Universiteit Leiden

Master Media Technology

Cross-Modal Translation with β-VAE

Name: Vinnayakk Bangarwaa
Student ID: s3773728

Date: 25/03/25

1st supervisor: Dr. G.J. Wijnholds
2nd supervisor: Dr. R. Saunders

Master's Thesis in Media Technology

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This thesis examines how structured representations learned through β-Variational Autoencoders (β-VAEs) can improve the understanding of symbols and help systems generalize better when combining different types of inputs, such as video and audio. It specifically looks at cross-modal translation, which means mapping features from videos to their corresponding audio descriptions, by using a β-VAE along with a modified Sequence-to-Sequence (Seq2Seq) model. An important challenge is finding the best configuration for the β-factor, which controls the balance between accurately reconstructing data and separating features in the latent space. The study tested several β scheduling methods and found that a gradual ramp-up approach works best to balance reconstruction and disentanglement in our scenario.

The methodology starts with the β-VAE learning meaningful representations from video and audio inputs. These disentangled representations capture important meaning while removing unnecessary details like noise. The study then feeds these meaningful representations into the Seq2Seq model to improve the accuracy of cross-modal translation. The results show that the β-VAE combined with the Seq2Seq model performs significantly better than the baseline model that uses raw features. This improvement is shown through better alignment metrics, such as lower cosine distances, stable test losses and the ability to generalize to new object–action pairs. The research suggests that using structured representations improves the understanding of symbols and allows the system to generalize beyond what it learned in training. These findings could be helpful in fields like robotics and human-computer interaction, where meaningful learning from multiple types of information is vital. Future research work will look at how to apply this method to more diverse and complex real-world scenarios and explore automating the tuning of the β-factor for producing even better performance.

Contents

1 Introduction

As a more digital, connected world unfolds, machines' ability to understand and interpret human languages, images, and sound at the same time becomes important. Virtual assistants and autonomous robots all need artificial intelligence (AI) systems to understand symbol meaning, whether in sound, pictures, or movements. This is significant because meaning isn't inherent to symbols but something learned from the interactions within an environment. We assign symbols their meanings. This phenomenon is known as symbol grounding [Harnad, 1990]. Traditional AI models have performed well on tasks like image recognition, speech synthesis, and natural language processing (NLP). Yet most of these models work separate from each other, good at a single task but unable to mix knowledge across different modalities [Bisk et al., 2020]. The main issue is a lack of compositional generalization—the ability to recombine symbols in a meaningful way to explore or generate new instances [Lake and Baroni, 2018]. This limits the ability of AI to display human-like reasoning and adaptability. This study explores a new approach to improve symbol grounding in AI models by using a β-Variational Autoencoder (β-VAEs) and a Sequence-to-Sequence (Seq2Seq) framework. The question is whether or not it will enable compositional generalization and learn to make accurate predictions for unseen combinations. It is a consideration in applications related to robotics, multimodal artificial intelligence, and cognitive science because generalizing outside training data is an essential feature of an intelligent system.

1.0.1 Definition of key concepts

Symbol Grounding - Symbol grounding refers to the process by which an artificial system creates a link between abstract symbols like speech, image, text and audio to their meanings in the physical world, supported by learned mappings to the sensory inputs. A system would be said to have achieved symbol grounding when it is capable of translating between multiple modalities, as in mapping video features to their corresponding audio descriptions.
Compositional Generalization - The ability to restructure already learned knowledge, including objects, actions, and their relations, so that an artificial system can understand or predict correct instances that were not present during training is called compositional generalization. This is observed when a system is able to properly use what it learned to represent new object-action combinations.

1.1 Societal Context

The rapid advancement in AI has brought variety of changes to society. AI supported systems are everywhere in our daily life now, from voice assistants like Siri and Alexa to recommendation mechanisms that personalize material on platforms like Netflix and YouTube. The reason behind this technological shift has been deep learning, which got popularized through architectures like Convolutional Neural Networks (CNNs) for computer vision tasks and Transformer models for handling natural language processing [Vaswani et al., 2023]. For all these developments, AI is still failing to learn basic cognitive abilities that are inherent to humans. Symbol grounding is one such ability which allows attaching abstract symbols to actual objects in the world. Imagine children being taught new words. They don't just learn to associate, but also extend that generalization to new situations. For example, having learned the meaning of "jump," a child will be able to understand "jumping on a trampoline" or "jumping on the bed" without specific training or examples. AI models, however, generally cannot display such

an understanding which restricts their use in dynamic settings [Marcus, 2018]. This restriction is experienced more commonly when dealing with robots, in which the AI system needs to explore in the real world settings. The robot that has learned to classify different categories of things and can carry out different types of activities should be able to generalize over unseen situations—i.e., classifying a new type of cup or performing an alternate version of a learned action. Lack of compositional generalization is a well-known bottleneck to bringing AI into the real world setting [Goyal et al., 2020].

1.2 The Problem

One of the major problems in AI research is developing models that can ground symbols, connect words, images, and actions in a way that leads to meaningful interaction. Most of today's models depend on huge amounts of labeled data and hence are data hungry and computationally intensive. Most importantly, they can't generalize beyond training, i.e., aren't able to understand unseen combinations of known objects and actions [Li et al., 2024]. The problem is even more challenging with multimodal learning, in which AI must be able to map across more than one source of information. For example, in human-robot interaction, a robot would be required to recognize an audio command, translate the corresponding video, and generate an appropriate response. Symbol grounding is required by these systems because otherwise they cannot link sensory inputs and abstract knowledge in an environment. [Bisk et al., 2020]. Recent research suggests that Variational Autoencoders (VAEs) can provide a solution by learning structured latent representations. β-VAEs specifically introduce a regularization term that encourages disentanglement, which allows models to separate meaningful features in the latent space [Higgins et al., 2016]. This disentangled representation can enable improvements in compositional generalization by learning different concepts (e.g., "jump" and "ball") represented separately, so that it becomes easy to reconstruct them in new combinations. A number of studies have attempted to tackle the problem of symbol grounding and compositional generalization. [Ponte and Rauchas, 2022] introduced a sequence-to-sequence (Seq2Seq) model for mapping video features to audio descriptions. The results from the study demonstrate that the model achieved symbol grounding but not compositional generalization. [Li et al., 2024] also introduced a softened symbol grounding model that combines neural perception with symbolic reasoning, but their model relies on pre-defined symbolic modules and is therefore less dynamic. [Hamilton et al., 2024] introduced DenseAV, a self-supervised model that acquires cross-modal relations between video and audio features. As mentioned in the study, the model improved compositional learning but performed poorly when tested with rare object-action pairs. [Hemadri et al., 2024] also developed an intelligent communication system by using a CNN-VAE architecture, which was capable of providing low error rates but was not evaluated in dynamic environments. In the context of Variational Autoencoders, [Burgess et al., 2018] proved that β-VAEs promote disentangled representation, leading to better understanding of learned features. [Asperti and Trentin, 2020] also optimized VAE models by minimizing KL divergence and reconstruction loss. This in turn ensures better generalization. More recently, [Shakya et al., 2024] compared traditional VAEs with β-VAEs and found that tuning β factor must be done with care so that disentanglement and reconstruction accuracy are both maintained. Even after all the mentioned advancements, though, to this day no such work has achieved the integration of β-VAEs into the multimodal learning pipeline for compositional generalization and symbol grounding. This work builds upon earlier findings by exploring the capability of β-VAEs to produce meaningful latent representations with which a Seq2Seq can

learn to map between video and audio inputs.

