AVJEPA: Exploring Multimodal Representations using Joint Embedding Predictive Architecture

BOLUO GE and JOHN ZHU, North Carolina State University, USA

In this study, we investigate the potential benefits of multi-modal feature prediction through self-supervised learning. We introduce Audio-Video Joint Embedding Predictive Architecture (AVJEPA), a self-supervised learning approach to generate quality multi-modal (audio and video) feature representations. We evaluate AVJPEA on video classification and video/audio mask prediction. Our results shows that single-modal representations outperformed AVJEPA on video classification tasks and we were able to produce low-fidelity multi-modal mask predictions. Our code is available at source code.

1 Introduction

Human neural systems demonstrate an extraordinary capacity to transform raw sensory input into rich semantic understanding. Critically, this is done without any explicit labels. Understanding the principles behind such unsupervised learning capabilities has been a big challenge in machine learning research. One compelling work is the predictive feature principle [8], which suggests that effective representations can emerge from learning to predict features across temporally adjacent stimuli. Building on this, Jepa series work [1, 3] explore feature prediction as an independent objective for learning visual representations from unlabeled vision data. This paper extends JEPA approach to multimodal learning, investigating whether feature prediction alone can effectively learn joint representations across both visual and audio domains. Further, we evaluate the capability of our model using two downstream tasks.

To that end, we introduce the AudioVideo joint-embedding predictive architecture or AV-JEPA, a framework that combines masked modeling prediction with joint-embedding prediction. Given computational constraints, we pretrain three variants of compact AV-JEPA models on 30,000 videos sourced from public datasets. We evaluate our models through fine-tuning on downstream tasks including video classification and video prediction.

2 Related Work

2.1 Self-Supervised Learning

Self-supervised learning is a fundamental but still growing collection of work. A close work to ours is data2vec[2] which presents a framework of self-supervised learning applicable across modalities based on prediction of masked input data. They also used the transformer architecture for their approach, and

Authors' Contact Information: Boluo Ge; John Zhu, North Carolina State University, Raleigh, NC, USA.

completely avoid handcrafted augmentations. In addition, the Joint Embedding Predictive Architecture introduced by Assran et al. [1] has demonstrated effective self-supervised learning for visual data by leveraging predictive coding principles.

2.2 Representation Learning

Previous work done in representation learning includes [9] who leverages self-supervised feature learn lip-reading and automatic speech recognition. Other work including [10] for LSTM unsupervised learning and [11] for unsupervised learning representations on unlabeled video data. These works examplify a family of approaches that apply a predictor network to map cross time-step representations, however these are trained on frozen encoders as compared to the JEPA approach of end-to-end pretraining. The usage of vision transformers [4] can be seen in [6] where vision transformers are used to learn spatiotemporal representations, effectively encapsulating video data. However, most core and related to this work is the JEPA as seen in [5] which not only describes JEPA but also higher order variations of the framework that can be potentially applied to achieve higher level reasoning through heirarchal representations.

2.3 Multimodal learning

As explored in works like CLIP [7], highlights the importance of aligning representations across modalities. This research laid the groundwork for integrating audio and vision data in a unified embedding space. Similarly, advances in transformer-based architectures, like those proposed by [4], have demonstrated the scalability and adaptability of self-supervised learning in both vision and audio domains.

3 Background

3.1 JEPA

The Joint Embedding Predictive Architecture [1] is a non-generative approach for self-supervised learning on data. The approach learns in *representation space* rather than in low-level outputs (pixel-level, text token-level, etc.). This flexibility allows JEPA trained models to serve as a generalized information encoder, where the latent space generated constitute quality representation of input data features. We describe the JEPA frame work as in [1]:

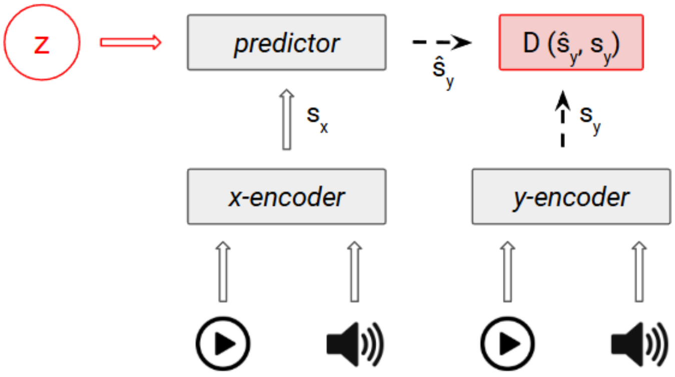
3.1.1 Target. Here, we describe the process of creating training targets in the JEPA framework. A non-transformed input x is fed through the *target-encoder* to produce a token level representation $s_y = \{s_{y_1}, \ldots, s_{y_N}\} \cdot s_{y_k}$ is the target-encoder generated representation of the k^{th} token. Without any transformation, the target-encoder is provided the full scope of information from our input.

