training dataset would be denoted as $\mathcal{D}_{A\cup B} = \{(x_i^A, x_i^B, \boldsymbol{y}_i)\}_{i=1}^N$, and the training objective for this joint training stage is defined as:

$$\mathcal{L}_{\text{joint}}(\theta_A, \theta_B; \mathcal{D}_{A\cup B}) = \frac{1}{2|\mathcal{D}_{A\cup B}|} \sum_{(x_A, x_B, \boldsymbol{y})\in\mathcal{D}_{A\cup B}} P(f_A(x_A;\theta_A) \mid f_B(x_B;\theta_B)) + P(f_B(x_B;\theta_B) \mid f_A(x_A;\theta_A)) \tag{4}$$

The overall training objective becomes a combination of modality-specific projection losses and this joint training loss. Optionally, the optimization can also include parameters from $\mathcal{C}$, so that the abstract knowledge learned in the concept space is adjusted based on modality-specific information. Then the objective becomes $L'_{\text{joint}} = L_{\text{joint}} + \beta L_{\mathcal{C}}$ where $L_{\mathcal{C}}$ denotes the KL divergence loss of the concept space.

3.4 Adapting to Downstream Tasks

With an abstract concept space and decoupled projection models, our proposed learning framework naturally accommodates various downstream tasks involving single or multiple modalities. Regardless of the specific downstream tasks, their inference process consists of two stages: creating projections and relating them to learned knowledge. This approach more closely resembles human learning than traditional black-box models. In our daily interactions with objects, we process external stimuli like vision by creating abstract mental entities for objects we see. We then comprehend these mental entities using our understanding of the world, or, in other words, our concept space (Gärdenfors, 2014). In Section 4, we use an Image-Text Matching task involving multi-modality and a Visual Question Answering task with a single-modality-focused approach to illustrate the functionality of the proposed framework.

4 Implementation and Experiments

We base our evaluation on three datasets: CLEVR (Johnson et al., 2017a), COCO (Lin et al., 2014), and GQA (Hudson & Manning, 2019) where their concepts are formed from original and supplemental annotations. Both attribute and categorical concepts are present in COCO and GQA whereas CLEVR only contains attribute concepts. More details on the datasets and preprocessing steps can be found in Appendix B. Our experiments follow the same train and validation splits as the original datasets. The proposed framework is pretrained on the train sets and tested on the validation sets.

4.1 Pretraining

Concept Space. To ensure that each concept box always has a valid set of lower boundaries smaller than its upper boundaries, we use two vectors, $(\omega_{\min,y}, \omega_{\Delta,y}) = \Omega_y$, instead of $(\omega_{\min}, \omega_{\max})$ to represent a box in our actual experiments, where $\omega_\Delta \in \mathcal{K}_{\geq 0}$ is restricted to non-negative values. A box's upper boundaries can be obtained by $\omega_{\max} = \omega_{\min} + \omega_\Delta$. We set the dimension of $\mathcal{K}$ to 50, based on empirical experiments. Initial parameters for $\mathcal{C}$ are sampled from two uniform distributions. As for the negative sampling method, in CLEVR, the only negative concept pairs come from combinations of concepts residing in the same-attribute families, such as (*red*, *blue*) in the `color` family. For COCO and GQA, negative samples are randomly selected from all concepts. The concept space is trained for just two epochs for each dataset with a batch size of 256 using an AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 10^{-3}. The training of this concept space can be completed quickly as there are only thousands of parameters for a moderately-sized concept space.

Projection Models. In adapting our framework to the datasets featuring vision and natural language modalities, we incorporate a vision projection model f_{vision} based on a Vision Transformer encoder (Dosovitskiy et al., 2020) and a natural language projection model f_{NL} based on a BERT encoder (Devlin et al., 2018). Both models utilize their encoders' outputs on `[CLS]` tokens to generate projection boxes in $\mathcal{K}$. The outputs e with a dimension of 768 are divided into two equal chunks, $h_{\min}$ and h_{Δ}, each with a dimension of 384. These chunks are then input into two fully connected layers to produce $\omega_{\min}$ and ω_{Δ} for their respective projection boxes. To ensure ω_{Δ} is always a non-negative vector, an additional ReLU layer is applied. The complete projection process for inputs from the vision modality is outlined in Algorithm 1. 0

Algorithm 1 Illustration of a ViT-based projection model f_{vision} which projects vision modality inputs to the knowledge space $\mathcal{K}$

input modality-specific input x_{vision}

Ensure: $\omega_{\Delta}^{\text{vision}} \in \mathcal{K}_{\geq 0}, \Omega^{\text{vision}} \subset \mathcal{K}$

$e_{\text{vision}} \leftarrow \text{ViT}(x_{\text{vision}})$

$h_{\min}^{\text{vision}}, h_{\Delta}^{\text{vision}} \leftarrow \text{split}(e_{\text{vision}})$

$\omega_{\min}^{\text{vision}} \leftarrow \text{Linear}_{\min}^{\text{vision}}(h_{\min})$

$\omega_{\Delta}^{\text{vision}} \leftarrow \text{ReLU}(\text{Linear}_{\Delta}^{\text{vision}}(h_{\Delta}))$

output $\Omega^{\text{vision}} = (\omega_{\min}^{\text{vision}}, \omega_{\Delta}^{\text{vision}})$

For each object i in the CLEVR dataset, its attribute prediction for a specific attribute family $\boldsymbol{z}$ (e.g., `color`) is generated by $\bar{y}_i^z = \text{argmax}_{y \in \boldsymbol{z}} P(\Omega_y | \Omega_i)$. For each object i in COCO and GQA, a threshold is applied to $P(\Omega_y \mid \Omega_i), y \in \mathcal{Y}^{\text{attr}}$ to obtain attribute predictions, and category prediction is generated by $\bar{y}_i^{\text{cat}} = \text{argmax}_{y \in \mathcal{Y}^{\text{cat}}} P(\Omega_y \mid \Omega_i)$.

We establish a baseline by replacing the concept space with a traditional Multilayer Perceptron (MLP) at the classification head of f_{vision}. Additionally, we implement the vision-modality projection model using a ResNet model (He et al., 2015) as the backbone to showcase the flexibility of the proposed framework. Results summarized in Table 1 show that our proposed framework achieves comparable performance to traditional models while leveraging a novel concept space with interpretable learned knowledge.