1.3 Research Questions, Hypothesis, and Methodological Contributions

This investigation is motivated by the following research questions:

1. What is the most suited β-factor that best balances the disentanglement and reconstruction processes in the latent space?

2. Can a β-VAE's structured latent representations be used to make symbol grounding more effective when used in a Seq2Seq model?

3. Do the structured latent representations from a β-VAE successfully allow generalization to completely unseen object-action combinations when used in a Seq2Seq model?

This investigation hypothesizes that a β-VAE based model for disentangled representation learning is able to establish connections between visual and auditory inputs without requiring explicit supervision. The initial methodology required the incorporation of a Dual-Stream Transformer together with Cross-Attention Congruence Regularization (CACR), in addition to a β-VAE, to achieve cross-modal alignment. However, the experiment results show that the latent vectors taken only from the β-VAE are sufficient to learn symbol grounding and to enable compositional generalization. Therefore, this work focuses on using these latent vectors as part of an adapted Seq2Seq system from [Ponte and Rauchas, 2022] to achieve better symbol grounding and ultimately enable compositional generalization by accurately mapping video features to their respective audio components.
The following contributions are made to test the hypothesis and answer the research questions above:

1. A rigorous experimental analysis of various β-factor settings was performed to determine the best parameters that could properly balance disentanglement and reconstruction in the latent space. Careful tuning of the β-factor is important to ensure that the latent space maintains meaningful relationships and desired level of detail required to reconstruct accurately.

2. The initial approach included a Dual-Stream Transformer coupled with Cross-Attention Congruence Regularization (CACR) and a β-Variational Autoencoder (β-VAE). However, the empirical results show that the β-VAE produced latent vectors by themselves are sufficient to represent the inherent relationship between the visual and the audio inputs. This improvement in the overall performance of the framework is based majorly on the structured latent space developed with the help of β-VAE.

3. This methodology uses these carefully tuned latent vectors in a modified Seq2Seq architecture—building on the system laid out by [Ponte and Rauchas, 2022]—to allow the translation of video features to their respective audio components. This end-to-end learning system improves symbol grounding and enables compositional generalization.

4. As mentioned in Section 3.3, a larger test set was used in which the model was tested on completely unseen object–action pairs. The findings from the test support the claim that the disentangled latent representation improves symbol grounding and enables sufficient

generalization outside the training set, and thereby support the hypothesis of enabling compositional generalization.

This modified methodology, which passes the latent vectors from the β-VAE into a Seq2Seq setup, shows better modality alignment, as also displayed in Table 3 showing lower cosine distance compared to the baseline method. The results confirm that the β-VAE based method significantly improves symbol grounding as well as the ability of the model to generalize to new compositions. The outcomes show that the proposed method successfully addresses the research questions by displaying strong symbol grounding together with enhanced compositional generalization.

2 Background/Related Work

Several works have been conducted in the past to develop systems that can ground symbols effectively. However, a smaller percentage of works involve testing with multimodal data and compositional generalization. This section assesses the existing research and classifies it into two main categories–Symbol Grounding in AI Models, and Compositional Generalization and Learning in AI Models. After that, it will discuss the workings of the β-VAE model and other considerations that went into this study.

2.1 Symbol Grounding in AI Models

Several models have attempted to achieve symbol grounding by using deep learning techniques. One such significant contribution in this area was provided by [Ponte and Rauchas, 2022], where they used a Sequence-to-Sequence (Seq2Seq) model to ground words in visual interaction. This model is trained on a custom dataset that includes five objects and five actions corresponding to each object. The dataset includes videos paired with synchronized audio descriptions that either describe the motion or speak the name of the object.

The Seq2Seq model used in this research utilizes video feature representations as inputs and seeks to align them with their corresponding audio features. In order to help the extraction of these feature representations, Contrastive Language-Image Pretraining (CLIP) is used for video processing and Wav2Vec for audio processing. The features are then converted into one-hot encoded integers before being passed into the model. The inherent encoder-decoder structure of the Seq2Seq model helps in the learning process by encoding video features into a latent space before decoding them into corresponding audio descriptions.

The results show that the Seq2Seq model displays symbol grounding abilities, as it is able to recognize correlations between video features and their corresponding auditory representations. This is reflected in its generalization abilities across paired modalities. However, the model suffers from a strong limitation in terms of compositional generalization. In particular, when tested in new situations involving unseen combinations where specific object action pairs were not included in the training data, the model did not generate correct descriptions. This shows that the model is not using the components it learned to construct new outputs, even though it has performed well in translating video features into their corresponding audio descriptions during symbol grounding tests.

Another experiment by [Li et al., 2024] proposes a softened symbol grounding approach which bridges neural network training and symbolic reasoning. The work creates symbolic states using a Boltzmann distribution, a statistical mechanics probabilistic model describing how systems distribute over different states based on energy levels. Here, it helps the system to calculate probabilities for different symbolic interpretations rather than arriving at one fixed mapping. The constrained symbolic AI models require exact mappings; this approach supports more flexible knowledge representation.