3.1.2 Context. Here, we describe the functionality of the core of JEPA learning: Context Encoding. Rather than being provided the raw input x as in the target encoder, the context encoder is provided *transformed* data x'; in our case masked video and audiospectrogram data. As in the target encoder, a corresponding set of number of outputs $s_x = \{s_{x_1}, \ldots, s_{x_N}\}$ will be produced where s_{x_k} represents the k^{th} context embedded token. Given the transformation prior to input, the context-encoder is only provided a portion of information from the input.

3.1.3 Predictor. Here, we describe the predictor in JEPA. After being passed through the context encoder, we have a set of information s_x that represents only part of the input data. It is thus the job of the predictor to predict, *in the embedding space*, the missing pieces of generated by the target encoder s_y. Thus after the predictor pass we will have a context/predictor embedded representation $\hat{s}_y$ and a target encoder embedded representation. We learn on how well the context/predictor performed with the target-encoder representation as the target.

3.1.4 Training. While the context encoder is updated using traditional gradient approaches, to maintain difference across the target and context models, the target encoder's parameters are updated through a exponential moving average of the context encoder parameters. In this way, JEPA is able to avoid mode-collapse as constant output learning becomes impossible.

Fig. 1. AVJEPA Architecture



4 Methodology

4.1 Motivation

This paper aims to explore the construction and potential advantage of a multi-modal embedding space. Our strategy involves training single (video) and multi-modal (video and audio) JEPA architectures on a variety of model sizes, then evaluating these frozen encoder/predictors on classification and mask prediction downstream tasks. Inspired by the work in [7], by pretraining these models with paired vision-audio data, we hope these approaches can provide an understanding of the advantages and challenges of multi-modal embedding and reveal any emergent properties of such an embedding space.

5 Implementation

5.1 Framework

The code of this paper was built as a fork of the original VJEPA project. While VJEPA has publicly released code for its pretraining and evaluation oracles, encoder architectures, and video masking approach, it keeps the decoder architecture unreleased.

5.2 Models

The models evaluated in this project were: ViT-Tiny, ViT-Small, ViT-Base and custom Predictor ViT [4]. For downstream task learning, we employed a variety of linear/attention probes. For mask prediction, we used a simple attention probe consisting of single, linear projection, multi-headed attention, and quary pooling layers to decode representation space into pixel space.

5.3 Dataset

Given the ease of use of it's data container format (.mp4) we chose to use a subset of the Kinetics-400 dataset.

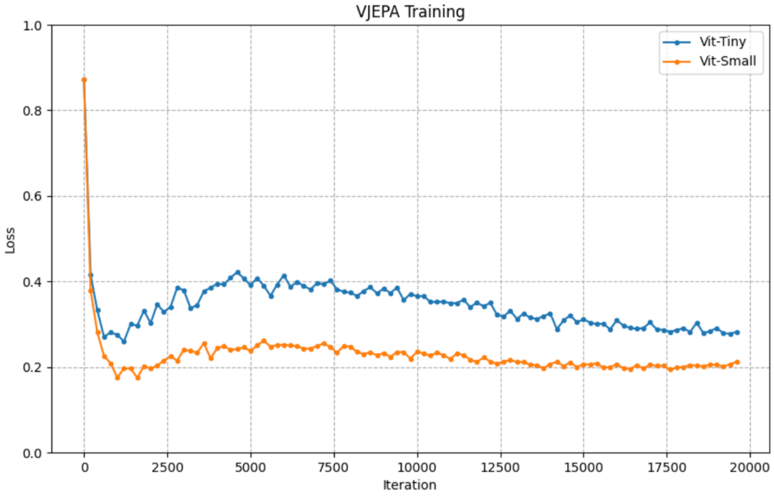
5.4 Challenges

There are 3 main points of challenge in this paper.

(1) Compute: Compared to previous works [3] we had access to significantly less compute available, in GPU processing, CPU processing and available memory. To deal with this, we chose smaller variations of models compared to previous as well as trained and evaluated on a smaller scale of data. During overnight training of a single model, we observed the double descent phenomenon (as illustrated in Figure 2). Based on this observation, we decided to early stop at iteration 2000.

(2) Multi-modal Engineering: Incorporating multiple modes of data, particularly those that are of differing dimensionality incurs challenges in maintaining quality tokenization and encoding of input data. This includes both loading data and performing necessary transforms such as masking strategy and positional encoding. Time-wise, this engineering task became the majority of our efforts. In order to implement AVJEPA, complete "refactoring" of existing training loops, mask generation objects and a new dataset object were needed. Existing implementation did not support the dynamic splitting of video/audio tokens required for our objectives, which we thus implemented as needed. We've had 3 major iterations of AVJEPA, first with no masking policy implemented for the audio data, a second failed attempt with audio-masking (we used a single mask set per batch with no truncation, however

Fig. 2. Double Descent



we saw no convergence and consistent loss oscillation during training), and a third successful attempt using unique mask set per entry in batch with truncation of masks for shape matching.