Backbone	Method	CLEVR	COCO		GQA	
		Accuracy	Accuracy	F1 Score	Accuracy	F1 Score
ResNet	Baseline	$0.997_{\pm 1.0e^{-4}}$	$0.897_{\pm 1.7e^{-3}}$	$0.625_{\pm 2.4e^{-3}}$	$0.733_{\pm 4.5e^{-3}}$	$0.401_{\pm 2.8e^{-3}}$
	f_{vision}	$0.990_{\pm 2.7e^{-3}}$	$0.900_{\pm 7.8e^{-4}}$	$0.621_{\pm 2.1e^{-3}}$	$0.724_{\pm 1.7e^{-3}}$	$0.429_{\pm 3.1e^{-3}}$
ViT	Baseline	$0.999_{\pm 3.9e^{-5}}$	$0.956_{\pm 2.0e^{-3}}$	$0.663_{\pm 1.7e^{-3}}$	$0.841_{\pm 2.3e^{-3}}$	$0.567_{\pm 1.7e^{-3}}$
	f_{vision}	$0.999_{\pm 4.0e^{-5}}$	$0.955_{\pm 1.4e^{-3}}$	$0.658_{\pm 2.4e^{-3}}$	$0.839_{\pm 3.0e^{-3}}$	$0.574_{\pm 1.1e^{-3}}$

Table 1: A comparison with baseline models on classification performance of vision-modality inputs. Category concepts are evaluated with accuracy (%) and attribute concepts with f1 score. 2-sigma errors over five trails of experiments are reported

Apart from featuring a concept-centric learning scheme, the proposed framework can also learn modality-specific knowledge faster by referencing learned knowledge from the modality-agnostic concept space as indicated in Fig. 2. This more natural learning process of our framework bridges the efficiency gap between traditional machine learning methods, which often demand extensive data, and human learning, which excels at adeptly and efficiently extracting modality-specific representations and associating them with mental entities of abstract knowledge. To fully evaluate the impact of this transparent, modality-agnostic concept space on the learning of modality-specific projection models, we conduct an ablation study on it in Sec. 5.

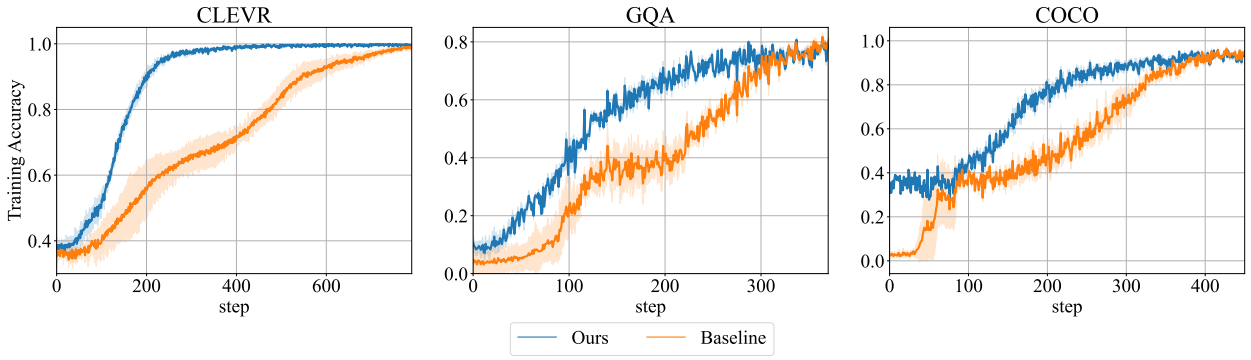

Figure 2: Learning curves of proposed projection models and baseline models. Shaded area in plots represents 2-sigma errors over five trails of experiments. During the learning process, the proposed vision-modality projection model converges faster compared to the baseline thanks to the universal concept space that already has abstract knowledge embedded in it. This faster learning process of our framework bridges the efficiency gap between traditional machine learning methods, which require a huge amount of data, and human learning that excels at extracting modality-specific representations and linking them to mental entities of abstract knowledge.

Projection models for the natural-language modality achieve highly accurate performance ($\geq 99\%$) thanks to the clearly structured description sentences. Further implementation and training details of projection models can be found in Appendix C.

Now, we focus on our proposed framework's adaptation to two downstream tasks: Image-Text Matching involving cross-modality references and Visual Question Answering with a single-modality-focused approach.

4.2 Image-Text Matching

Image-text matching is a binary classification task on whether a natural language sentence describes an image. Our framework can naturally adopt a common approach involving creating representations for sentences and images in a shared latent space. In contrast to those works, however, our latent space is a knowledge-embedded concept space that supports efficient probing. Specifically, given an image-text pair $(x_m^{\text{vision}}, x_n^{\text{NL}})$, their representations in the learned concept space $\mathcal{C}$ are generated by $f_{\text{vision}}(x_m^{\text{vision}}) = \Omega_m^{\text{vision}}$ and $f_{\text{NL}}(x_n^{\text{NL}}) = \Omega_n^{\text{NL}}$. The probability that $(x_m^{\text{vision}}, x_n^{\text{NL}})$ is a positive pair can be determined by the cross entailment probability as follows:

$$P(\text{matched} \mid (x_m^{\text{vision}}, x_n^{\text{NL}})) = \frac{1}{2}\left[P(\Omega_m^{\text{vision}} \mid \Omega_n^{\text{NL}}) + P(\Omega_n^{\text{NL}} \mid \Omega_m^{\text{vision}})\right]$$

This inference process is demonstrated in Fig. 7 in Appendix.

In our experiments, we employ two methods to create negative image-text pairs: swapping whole description sentences and swapping attributes. Specifically, for the first method, we replace 50% of images' description sentences using random sampling. For example, an original description sentence of a CLEVR object might be changed from "There is a large, metal, red cube" to "There is a *rubber, small, yellow sphere*." On the other hand, swapping attributes involves changing only a subset of attributes that describe an object, creating a more challenging image-text matching task. For instance, the same description sentence would be changed to "There is a *small*, metal, red cube."

To compare our framework's performance, we implement other benchmark multi-modality models with applications in the Image-Text Matching task. The results are summarized in Table 2. In contrast to those models with traditional black-box architectures, our framework displays a more efficient learning process and adopts a more transparent inference process without sacrificing its performance. Details of this experiment can be found in Appendix D.