Transitions within the symbolic space require a Markov Chain Monte Carlo (MCMC) algorithm, a computational technique to sample from computationally costly probability distributions when direct computation is not practical. Since the Boltzmann distribution includes probabilistic symbol grounding, MCMC enables efficient search among different suitable representations for the most likely symbolic interpretation. This is followed by an SMT (Satisfiability Modulo Theories) solver, a solver based on logic that determines whether a specified set of constraints is satisfiable. When translating symbols, there are certain conditions based on logic

which must be true (e.g., "a cat is an animal" must always be the case). The SMT solver makes sure that the symbols provided are logically consistent.
One of the main advantages of this approach is that it is capable of more flexible and probabilistic symbol grounding compared to fixed rule-based systems. One major disadvantage, though, is that the framework depends on pre-defined symbolic reasoning modules, so it is less adaptive and its ability to generalize to new contexts is limited. Another disadvantage is that this approach does not learn symbolic rules autonomously, so it is not adaptable.
Another recent method by [Hamilton et al., 2024] presents DenseAV, a self-supervised model that learns cross-modal relations between video and audio features. The model uses DINO (Self-Distillation with No Labels) and HuBERT (Hidden-Unit BERT), two prominent self-supervised learning models in computer vision and speech processing, respectively. DINO helps neural networks learn object and pattern recognition from raw images without the need of human-labeled data, thus being suitable for symbol grounding in visual representations. HuBERT learns speech representations from unlabeled data by grouping audio features into separate independent units, making symbolic reasoning in speech tasks possible. While these models were state-of-the-art when they came out, more recent advancements such as DINOv2 and wav2vec 2.0 have since improved upon them. DenseAV uses multi-head aggregation mechanism to classify between spoken word and common ambient noises. This improves the capability of the model to associate verbal data as well as non-verbal data with visual interaction. Although DenseAV has the advantage for cross-modal learning, compositional generalization cannot be achieved because the model cannot predict new object-action pairs based on its training. The results show error in processing new combinations of rare objects and actions.

2.2 Compositional Generalization and Learning in AI Models

Compositional generalization, as defined in Section 1.0.1, was tackled by [Hemadri et al., 2024], where they proposed an intelligent communication system using CNN-VAE (Convolutional Neural Network-Variational Autoencoder) architecture. The system introduces ϵ-VAE, a variant of VAEs designed to improve latent space disentanglement. This results in a highly organized latent representation with minimal errors in learned representations. But the disadvantage with it is that the model was not trained in dynamic scenarios where objects and actions can change over time. Therefore, even though the system generalized within its training environment, it was never tested on new compositions, limiting its application in the real world setting.
Another approach was proposed by [Hristov et al., 2018], who introduced a framework for interpretable latent space learning with β-VAEs. The method in the study used weak supervision, forcing the model to encode specific properties such as object size, color, or shape in separate latent dimensions by using supplementary classification tasks. This organized latent representation led to improved generalization and interpretability in the model. However, the model was not tested with higher-dimensional multimodal inputs, and hence it cannot be determined whether it would generalize well to both visual and auditory signals.
In another attempt, [Asperti and Trentin, 2020] proposed a method for balancing reconstruction loss and KL divergence in VAEs. The KL divergence is a measure of how much the learned latent distribution varies from a specified prior distribution, and it is important to have an appropriate balance to avoid over-regularization or poor reconstructions. The method calculates KL divergence directly from mean reconstruction error, without decreasing generative quality while building a structured latent space. The results from the study show that the system can balance the loss terms effectively, leading to faster convergence and improved generative

quality. This paper suggests that adaptive control of KL divergence is an important factor to achieve improved generalization without losing reconstruction accuracy.
Similarly, [Shakya et al., 2024] compared baseline VAEs to β-VAEs, pointing out that increasing values of β-factor promotes disentangled latent spaces but that excessive disentanglement can cause poor-quality reconstructions. The study focuses on the importance of balancing these competing objectives by fine-tuning the β-factor. Also, [Burgess et al., 2018] describes β-VAE as a rate-distortion problem and showed that larger values of β-factor produce more independent and interpretable latent representations, but it will have lower reconstruction quality. [Higgins et al., 2016] followed up by showing that β-VAEs can learn disentangled representations without supervision, forcing the model to disentangle various abstract components in the latent space. From the findings reported in these studies, it can be safely assumed that latent space structuring is important to enable compositional generalization.

2.3 β-VAE Model

Variational autoencoders are well-known models that are capable of encoding input data into a well-structured latent space. They provide clear separation of features in a latent space. A β-VAE improves the capabilities of a VAE by introducing a β-coefficient in the model. A low β-factor value close to 1 favors the model by focusing on reconstruction accuracy; however, such prioritization can affect the ability of entangled representations to classify between meaningful features. A high β-factor value instead forces the latent space towards disentangled representations but reduces the model's ability to focus upon detailed complexity during the reconstruction process. A dynamic β-factor which increases gradually might help overcome such trade-offs. Such a strategy ensures the early stages of model training focus upon reconstruction and later shifts attention towards disentanglement and the creation of structured latent representations.
The β-VAE model used in this study was developed by learning from other available implementations and online tutorials. The initial reference was taken from an online tutorial on Medium[1], which provided insight into the architecture and training dynamics of the β-VAE model. For further optimized implementation, two other open-source repositories on GitHub by matthew-liu[2] and 1Konny[3] were studied, with differences in model structure, hyperparameter tuning, and optimization methods. By learning from these sources, a custom β-VAE implementation was built especially adapted for the needs of this work. The final model uses adjustments in the encoder-decoder architecture, optimization of the latent space, and β-factor scheduling. This ensures desired level of disentanglement and reconstruction trade-off.
The objective of the model is to learn structured latent representations of the video and audio features, which are then extracted and saved in the form of latent vectors. These latent vectors are then used by the Seq2Seq model for testing compositional generalization. The audio and video features are processed by separate encoders into a shared latent space, which learns the nature of interaction between the two modalities. Both encoders consist of a series of 1D convolutional layers followed by batch normalization and max-pooling. An 8-dimensional latent space is provided for each modality (video and audio) to map the encoded representations. Together, it results in an output of 16-dimensional latent vectors, which are computed using

[1] https://medium.com/@rahuldasari7502/building-a-beta-variational-autoencoder-%CE%B2-vae-from-scratch-with-pytorch-c5896ecc4dee
[2] https://github.com/matthew-liu/beta-vae
[3] https://github.com/1Konny/Beta-VAE

the mean and log variance of both audio and video features.
Hence, the latent distributions can be defined as:

$$\mu,\ \log \sigma^2 = \text{Encoder}(X)$$

where X is the input feature vector (in the form of video or audio) and μ and σ^2 represent the parameters of a latent Gaussian distribution.
The latent variables z are then sampled using the reparameterization trick, which can be defined as:

$$z = \mu + \epsilon \cdot \exp\left(0.5\ \log \sigma^2\right)$$

where ϵ is randomly sampled from a standard normal distribution.
The decoders are also designed similarly but are used to recover the original input (audio/video) features from their corresponding latent vectors. To preserve the meaningful relationship between multimodal features in the latent vectors, the video and audio decoders work independently.
There are three main loss functions used in the model, which help regulate the balance between disentanglement and reconstruction in the latent space:

1. The reconstruction loss is used to control the quality of the reconstruction of the input features by the decoder. Mean Squared Error (MSE) is calculated after every training epoch to analyze the reconstruction loss.

2. The KL (Kullback-Leibler) Divergence loss helps structure the latent representations by preventing over-fitting and forcing meaningful feature alignment between the two modalities[Asperti and Trentin, 2020].

3. Capacity-based loss for KL balancing: This loss gradually increases the KL divergence, which keeps the latent representations structured but still meaningful.

Therefore, the total loss function for the model can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \cdot |\mathcal{L}_{\text{KL}} - C|$$

where C is an adaptive capacity term, $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, $\mathcal{L}_{\text{KL}}$ is the KL divergence loss, and $\mathcal{L}$ is the total loss.

2.4 Other Considerations

Lastly, prior to the selected dataset, the something-something v2 dataset from [Goyal et al., 2017] and creating PyBullet simulations for data were also considered for the experiments. PyBullet was not required since no physics-based interactions were specifically desired. Something-something v2 dataset, while perfect for symbol grounding experiments and compositional semantics, lacks any corresponding audio. Nvidia Tacotron2[4] and Google TTS[5] were considered to generate synthetic speech from the textual labels but could not be used due to computational constraints. In lieu of that, the research is based on a custom-built symbol grounding dataset enabling controlled experiment trials.

[4] https://catalog.ngc.nvidia.com/orgs/nvidia/resources/tacotron_2_and_waveglow_for_pytorch

[5] https://pypi.org/project/google-tts/

This thesis builds on the above mentioned findings by integrating a β-VAE with a Seq2Seq model to tackle the problem of symbol grounding and compositional generalization. By using structured latent representations and evaluating the model on unseen object-action combinations, this work will attempt to close the gap between multimodal learning and generalization.

3 Methods

3.1 Baseline Model

The Seq2Seq model in the previous research by [Ponte and Rauchas, 2022] is used as the baseline model for comparing the results of our method. As described in the Section 2.1, it uses raw feature files extracted from CLIP for video and from Wav2Vec for audio. These feature sets are preprocessed and converted into one-hot encoded integers. The model then attempts to translate video to audio directly from the features. As we detail in Section 4.4, it successfully performs symbol grounding but fails to show any signs of compositional generalization when tested on unseen object-action pairs.

3.2 Proposed Model with β-VAE

Instead of using the raw features directly in the Seq2Seq model for mapping video to audio, the β-VAE first learns a latent representation of these features in a shared latent space. It uses separate encoders for both video and audio, which makes sure that each modality learns a well-structured latent representation before passing through the Seq2Seq model.
The main improvements are:

1. Semantic representation is improved through latent space learning.
2. Generalization is improved by forcing regulated disentanglement in the learned representations.
3. An additional layer of abstraction is introduced by using learned representations instead of raw features.
4. β-factor optimization is used to balance reconstruction quality and disentanglement in the latent space. Several ways of adjusting the β-factor are tested before finalizing the most suitable β-VAE model.

To answer the research questions and test the hypothesis as described in Section 1.3 , the experimental setup is divided into three stages which are: (i) Data and Preprocessing, (ii) Learning Structured Representations with β-VAE, (iii) Cross-Modal Translation using Seq2Seq.

3.3 Data and Preprocessing

[Ponte and Rauchas, 2022] created the dataset used in this study. It was created specifically for their own experiments that also investigate symbol grounding and compositional semantics. The dataset uses both human-generated and artificial voices for the audio. The videos are 3 seconds long and have a resolution of 180×180 pixels, and the audio signals are sampled at 16 kHz. [Ponte and Rauchas, 2022] also describes applying several data augmentation techniques that further increase the dataset's variability. In total, it is an audiovisual dataset that comprises 36,000 object-action combinations and is available publically for download[6].
The original study uses only 14,500 audio-video pairings from the dataset. One of the evaluation methods in the study requires decoding the feature sets back to text using Wav2Vec so that they can be manually inspected. Therefore, the videos in which voices were unclear, either due

[6] https://www.kaggle.com/datasets/fabiodeponte/symbolgrounding

to accent or the data augmentation technique used, were removed from the original study. However, my study uses the full dataset which provides a larger training set and test set for evaluation.
Each video in the dataset is paired with its corresponding audio, which helps define what is happening in the video. For example, if it is an object like a pen, or if an action is being applied to an object like a pen, the audio helps clarify what is happening in the video.
The dataset consists of five objects that appear in the following order: (i) Pen, (ii) Phone, (iii) Spoon, (iv) Knife, (v) Fork
It consists of five actions, applied to each object, that appear in the following order: (i) Left, (ii) Right, (iii) Up, (iv) Down, (v) Rotate
Hence, the audio files paired with the videos contain descriptions such as "This is a pen" for objects and "Move the pen to the left" for actions performed on those objects. The data set structure consists of five instances of each object, followed by five actions on each object. Upon manual inspection of the data set, it is observed that this pattern repeats every 50 iterations for 36,000 rows. This data set is further prepared for experiments by normalization, segmentation, and group-based data splitting for training and testing.
First, to maintain consistency, the raw video and audio features are min-max normalized to keep the values between 0 and 1. This normalization is performed using the scikit-learn MinMaxScaler, and the output feature arrays are saved for the next step. Second, it is important that the video-audio inputs are temporally aligned so they are segmented into 30 time steps, where each time step represents a window of 0.1 seconds. Each sample is either truncated or padded to match the sequence length. The final outputs are saved for group-based splitting.
The group-based splitting approach is used to maintain balance in the training and test data set. This makes sure that the training set does not contain too many instances of "still objects" while the test set contains only "actions on the objects" or vice versa. This approach also allows for separating "still object" instances from testing and "actions on the objects" instances from training, for every object in the dataset. Thus, it enables testing for compositional generalization on data that is "almost" truly unseen. The data set is divided into an 80/20 ratio, with 80% allocated for training and 20% for testing. The samples are assigned to the training and test splits using the GroupShuffleSplit function, which prevents overlap between groups.