Method	Fine-tuned?	CLEVR		COCO		GQA	
		sent.	attr.	sent.	attr.	sent.	attr.
BLIP (Li et al.)	✓	0.999	0.999	0.992	0.536	0.979	0.576
CLIP (Radford et al.)	✗	0.997	0.997	0.974	0.587	0.945	0.532
FLAVA (Singh et al.)	✓	0.998	0.998	0.992	0.505	0.980	0.536
ViLT (Kim et al.)	✓	0.994	0.994	0.985	0.515	0.965	0.555
Ours	✗	0.995	0.995	0.970	0.552	0.929	0.536

Table 2: A comparison with state-of-the-art multi-modality models on the Image-Text Matching Task. We test these models and our framework using two variants of the matching task: swapping whole sentences (sents.) and swapping attributes (attr.). Classification accuracy (%) is reported.

4.3 Visual Question Answering

Visual Question Answering (VQA) evaluates an AI system's ability to reason about images by answering questions related to those images in a natural language format. For this task, we focus on the CLEVR dataset, whose questions are designed to include attribute identification, counting, comparison, spatial relations, and logical operations. Recently, several works (Johnson et al., 2017b; Yi et al., 2018; Mao et al., 2019; Li et al., 2020a; Mei et al., 2022) have focused on a neural-symbolic reasoning approach, using chains of symbolic programs to predict answers to these questions. Our framework's adaptation to VQA involves using a similar set of symbolic programs, but these programs operate on the knowledge space $\mathcal{K}$ containing interpretable concepts in $\mathcal{C}$ instead of the high-dimensional latent spaces used by previous works.

Problem Formulation. Given an image-question pair $\{X_i^{\text{vision}}, q_i\}$ where X_i^{vision} is an original CLEVR image as shown in Fig. 6 and q_i is a natural language question such as *"Are there more cubes than yellow things?"*, an AI system needs to generate an answer o_i in the natural language format such as *"Yes"*.

Symbolic Programs. We design our symbolic programs as deterministic functions operating on $\mathcal{K}$. Precisely, we follow the same program definitions as proposed by Johnson et al. (2017a).

Program Generator. An LSTM model π is used to process questions into sequences of programs: $\hat{z}_i = \pi(q_i)$. We follow the same pretraining procedure used in (Johnson et al., 2017b) to train this program generator. However, as there is no fine-tuning stage in our adaptation, the parameters in π are frozen once pretraining is finished.

Object Detection and Projection. Similar to our pretraining process, we use $f_{\text{detection}}$ to obtain a set of single-object images $\boldsymbol{x}_i^{\text{vision}}$ from X_i^{vision} which are then fed into f_{vision} so their projections can be obtained. Additionally, each single object's coordinates predicted by $f_{\text{detection}}$ are attached to its projection box so questions involving spatial relations can be inferred.

Inference Process. A correctly predicted program sequence $\hat{z}_i$ starts with a `Scene` function that returns all objects in an image and ends with a program that outputs the answer o_i. Intermediate programs takes output from previous programs as inputs, which is a reoccurring process until the last function. Our concept space $\mathcal{C}$ is mainly involved in attribute identification which follows the same rule as used when evaluating projection models' performance in Sec. 4.1. The complete inference process is also demonstrated in Fig. 8 in Appendix.

Results. We perform no fine-tuning on the concept space $\mathcal{C}$ and vision-modality projection model f_{vision} for the VQA task. A comparison to benchmark models summarized in Table 3 shows our framework achieves performance levels on par with those fine-tuned benchmark models.

5 Ablation Study

We discover that using a pretrained concept space with learned abstract knowledge helps modality-specific projection models converge faster compared to the ones without the access. Specifically, we cut our

Method	Accuracy	Fine-tuned?
SA+MLP (Johnson et al.)	73.2	✓
Dependency Tree (Cao et al.)	89.3	✓
Human (Johnson et al.)	92.6	N/A
RN (Santoro et al.)	95.5	✓
IEP (Johnson et al.)	96.9	✓
MDETR (Kamath et al.)	99.7	✓
NS-VQA (Yi et al.)	99.8	✓
Ours	96.5	✗

Table 3: A comparison between our framework's performance and state-of-the-art models.

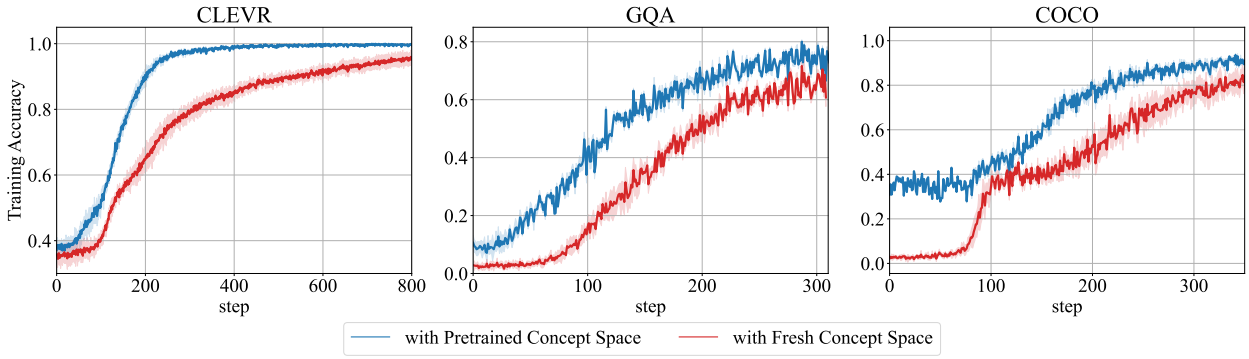

Figure 3: Ablation study on the pretrained concept space. We cut our projection models' access to the pretrained concept space and the learning of this concept space is combined into training processes of the projection models. Shaded area in plots represents 2-sigma error over five trails of experiments. Their classification accuracy is used to compare the ablated version and the original framework.

framework's access to the pretrained concept space $\mathcal{C}$. Instead, the framework is only provided with a freshly initialized concept space $\mathcal{C}'$ and the loss function during pretraining of the vision-modality projection model is changed to $\mathcal{L}'_{\text{vision}} = \mathcal{L}_{\text{vision}} + \mathcal{L}_{\mathcal{C}}$. Fig. 3 shows that the original framework's projection models can converge faster than the ablated version. Based on this evidence, we conclude that the abstract knowledge shared by the pretrained concept space streamlines the learning process of modality-specific projection models.