3.4 Learning Structured Representations with β-VAE

A β-VAE is used for learning structured multimodal representations for improving symbol grounding and enabling compositional generalization. The β-VAE acts as a "dimensionality reduction" method to map unprocessed audio and visual features into a shared latent space and later pass them through a sequence-to-sequence (Seq2Seq) model for cross-modal translation. Rather than directly passing raw feature representations into the Seq2Seq model, the β-VAE first maps the features into a reduced dimension latent space. This type of representation suppresses the undesired information which is irrelevant to the semantic meaning of the data. It also improves interpretability by enforcing controlled disentanglement of features and enables compositional generalization by forcing the model to learn useful structured representations instead of just relying upon simple raw correlations.
The model uses two encoders: the video encoder maps features consisting of 17 channels from the video into a latent space with 8 dimensions, and the audio encoder maps features consisting of 5 channels from the audio into a different latent space with 8 dimensions. After the encoding

process, the latent representations obtained from the two modalities are combined into a single 16-dimensional latent space that captures the relational structure between the two streams. In order to create a meaningful and structured latent space, several configurations for the β-factor are tested. The list of different configurations tested are as follows:

1. Linear Scheduling: β increases linearly over epochs.
2. Fixed Increment: β increases in fixed step sizes until a max threshold.
3. Incremental Step: β increases in small, discrete steps.
4. Epoch-Based Scaling: β scales proportionally to epoch count.
5. Alternative Fixed Increment: Similar to Fixed Increment, but with smaller step sizes.
6. Gradual Ramp-Up: β slowly increases over time, ensuring a smooth transition.
7. Slow Increase Over Time: A slower version of the Gradual Ramp-Up.
8. Higher Fixed Increment: A more aggressive increase in β.
9. Exponential Decay: β starts high and gradually decreases over time.

As we can observe in the Table 2, the final selected choice is Gradual Ramp-Up, which after thorough testing of all strategies turned out to be the best configuration. It improved performance because of its ability to allow the model to focus first on accurate reconstruction during the initial training phase. Gradual Ramp-Up then introduces disentanglement restrictions incrementally, avoiding random changes that can destabilize learning. The approach enables structured latent representations to develop naturally, without any excessive loss of information.
In comparison with other methods, alternatives like Fixed Increment and Exponential Decay caused over-regularization, which is discouraging for reconstruction. Incremental Step, though more stable, lacked intricate control over disentanglement. Gradual Ramp-Up ultimately provided the best trade-off between disentanglement, reconstruction, and generalization.
As described in Section 2.3, the model also uses a capacity-based loss. Initial training experiments revealed that the KL divergence term often sees sudden abrupt increases in the middle stages of training, leading to challenges in learning and destabilization of the latent representations. To tackle this problem, a capacity loss term is included in the loss function. The additional term manages the effective capacity of the latent space by gradually increasing an adaptive clipping level across training epochs. The model thus focuses more on exact reconstruction in the early stages of training and later applies gradual increase in regularization of the latent space. This approach to capacity-based loss encourages structured and disentangled learning by the β-VAE without falling for over-regularization, thus improving training stability and feature learning standard.

3.5 Cross-Modal Translation using Seq2Seq

Before Seq2Seq model training, the latent vectors are extracted using the pre-trained β-VAE model. It is used to encode video and audio data into their corresponding latent vectors. The mean embeddings of both modalities are extracted to force stable feature representations. The

latent vectors are then stored and used as input for training the Seq2Seq model. The extracted latents are then reshaped into a sequence, with each sample having one timestep which prepares them for sequence learning. These latent vectors are then passed through a modified version of the Seq2Seq model that was used in the study by [Ponte and Rauchas, 2022]. The Seq2Seq model translates or maps the video latent vectors onto the audio latent vectors. In this process, the video features serve as input, while the corresponding audio features, predicted by the model, represent their translation into a different modality.
The model architecture consists of an encoder and a decoder. Both the encoder and decoder use a single-layer LSTM (Long Short-Term Memory) network. The encoder takes in the input latent vectors (video) to generate a context vector. The decoder is then initialized with the context vector from the encoder, which in turn helps the model reconstruct the audio latent vector sequence. The final output is generated using a dense output layer, which applies a linear transformation.

3.6 Evaluation Approach

For inference, the Seq2Seq trained model is used for generating audio representations from unseen video sequences. It involves passing a new video sequence through the trained Encoder LSTM for generating the context vector. The audio representation corresponding to it is generated through the decoder LSTM based on the co-relations which were learned earlier.
The experiments are evaluated using the following metrics to measure how well the predicted audio representation aligns with the actual ground truth:

1. Mean Squared Error (MSE) helps understand the difference between the original output and the model's prediction.

2. The alignment between predicted and original sequences is analyzed with the help of Cosine Distance.

3. To test for compositional generalization, the model's performance is evaluated on unseen object-action combinations.

The model's capability to display generalization on truly unseen combinations of object-action pairs is tested by holding out each object's action instances from the training set one by one. The training set contains all the "still" instances of the object, and the test data contains all the "action" instances of the object. This forces the learned representations, or the latent vectors extracted from the β-VAE model, to display the emergence of compositional generalization beyond the seen combinations in the training set.
For each "action" instance of an object held out from the training set, the experiment is run three times, and the average scores are reported in the Table 4 to reduce randomness. Cosine Distance is used to analyze the predicted outputs against the ground truth.

4 Experiments and Results

This section describes the setup and results of the experiments that test whether structured representations from a β-VAE help with symbol grounding and compositional generalization. It details the training configurations for both the β-VAE and the modified Seq2Seq model, along with implementation details and guidelines for reproducibility. The results are presented for various β-factor scheduling strategies, and the best-performing setup is analyzed using quantitative evaluations, including tests for symbol grounding and compositional generalization with new object-action pairs.

4.1 Training Settings

Parameter	**β-VAE**	**Seq-to-Seq**
Learning Rate	0.0005	0.0003
Batch Size	64	64
Training Epochs	70	25
Loss Function	Recon Loss + KL	MSE + Cos Distance
Optimizer	Adam	Adam
Input	Preprocessed features	Latent vectors

Table 1: Training settings for β-VAE and Sequence-to-Sequence training.

4.2 Implementation Details

The β-VAE model is implemented using the PyTorch framework, and the Sequence-to-Sequence model has been implemented using TensorFlow/Keras. NumPy, scikit-learn, and Matplotlib are used for data management, debugging, preprocessing, and evaluation. The experiments are run in Google Colab notebooks and use the T4 GPU provided with the free version of Google Colab. All the training and testing processes are accelerated using CUDA wherever possible or required.

4.3 Reproducibility

The experiments were run on Google Colab from 14th January 2025 to 18th February 2025 in Leiden, NL. The training used a T4 GPU with 16GB of VRAM and approximately 25GB of allocated RAM. The dataset for these experiments is taken from the study by [Ponte and Rauchas, 2022] and is available publicly under Creative Commons License. All preprocessing steps, including normalization and segmentation, are run in Google Colab.
If running the experiments on a local system, the required software dependencies are Python 3.8 or higher, PyTorch 1.10 or higher, and TensorFlow 2.8 or higher. NumPy, scikit-learn, and matplotlib for preprocessing and visualization. The β-VAE model is trained using the configurations described in the methods section. The adapted Seq2Seq model uses latent representations from the trained β-VAE model instead of raw features. The latent vectors can be extracted and saved using the provided scripts. The storage requirements are between 30GB and 40GB for raw datasets, feature files, processed datasets, and model checkpoints. All the

mentioned codes above can be accessed via Github[7]. To replicate the experiments in Google Colab or a local machine:

1. Access Google Colab and ensure GPU acceleration is enabled (`Runtime > Change Runtime Type > GPU`) or load the required libraries in your preferred environment.
2. Download and preprocess the dataset using the provided scripts.
3. Train the β-VAE model and extract the latent vectors.
4. Train the Seq2Seq model using the extracted latent vectors.
5. Run evaluations and compare the results.

4.4 Results

The technical descriptions of β-configurations tested are as follows:

1. Linear Scheduling: min(0.05, epoch / 20)
2. Fixed Increment: min(0.2, epoch / 50)
3. Incremental Step: min(0.07, 0.005 * epoch)
4. Epoch-Based Scaling: min(0.07, epoch / 20)
5. Alternative Fixed Increment: min(0.07, 0.003 * epoch)
6. Gradual Ramp-Up: min(0.07, 0.0025 * epoch)
7. Slow Increase Over Time: min(0.07, (epoch / 30) * 0.07)
8. Higher Fixed Increment: min(0.1, 0.005 * epoch)
9. Exponential Decay: 0.07 * (1 - np.exp(-epoch / 10))

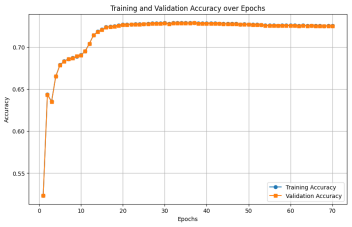
All the above mentioned configurations were tested with the β-VAE model to arrive at the best performing version of the model that can be used for extracting latent vectors for cross-modal translation experiments. The final β-factor configuration used is the Gradual Ramp-up strategy because it allowed for the most suitable balance between keeping reconstruction loss minimized and creating a structured latent representation of the video and audio inputs. Table 2 summarizes the results of all the configurations:

[7] https://github.com/vinnayakk/crossmodal_translation/

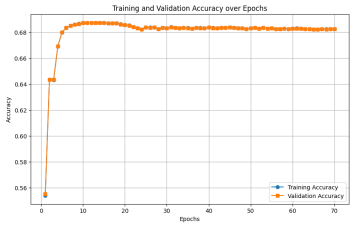
β Strategy	Train Acc.	Val. Acc.	Test Acc.	Test Cosine Distance
Linear Scheduling	71.59	71.56	72.76	27.24
Fixed Increment	68.07	68.06	68.34	31.66
Incremental Step	72.89	72.84	74.15	25.85
Epoch-Based Scaling	71.60	71.57	72.70	27.30
Alt. Fixed Increment	73.22	73.16	74.44	25.56
Gradual Ramp-Up	**73.38**	**73.31**	**74.55**	**25.45**
Slow-Incr. Over Time	73.43	73.37	74.45	25.55
Higher Fixed Incr.	71.80	71.76	72.57	27.43
Exponential Decay	72.22	72.17	73.09	26.91

Table 2: Comparison of different β scheduling strategies based on training, validation, and test accuracy, along with test cosine distance. The Gradual Ramp-Up strategy performs the best with the highest test accuracy (74.55%) and the lowest test cosine distance (25.45). In contrast, the Fixed Increment strategy shows the lowest test accuracy (68.34%) and the highest test cosine distance (31.66). These results highlight the importance of β tuning for optimal model performance.

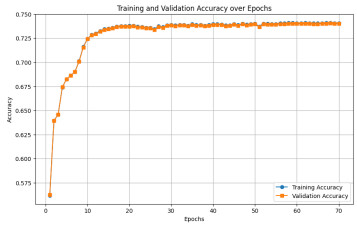
Below given are the Training and Validation Accuracy graphs for the β-VAE model:



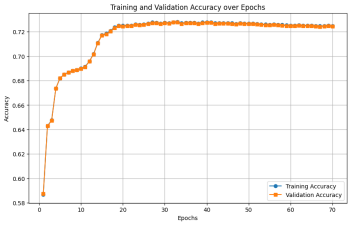
(a) Linear Scheduling



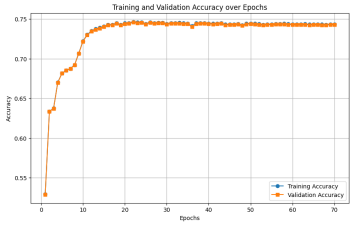
(b) Fixed Increment Scheduling



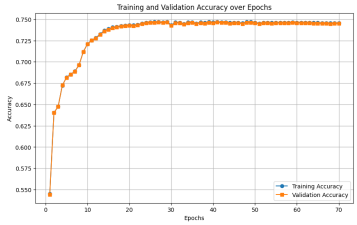
(c) Incremental Step Scheduling



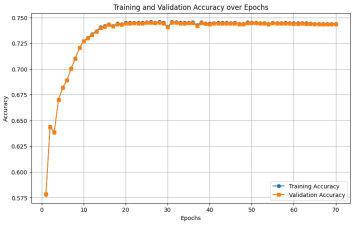
(d) Epoch Based Scaling Scheduling



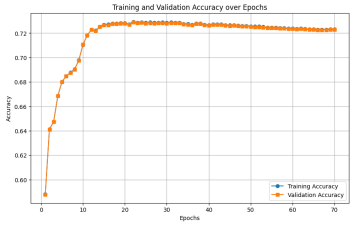
(e) Alternative Fixed Increment Scheduling



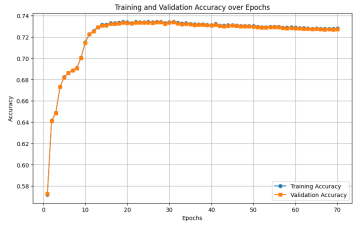
(f) Gradual Ramp-Up Scheduling



(g) Slow Increase Over Time Scheduling



(h) Higher Fixed Increment Scheduling



(i) Exponential Decay Scheduling

Figure 1: Scheduling Strategies

4.4.1 Symbol Grounding Tests

For the symbol grounding tests, the best model configuration from the previous research by [Ponte and Rauchas, 2022] is compared with the proposed model. Table 3 shows the performance comparison between the baseline model (Seq2Seq with raw features) and the proposed model (β-VAE + Seq2Seq):

Model	**Baseline (Seq2Seq)**	**β-VAE + Seq2Seq**
Test Cos Distance	50.12	0.17

Table 3: Comparison between the baseline Seq2Seq model and the proposed β-VAE + Seq2Seq approach based on the test cosine distance value. The baseline model shows a decently higher test cosine distance (50.12), indicating poor alignment between predicted and actual representations. In contrast, the β-VAE + Seq2Seq model achieves a noticeably lower cosine distance (0.17), showing better cross-modal translation. These results highlight the importance of latent representations for improving multimodal learning.