6 Discussion

A Cognition-Inspired Learning Paradigm. Most current multi-modality learning frameworks, and even the broader landscape of machine learning systems, rely on a learning paradigm that differs substantially from those observed in human cognition. When exposed to new knowledge, we instinctively form a concept and associate it with the external stimuli tied to that information. This newly formed concept is then integrated into our existing body of knowledge and stored as persistent memory in the mind. In contrast, most machine learning frameworks encode knowledge into large sets of model parameters that are difficult to interpret without specific model input. As a result, the activation of learned knowledge in such systems is often transient and dependent on specific inputs. This fundamental difference presents a significant challenge in designing systems that can explicitly form, retain, and reason over interpretable concepts in a manner analogous to human cognition.

The inclusion of a structured concept space and the use of concept-grounded inference may initially appear restrictive, particularly when compared with conventional models that rely on dense, task-specific representations optimized end to end for performance. However, we view this design as a deliberate and principled choice. By introducing a concept space that reflects structured, abstract knowledge similar to how humans form and retain concepts, the framework gains several benefits that are otherwise difficult to

achieve. These include more efficient learning, natural generalization across modalities, and interpretability through explicit probing. The concept space acts as an inductive bias consistent with human cognition, enabling machine learning systems to operate in a more principled and cognitively grounded manner. We believe this work highlights a compelling direction for rethinking learning systems to more closely mirror human intelligence.

Addressing Bias. Hidden biases learned from datasets often hinder the trustworthiness of ML systems (Amodei et al., 2016; Lederer, 2023; Kaur et al., 2022; Knott et al., 2023). For example, NLP models often tend to associate the word "monarch" more with the word "male" than "female," reflected, for instance, in higher similarity scores between embeddings of "monarch" and "male." Our proposed framework facilitates effective probing into the model's learned knowledge and offers the capacity to rectify such biases.

Concept 1	Concept 2	Concept Space	Ground Truth
Orange	Bus	0.043	0.043
Old	Building	0.032	0.048
Smiling	Person	0.074	0.073
White	Snow	0.910	0.974
Parked	Car	0.228	0.244
Cloudy	Sky	0.173	0.192

Table 4: Sample Entailment Relation Queries of Concepts in Learned GQA Concept Space

Table 4 shows probing of a learned concept space fitted to the GQA dataset in action. Our framework enables easy querying of targeted concept pairs, which would be computationally expensive, if not infeasible, in traditional latent spaces. Further demonstrations of probing into the learned concept space can be found in Appendix A.3.

Revisiting the earlier example of the concept pair "monarch" and gender, the bias can be addressed directly in our framework by adjusting the ground-truth entailment probabilities. Specifically, ensuring equal entailment probabilities between "monarch–male" and "monarch–female" mitigates representational bias, a correction that can be easily applied through user-guided specification.

Scalability of the Concept Space. In our experiments, the concept space is constructed to reflect ground-truth entailment probabilities observed in training data. This approach can scale to larger and more diverse sets of concepts. Prior work (Vilnis et al., 2018; Li et al., 2018; Lai & Hockenmaier, 2017) has shown that similar embedding structures can learn entailment relations for large ontologies such as WordNet (WordNet). Although scaling introduces challenges in generating ground-truth probabilities, textual corpora offer a promising resource for extracting such relations, as demonstrated by He & Peng (2020). To assess scalability, we fitted a concept space to the full set of WordNet noun entries, totaling 10,765 concepts. The resulting space achieved a KL divergence of 0.1308 with respect to the ground truth, compared to 0.1172 for the GQA concept space.

Call for Concept-Focused Datasets. A major bottleneck in concept-centric learning is the lack of high-quality datasets with accurate concept annotations. In our experience, even after preprocessing, concept and attribute labels in datasets such as COCO and GQA contain significant noise. This limits not only the performance of our framework but also that of other systems. We believe that future datasets with richer, more reliable concept annotations would greatly support the development of interpretable and trustworthy AI systems.

7 Conclusion

This work is motivated by the observation that humans are capable of forming a coherent, structured understanding of the world and applying this knowledge across diverse tasks and modalities. Inspired by this cognitive capability, we proposed a concept-centric multi-modality framework centered around a modality-agnostic concept space that captures universally applicable knowledge.

Our primary technical contribution lies in the design of this framework, which integrates a shared concept space with a flexible set of modality-specific projection models. This design allows knowledge to be reused across modalities and task domains, enabling more interpretable, generalizable, and modular learning. Unlike traditional end-to-end learning systems that encode knowledge implicitly within dense parameter spaces, our framework embeds knowledge explicitly into a structured concept embedding space, enabling interpretability through efficient probing of concept entailment probabilities.

Experimentally, we showed that the proposed framework supports more efficient learning. Specifically, as demonstrated in the vision modality, our projection model converges significantly faster than a baseline model based on a traditional architecture. This gain in efficiency is attributed to the fact that the concept space already encodes structured, abstract knowledge that the projection model can adapt to. We further validated this effect through an ablation study, in which the concept space did not contain prior knowledge and was learned jointly with the projection model. The results confirm that the presence of structured abstract knowledge in the pre-trained concept space facilitates faster convergence, compared to learning the concept space from scratch alongside the projection model. Additionally, we evaluated the framework on two downstream tasks, Image-Text Matching and Visual Question Answering, and demonstrated that our method achieves performance comparable to state-of-the-art methods even without task-specific fine-tuning. While our goal is not to surpass existing benchmarks in raw performance, these results support the viability of a cognitively inspired learning paradigm. Rather than optimizing solely for accuracy, our framework emphasizes learning efficiency, interpretability, and structural alignment with human cognition. These qualities are increasingly important as machine learning systems are deployed in more complex and dynamic environments.

The broader implication of our work is a call to reimagine how machine learning systems acquire and represent knowledge. By introducing a concept space as an inductive bias, this framework opens a promising research direction toward building systems that "think" in a way that more closely resembles human reasoning. Such systems may offer greater transparency, flexibility, and the ability to generalize knowledge across tasks and modalities.

Looking forward, several improvements can further enhance this approach. First, scaling the concept space to support larger vocabularies and richer relational structures beyond entailment, such as compositional or causal relations, would expand its expressive power. Second, applying the framework to new task domains such as concept-grounded Text-to-Image generation presents a natural extension. At a broader level, a key open challenge lies in the need for high-quality concept annotations, which are often unavailable in practice. This highlights the importance of research into methods for unsupervised concept discovery from raw data, as well as more effective strategies for organizing and structuring the discovered concepts. Addressing these challenges will be critical for building more autonomous and cognitively aligned learning systems.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016.

Dana Angluin. Queries and concept learning. *Machine learning*, 2:319–342, 1988.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. *ArXiv*, abs/2212.07525, 2022a.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *International Conference on Machine Learning*, 2022b.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, and Liang Lin. Visual question reasoning on general dependency tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7249–7257, 2018.

Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation? *AI magazine*, 14 (1):17–17, 1993.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Peter Gärdenfors. *The Geometry of Meaning: Semantics Based on Conceptual Spaces The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Xiangteng He and Yuxin Peng. Fine-grained visual-textual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:520–531, 2020.

Eric Hsiung, Hiloni Mehta, Junchi Chu, Xinyu Liu, Roma Patel, Stefanie Tellex, and George Konidaris. Generalizing to new domains by mapping natural language to lifted ltl. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 3624–3630, 2022. doi: 10.1109/ICRA46639.2022.9812169.

Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506, 2019. URL `http://arxiv.org/abs/1902.09506`.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General Perception with Iterative Attention. In *International Conference on Machine Learning*, 2021.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017a.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision*, pp. 2989–2998, 2017b.

Neha Kalibhat, Shweta Bhardwaj, Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. *CoRR*, abs/2104.12763, 2021. URL `https://arxiv.org/abs/2104.12763`.

Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy Artificial Intelligence: a Review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021a.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021b. URL https://api.semanticscholar.org/CorpusID:231839613.

Alistair Knott, Dino Pedreschi, Raja Chatila, Tapabrata Chakraborti, Susan Leavy, Ricardo Baeza-Yates, David Eyers, Andrew Trotman, Paul D. Teal, Przemyslaw Biecek, Stuart Russell, and Yoshua Bengio. Generative ai models should include detection mechanisms as a condition for public release. *Ethics and Inf. Technol.*, 25(4), October 2023. ISSN 1388-1957. doi: 10.1007/s10676-023-09728-4. URL https://doi.org/10.1007/s10676-023-09728-4.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models, 2020.

Lingjing Kong, Guangyi Chen, Biwei Huang, Eric P. Xing, Yuejie Chi, and Kun Zhang. Learning discrete concepts in latent hierarchical models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=bO5bUxvH6m.

Alice Lai and Julia Hockenmaier. Learning to predict denotational probabilities for modeling entailment. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 721–730, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1068.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-lLevel Concept Learning through Probabilistic Program Induction. *Science*, 350(6266):1332–1338, 2015.

Edith M. Lederer. Exec Tells First UN Council Meeting that Big Tech Can't be Trusted to Guarantee AI Safety, Jul 2023. URL https://apnews.com/article/artificial-intelligence-un-big-tech-first-5a184197c4281365866b5963d56f84ea.

Sharon Lee, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. Language-informed visual concept learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=juuyW8B8ig.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. URL https://arxiv.org/abs/2201.12086.

Qing Li, Siyuan Huang, Yining Hong, and Song-Chun Zhu. A competence-aware curriculum for visual concepts learning via question answering. In *European Conference on Computer Vision*, pp. 141–157. Springer, 2020a.

Qing Li, Siyuan Huang, Yining Hong, and Song-Chun Zhu. A Competence-Aware Curriculum For Visual Concepts Learning via Question Answering. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, pp. 141–157, Berlin, Heidelberg, 2020b. Springer-Verlag. ISBN 978-3-030-58535-8. doi: 10.1007/978-3-030-58536-5_9. URL https://doi.org/10.1007/978-3-030-58536-5_9.

Wei-Hong Li, Xialei Liu, and Hakan Bilen. Improving task adaptation for cross-domain few-shot learning. *CoRR*, abs/2107.00358, 2021. URL https://arxiv.org/abs/2107.00358.

Xiang Lorraine Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the Geometry of Probabilistic Box Embeddings. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:108301524.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept Neurons in Diffusion Models for Customized Generation. *ArXiv*, abs/2303.05125, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes Words and Sentences from Natural Supervision. *ArXiv*, abs/1904.12584, 2019.

Muhammad Arslan Masood, Markus Heinonen, and Samuel Kaski. Multi-modal representation learning for molecules. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025. URL https://openreview.net/forum?id=WT7BpLvL6D.

Lingjie Mei, Jiayuan Mao, Ziqi Wang, Chuang Gan, and Joshua B. Tenenbaum. FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic descriptions, and Conceptual Relations. *ArXiv*, abs/2203.16639, 2022.

Tom M. Mitchell. *Machine Learning, International Edition*. McGraw-Hill Series in Computer Science. McGraw-Hill, 1997. ISBN 978-0-07-042807-2. URL https://www.worldcat.org/oclc/61321007.

Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, 2016. URL https://api.semanticscholar.org/CorpusID:14849501.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017.

Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jvYXln6Gzn.

Bowen Shi, Abdelrahman Mohamed, and Wei-Ning Hsu. Learning Lip-Based Audio-Visual Speaker Embeddings with AV-HuBERT, 2022.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language And Vision Alignment Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022a.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022b.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91220–91259. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a5d2da376bab7624b3caeb9f78fcaa2f-Paper-Conference.pdf.

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

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 263–272, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1025. URL https://aclanthology.org/P18-1025.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=SGlrCuwdsB.

WordNet. Wordnet, a lexical database for english. URL https://wordnet.princeton.edu/.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

A Concept Space Details

A.1 Preliminary

A smoothing function for the concept space is defined as:

$$m^i_{\text{soft}}(\omega) = \frac{\text{softplus}(\omega^i)}{\text{softplus}(G^i_{max} - G^i_{min})} \tag{5}$$

where the denominator is a normalization term with G_{max}, G_{min} being the global maximum and minimum values at i dimension. In short, this smoothing function is introduced so a valid joint probability can be calculated even if two concepts/boxes are disjoint and we refer readers to Li et al. (2018) for its complete proof.