4.4.2 Compositional Generalization Tests

Table 4 shows the performance comparison of the model on holding out the objects one by one during compositional testing:

Model	**Baseline (Seq2Seq)**		**β-VAE + Seq2Seq**	
	Test Loss	**Cosine Distance**	**Test Loss**	**Cosine Distance**
Object 0	0.73	59.81	0.0016	1.90
Object 1	0.80	66.28	0.0040	3.23
Object 2	0.73	59.59	0.0029	2.64
Object 3	0.86	64.59	0.026	8.88
Object 4	0.97	61.47	0.021	6.01

Table 4: The baseline model displays higher cosine distances, ranging from 59.59 to 66.28. In comparison, the β-VAE + Seq2Seq model achieves noticeably lower cosine distances across all objects. The lowest cosine distance is observed for Object 0 (1.90), while the highest is for Object 3 (8.88). In addition, it can be observed that the test loss for Seq2Seq increases dramatically during compositional generalization testing. However, the test loss remains low and stable for the β-VAE + Seq2Seq model when tested for compositional generalization.

The table above shows that the Seq2Seq model with latent vectors from the β-VAE shows significantly better compositional generalization compared to the baseline model. For each held-out object, the baseline model has higher test losses and cosine distances, indicating that it struggles to generalize effectively to unseen object–action combinations. In comparison, the proposed model consistently achieves lower cosine distances and maintains lower test losses. This suggests that the structured latent representations learned from the β-VAE allow the model to capture and reuse underlying meaningful patterns more efficiently. By separating and compressing the multimodal inputs into a meaningful shared latent space, the model

is able to infer new combinations that go beyond its training data. The results show that adding a disentangled latent space improves accuracy and the system's capacity for learning abstract reasoning across different modalities, which is important for achieving compositional generalization.

5 Discussion

The initial inspiration for this study was to explore emergent behaviors in an artificial creature in a dynamic soundscape. The soundscape could be either real-time or curated and acts as an input for the artificial creature. The assumptions were that all of the mentioned settings would trigger reactions from the artificial creature, thus leading to some novel behaviors or situations. However, after further research, it was found that there is a strong need to establish a system that could enable such an interaction. The artificial creature needs to understand the environment in a meaningful way in order to begin interacting. This led to further investigation into the topics of symbol grounding and compositional generalization.

As a result, this study investigated whether structured latent representations help to improve symbol grounding and enable compositional generalization in multimodal systems. The β-VAE model helped to disentangle video and audio features in a shared latent space. This allowed the system to connect "what it sees with what it hears". The model learns an abstract representation of the video and audio features, which provides the capability of better interpretation and flexible mapping between the two modalities.

One takeaway is that this approach offers a fundamental benefit over the method of using direct feature mappings. When tested with unseen object-action pairs during video-to-audio mapping using Seq2Seq, the model showed strong signs of compositional generalization in the results. This suggests that a structured representation of disentangled features in a latent space can significantly improve the generalization capabilities of a system. For AI systems that are used in the real world, this can be very helpful. Several areas in the fields of human-computer interaction and robotics can benefit from this method to process dynamic real-world situations where it is required to deal with multisensory input while understanding meaningful relationships between them.

Another important takeaway is the fine-tuning of the β-factor. A balance between the precision of reconstruction and disentanglement is very important for an effective evaluation of the model's ability to identify meaningful relations. A β-factor that is too high creates excessively abstract representations, thus losing vital information. However, too-low β-factor gives rise to entangled representations, thus reducing the effectiveness of the model during generalization tests. The experiments performed show that gradual increases in β-factor growth result in the best performance. The improvement in performance occurs because it allows the model to focus first on producing high-quality reconstructions before it can be subjected to disentanglement constraints. However, one challenge that comes with fine-tuning of the β-factor is the manual tuning of it. Although the approach used in our experimental setup is effective, different learning environments and datasets can require different β schedules. Manual fine-tuning of the β-factor with each experiment and dataset, every time, can become a long and time-consuming process.

Despite the positive findings mentioned above, it is very important to identify and address the limitations of this study. One of the main limitations is the use of a small and predefined object-action dataset that has been specifically developed for symbol grounding and compositional generalization experiments. Although this controlled dataset makes it easier to experiment and analyze, it can still fail to capture the dynamic complexities that are present in real-world settings. Similarly, another limitation that deserves attention is the handling of noisy and unclear inputs. The assumption of the proposed system that the video and audio inputs are clear and preprocessed places constraints on it. Inputs in real-world applications are dynamic and can vary considerably in terms of diversity. They often get subjected to lags,

distortions, and/or background noise. AI systems applied in physical environments need to have the ability to adapt to low-quality inputs without reducing their understanding at the compositional level. In addition, while the study shows positive improvement in generalization ability, the model has not been tested in the real world, or even in dynamic settings. Many real-world applications, including robotics and human-computer interaction, require models that are able to properly handle continuously changing input. Another challenge, which is related to its practical applications, is the understanding of human-level interpretation of the learned latent representations. The results show that a structured latent space enhances generalization capabilities, but we also need methods to understand what every dimension encodes. Since abstractions are not clearly defined in β-VAE, the use of latent probing techniques can help to establish an understanding of the semantic content of the learned features.

Lastly, it is necessary to test the suitability of this approach outside of the video-to-audio translation scenario. Although this study focused on cross-modal relationships between video and audio signals, the method has not been tested with textual data, haptic inputs, or other sensory inputs used for multimodal interactions in robotics. It will be particularly helpful in tasks involving natural language grounding or robotic perception, where such systems are required to analyze and make sense of the complex environmental stimuli.

6 Conclusion

This study investigated the role of learned latent representations through β-VAE in improving symbol grounding and compositional generalization in multimodal systems. It follows the process of encoding video and audio features in a latent space which is followed by cross-modal translation using the Seq2Seq model. The results show improvement in alignment and meaningful connections between video and audio signals.

The first research question explores the best β-factor schedule to establish a balance between the disentanglement of latent space and the accuracy of reconstruction. The results from the experiments showed that the Gradual Ramp-up strategy is the most optimal for our experiments, followed by Alternative Fixed Increment and Slow Increase Over Time. It allows the model to first focus on reconstruction accuracy and then introduce constraints for disentanglement, thus avoiding information loss. As a result, this approach learns meaningful relationships between video and audio features, which helps to improve generalization.

The second research question investigated the improvement in symbol grounding when latent representations learned from the β-VAE model are used in the Seq2Seq model for cross-modal translation. The findings suggest that latent representations are better at understanding compositional relationships compared to direct feature mapping. It reduces redundancy, and the latent space helps in separating meaningful features from noise. This results in improved symbol grounding across video and audio inputs.

The third research question in this study addresses the system's ability to display compositional generalization when exposed to completely new object action instances during the testing phase. The results of the experiments offer strong evidence for successful compositional generalization, which shows that the proposed model with β-VAE outperforms the baseline model using raw feature mappings in a Seq2Seq model. The higher accuracy on unseen combinations and lower cosine distance figures indicate that the model is able to apply its learning of meaningful relationships beyond the training dataset.

There are many directions for future work that can arise from the findings achieved in this research. One important direction is an increase in the size and variety of training datasets used in the experiments. Future research should consider using larger and more diverse datasets, such as Something-Something V2. A dataset like Something-Something V2 contains more than 200,000 videos of "something being performed on something." The dataset has been created using videos from online sources that realistically reflect the real-world situations. Furthermore, using datasets that cover a larger variety of environmental settings and types of object interaction can help the model become more adaptable and responsive when testing its ability to generalize to complex, real-world object-action situations. Another area of importance for future work is the use of automation for tuning the β-factor. The use of automation through adaptive learning rate methods or optimization using reinforcement learning can help the system become more efficient, adaptable, and scalable. This will make it possible to implement the system across different datasets and modalities with ease and in a reasonable time. In addition, interpretability of the latent space is an open area of research. Future work can focus on exploring the model with a visualization approach or latent probing techniques. Different metrics can also be explored to measure disentanglement. This will help to analyze how individual latent dimensions reflect interpretable and compositional characteristics. Better understanding of the learned representations might be helpful in training more interpretable and transparent AI models. We can generalize this solution to other learning problems in addition to video-to-audio mapping. The method used in this research can also be suitable

for text-to-image grounding, tactile learning, and multimodal perception of robots. Future work can focus on these novel applications, which will help in acquiring an understanding of structured representations for other types of sensory input.

Lastly, future work can also explore how this method can be translated to the real world, especially for robotic simulations and interactive AI environments. It would be insightful to investigate and implement the model in an environment where a robot learns the object-action associations from that environment and can then apply this grounding knowledge to another robot through the use of transfer learning. We must explore how structured latent representations help in knowledge sharing and cooperation in artificial agents. This would help bridge the gap between AI research and real-world applications in dynamic environments. By focusing on the steps outlined above, the future work can significantly contribute towards smarter AI and multimodal systems. This advancement will enable AI to become more adaptable, more understandable, and able to generalize in dynamic real-world situations.

Bibliography

References

Andrea Asperti and Matteo Trentin. Balancing Reconstruction Error and Kullback-Leibler Divergence in Variational Autoencoders. *IEEE Access*, 8:199440–199448, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3034828. URL https://ieeexplore.ieee.org/document/9244048. Conference Name: IEEE Access.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL https://aclanthology.org/2020.emnlp-main.703/.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $$-VAE, April 2018. URL http://arxiv.org/abs/1804.03599. arXiv:1804.03599 [stat].

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent Independent Mechanisms, November 2020. URL http://arxiv.org/abs/1909.10893. arXiv:1909.10893 [cs].

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, June 2017. URL http://arxiv.org/abs/1706.04261. arXiv:1706.04261 [cs] version: 2.

Mark Hamilton, Andrew Zisserman, John R. Hershey, and William T. Freeman. Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Sound and Language, June 2024. URL http://arxiv.org/abs/2406.05629. arXiv:2406.05629 [cs].

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1): 335–346, June 1990. ISSN 0167-2789. doi: 10.1016/0167-2789(90)90087-6. URL https://www.sciencedirect.com/science/article/pii/0167278990900876.

Raghu Vamshi Hemadri, Akshay Rayaluru, Rahul Jashvantbhai Pandya, and Sridhar Iyer. AEVBComm: an intelligent communication system based on β-VAE. *CSI Transactions on ICT*, 12(4):107–118, December 2024. ISSN 2277-9086. doi: 10.1007/s40012-024-00401-9. URL https://doi.org/10.1007/s40012-024-00401-9.

I. Higgins, L. Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016. URL https://www.semanticscholar.org/paper/beta-VAE%3A-Learning-Basic-Visual-Concepts-with-a-Higgins-Matthey/a90226c41b79f8b06007609f39f82757073641e2.

Yordan Hristov, Alex Lascarides, and Subramanian Ramamoorthy. Interpretable Latent Spaces for Learning from Demonstration, October 2018. URL http://arxiv.org/abs/1807.06583. arXiv:1807.06583 [cs].

Brenden Lake and Marco Baroni. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2873–2882. PMLR, July 2018. URL https://proceedings.mlr.press/v80/lake18a.html. ISSN: 2640-3498.

Zenan Li, Yuan Yao, Taolue Chen, Jingwei Xu, Chun Cao, Xiaoxing Ma, and Jian Lü. Softened Symbol Grounding for Neuro-symbolic Systems, March 2024. URL http://arxiv.org/abs/2403.00323. arXiv:2403.00323 [cs].

Gary Marcus. Deep Learning: A Critical Appraisal, January 2018. URL http://arxiv.org/abs/1801.00631. arXiv:1801.00631 [cs].

F. Ponte and S. Rauchas. Grounding Words in Visual Perceptions: Experiments in Spoken Language Acquisition. 2022. URL https://www.semanticscholar.org/paper/Grounding-Words-in-Visual-Perceptions%3A-Experiments-Ponte-Rauchas/5f69235e767d2a7401fc6eff0d45038fd9a4f378.

Sahaj Shakya, Binod Maharjan, and Prabesh Shakya. From Entanglement to Disentanglement: Comparing Traditional VAE and Modified Beta-VAE Performance. *International Journal on Engineering Technology*, 2(1):38–48, December 2024. ISSN 3021-940X. doi: 10.3126/injet.v2i1.72491. URL https://www.nepjol.info/index.php/injet/article/view/72491. Number: 1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].