A.2 Concept Space Training Objective

We define a KL-divergence measure between a predicted conditional probability distribution $q(y_1|y_2)$ and a target $p(y_1|y_2)$ as:

$$D_{\mathbf{KL}}(P(y_1|y_2)||Q(y_1|y_2)) = \mathbb{E}_{(y_1,y_2)\sim P}\left[\log \frac{P(y_1|y_2)}{Q(y_1|y_2)}\right] \tag{6}$$

Let $\binom{\boldsymbol{y}}{2}$ denote a set of all concept pairs created from 2-combination from $\boldsymbol{y}$ The objective for training the concept space is formally described as the following:

$$\mathcal{L}_{\text{concept}}(\mathcal{C};\mathcal{D}_*) = \frac{1}{|\mathcal{D}_*|} \sum_{(x,\boldsymbol{y})\in\mathcal{D}_*} \frac{1}{2\cdot\left|\binom{\boldsymbol{y}}{2}\right|} \sum_{(y_1,y_2)\in\binom{\boldsymbol{y}}{2}} D_{\mathbf{KL}}(P(y_1|y_2)||Q(y_1|y_2)) + D_{\mathbf{KL}}(1-P(y_1|y_2)||1-Q(y_1|y_2)) \tag{7}$$

A.3 Probing into Concept Space

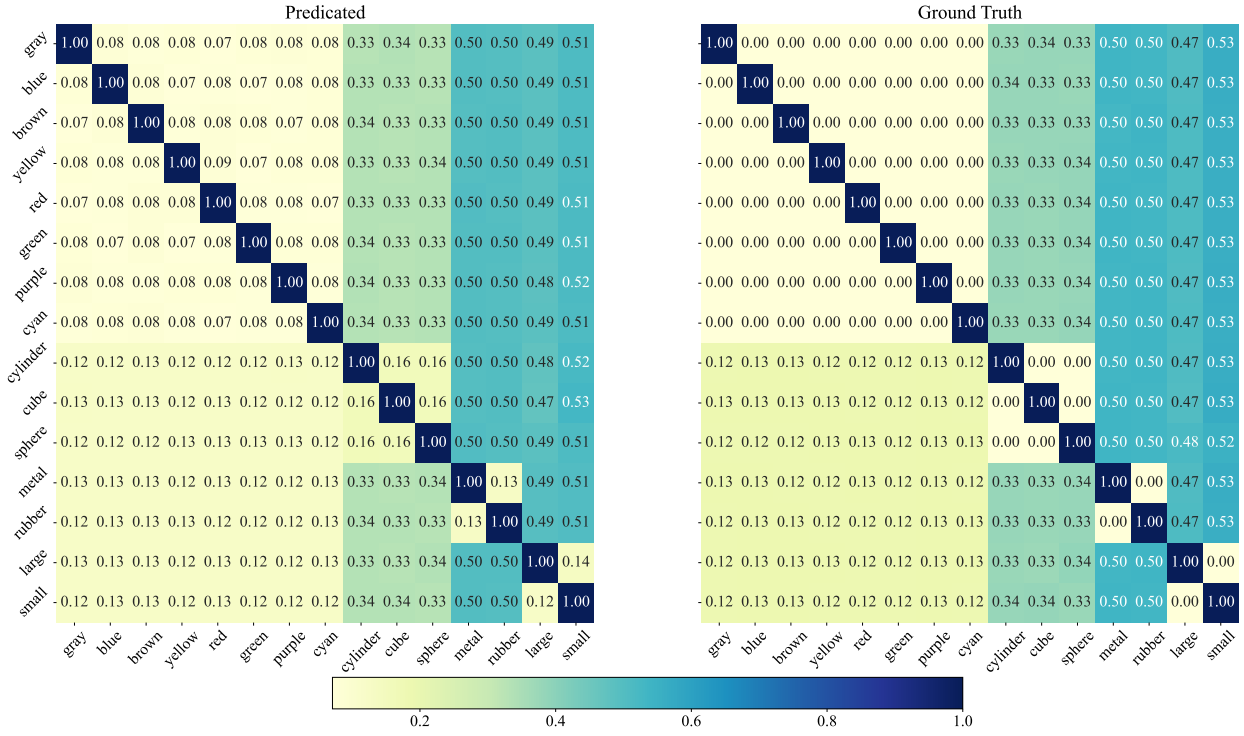

Figure 4: A comparison between the learned concept space's understanding of the CLEVR world and the ground truth relations illustrated via entailment probabilities of concept pairs. Such comparison allows simple probing into the knowledge learned by this abstract concept space. A SoftMax function is applied on entailment probabilities of same-attribute concepts conditioned on a single concept y so $\sum_{y'\in\text{attr}_i} P(y'|y) = 1$ is satisfied.

Figure 4 shows an example of probing into learned knowledge of the concept space exposed to CLEVR. Benefited from such efficient probing mechanism, this concept space offers more interpretability compared to traditional latent spaces or model parameters of previous learning frameworks.

B Evaluation Datasets and Preprocessing

We base our evaluations on three datasets:

CLEVR dataset comprises synthesized images paired with intricate questions testing a system's visual reasoning capabilities. We choose CLEVR for evaluation because it provides a highly controlled mini-world, where concepts are easily drawn from visual objects, and relationships between concepts are clearly defined. Each CLEVR image displays a scene with a random number of objects, each described by `color`, `shape`, `material`, and `size`, which produces 15 unique values, such as `blue`, `cube`, forming attribute concepts related to specific objects.

COCO dataset exposes our framework to a knowledge world resembling the real world better than computer-generated images from CLEVR. We use attribute annotations proposed by Patterson & Hays to establish

attribute concepts such as `soft`, `cooked`, and `parked` (2016). The original COCO classes are used as category concepts. We focus our evaluation on the top 35 frequent attributes and their associated categories to gain meaningful insights, resulting in 64 concepts.

GQA dataset is similar to COCO, providing a controlled sandbox mimicking the real world. We use the original attribute and category labels in GQA as concepts and filter out rare attributes and classes, resulting in the same amount of concepts as in COCO. Example attribute and category concepts include `happy`, `old`, `gray`, and `boy`.

Since each image in these datasets contains multiple objects, a preprocessing step is essential to isolate single objects. This isolation allows focused learning on targeted objects, reducing ambiguity. This process mirrors human learning, where attention naturally centers on a novel object while ignoring the surrounding environment Gärdenfors (2014).

Both COCO and GQA datasets already include object segmentation data. For the CLEVR dataset, we employ a MASK R-CNN model (He et al., 2017), denoted as $f_{\text{detection}}$, trained on a small amount of annotated data as an object detection model to generate segmentation. Visual object inputs are created by cropping original images to include only the objects of interest, as illustrated in Fig. 6.

In addition to object isolation, we generate a descriptive sentence for each object, introducing natural language as a new modality in the dataset. Each sentence of an object has the structure "*There is a*" followed by a sequence of values indicated by its attribute concepts in random orders to ensure diversity. Category concept values are added last to the sequence, except for CLEVR, where values from the `shape` attribute family are placed last for natural-sounding sentences.

C Projection Models Details

C.1 Architecture

ViT-based vision-modality projection models use a vision transformer (ViT-Base) pretrained on ImageNet-21k Dosovitskiy et al. (2020) as the backbone. The baseline MLP model is comprised of three fully-connected layers used as ViT's classification head, with each middle layer containing 128 neurons.

ResNet-based vision-modality projection models use a ResNet model (ResNet-50) pretrained on ImageNet-21k He et al. (2015) as the backbone. Because of ResNet's large feature vectors, the linear layer used to project feature vectors onto the concept space is expanded to a three-layer MLP, featuring two intermediate layers comprising 512 and 256 neurons, respectively. The baseline MLP model is comprised of three fully-connected layers installed after ResNet's layers, with each middle layer containing 128 neurons.

BERT-based nlp-modality projection models use a pretrained BERT encoder (BERT-base) Devlin et al. (2018) as the backbone.

C.2 Training Details

Vision modality projection models are trained for 10 epochs with a batch size of 256 with an exception of CLEVR whose models are only trained for 1 epoch. An AdamW optimizer with a learning rate of 10^{-4} is used. Learning rate schedulers are used to achieve warm-up for first epoch and then a process of 10^{-1} linear decrease over the remaining epochs.

Natural-language modality projection models are trained for 1 epoch using the same setup and hyper-parameters as used by the vision ones.

Thresholds for attribute identification are selected based on performances from training splits. Thresholds producing the best f1 score on training sets are used in tests.

D Image-Text Matching Experiment Details

D.1 Our Framework

We follow the cross-modality joint training method and train our vision and natural language projection models for only 1 epoch with a batch size of 256 and a learning rate of 10^{-4}.

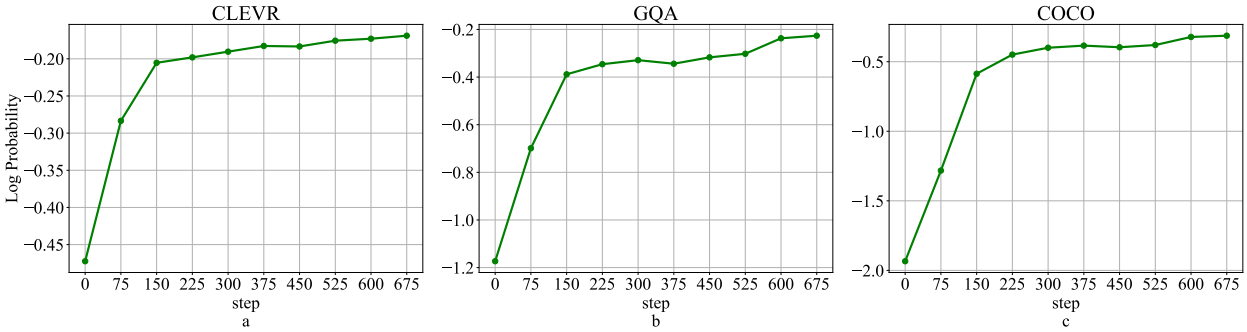

Figure 5: Cross-modality entailment probability of $P_{\text{cross}}(x_i^{vision}, x_i^{NL}) = 0.5 \cdot P(\Omega_i^{vision}|\Omega_i^{NL}) + 0.5 \cdot P(\Omega_i^{NL}|\Omega_i^{vision})$ over joint training steps. It can be observed that projection models of vision modality f_{vision} and natural language modality f_{NL} can quickly learn to produce overlapping projections for the same object in the concept space. Such quick convergence allows easy incorporation of new modalities/modalities into the proposed learning system. This joint training takes significantly less time and uses fewer GPU resources than the following BLIP and CLIP models.

Figure 5 illustrates the fast convergence of the proposed projection models on learning to produce overlapping representations of the same objects in the transparent concept space. This joint training also takes significantly less time and uses fewer GPU resources than the following BLIP and CLIP models.

D.2 BLIP

We follow the training method as stated in Li et al. (2022) and fine-tune the pretrained BLIP model directly on the Image-Text Matching task (swapping-sentence split) using both the image-text contrastive loss and a task-specific image-text matching loss produced by the image-text matching classification head in BLIP. We use a greater batch size of 512 as the calculation of image-text contrastive loss requires a large number of samples.

D.3 CLIP

We follow the training method as stated in Radford et al. (2021) and adapt the pretrained CLIP model to the general three datasets using the symmetric loss that favors larger similarity scores between positive image-text pairs and smaller scores for negative ones. We use a batch size of 512 as in BLIP during pretraining. Similar to our framework, CLIP model is not directly trained on the Image-Text Matching task.

D.4 ViLT

Similar to BLIP, we follow the training method as stated in Kim et al. (2021b) and fine-tune the pretrained ViLT model directly on Image-Text Matching task (swapping-sentence split) using a binary cross-entropy loss on the matching classification head.

D.5 FLAVA

We use the same procedures as used in ViLT to fine-tune a pretrained FLAVA model on the data domains appeared.

E Computation Resources

We run our experiments on a virtual machine (VM) hosted by Microsoft's Azure. This VM has four NVIDIA A100 PCIe GPUs with 320 GB of total memory.

F Additional Figures

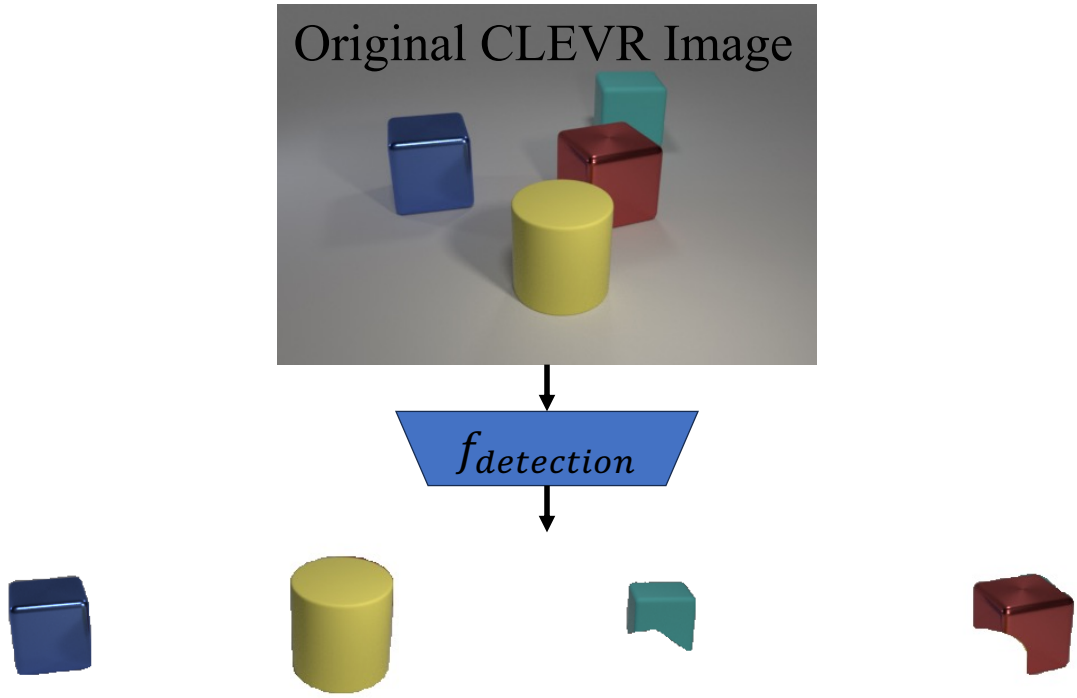

Figure 6: The segmentation masks generated by $f_{\text{detection}}$ are applied to the original CLEVR images to isolate each object from its surroundings environment. This preprocessing step enables our proposed framework to replicate the way we, as humans, naturally focus our attention on novel objects during the learning process.

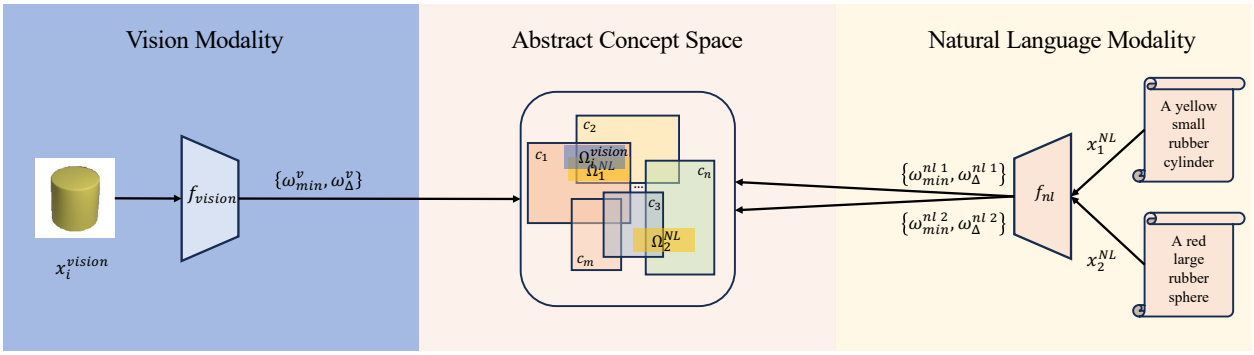

Figure 7: Application of the proposed framework on the Image-text matching task. An image x_i^{vision} of a yellow, small rubber cylinder and two description sentences $x_1^{\text{NL}}, x_2^{\text{NL}}$ are processed by their modality-specific models f_{vision} and f_{NL} which project modality-specific inputs onto a learned abstract concept space $\mathcal{C}$. We use the cross-entailment probability between projections of an image and a sentence to determine if they form a positive pair. While creating representations of images and sentences in a shared latent space is a common approach for the image-text matching task, our shared representation space is a knowledge-embedded concept space offering interpretability, which is in drastic contrast to the commonly used latent space with black-box structure.

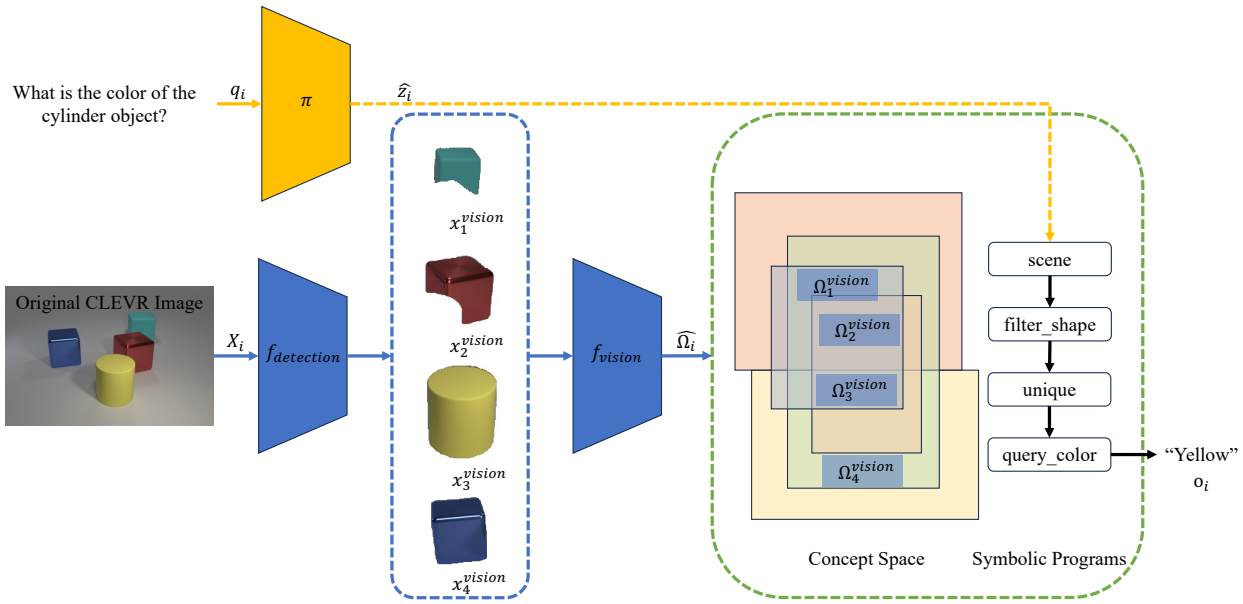

Figure 8: Application of the proposed framework to Visual Question Answering task. We reuse the object detection model $f_{detection}$ from the pretraining stage, which extracts a set of single objects $\boldsymbol{x}_i$ from an original CLEVR image X_i. The vision-modality projection model f_{vision} then projects $\boldsymbol{x}_i$ onto the $\mathcal{K}$. A program generator π is used to predict a sequence of symbolic programs $\hat{z}_i$ based on an input question q_i in natural language format. Programs in $\hat{z}_i$ operate on the concept space and produce an answer o_i to q_i.