(3) Downstream Task Evaluation: Fine-tuning probes/decoders for video classification and mask prediction is especially compute and design intensive. Firstly, for classification, memory and compute issues are common as the VJEPA/AVJEPA models handle only a few frames at a time, making downstream learning complex. For mask prediction, there is no provided decoder architecture and thus we must experiment and find a suitable architecture and fine-tuning approach.

5.5 Multi-Modal Integration

We incorporated audio into the existing VJEPA implementation through a few key steps:

(1) Audiospectrogram: We extract both video and audio output from the .mp4 format. Audio data is extracted, then cut to reflect the same temporal region covered by the 16 video frames. An audiospectrogram is then generated using short-time fourier transform and Mel spectrogram conversion. The result is the interpolated to fit the input shape.

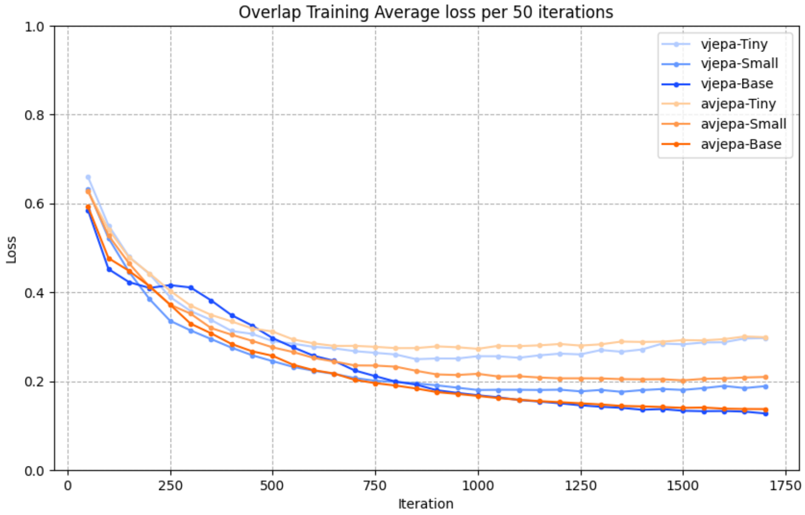
(2) Tokenization and Positional Embedding: The audiospectrogram is tokenized using 16x16 convolutions (similar to video input) and 2D sinusoidal positional embeddings are added to each token.

(3) Masking Strategy: We chose to adopt a "static" masking strategy for audio data. For a predetermined number of masks, we generate a random top left corner and mask a fixed shaped region of the audio spectrogram. Just as for video input, this masking is performed at the *token* level.

5.6 Evaluation

During evaluation, the pretrained encoder remains frozen while only the decoder is finetuned for specific tasks. For classification, this decoder is a simple classification head. For prediction tasks, the decoder transforms the encoder's learned latent representations back into pixel-level video frames. This paper primarily evaluates the training performance and downstream task performance of VJEPA and AVJEPA. We evaluate on 3 sizes of encoder models on a subsection of the Kinetics-400 dataset.

Fig. 3. JEPA Models Training



5.6.1 Self-Supervised PreTraining. Training was conducted using NVIDIA 4090 and NVIDIA 3070 graphics cards. These models were all trained on the same training set, using a parameter of 2 unique masks per context iteration for both the video and audio tokens (4 masks total). As in Figure 3, there exists only minor differences in modality during training cycle. Given that we only train 1 final version of each model, our current data cannot be used to come to any conclusion regarding degree of difficulty of training given a deeper modality. However, generally, we can observe that the larger models achieve a better performance (expected) even to a point where given our about 1750 iteration limit, potentially we could have seen further convergence on the Vit-Base variations if continued training.

5.6.2 Video Classification. We evaluated three vjepa models and three models mentioned above on video classification tasks. Each dataset - both the fine-tuning set and the validation set - contains 3,500 videos. Due to computational resource constraints, we limited our testing to 6,020 iterations across all evaluations.The evaluation result is shown in figure 4. For both VJEPA and AVJEPA models, accuracy improves as model size increases.

While we anticipated AVJEPA to outperform VJEPA, given enrichment through audio data, the opposite is demonstrated in our results. We suspect that the additional complexity of multimodal representation space, exacerbated by limited training compute/amount, bottlenecked AVJEPA more VJEPA, and thus it's performance on downstream tasks demonstrated.

5.6.3 Mask Prediction. We qualitatively evaluated a single avjepa model variation, Vit-Small, on the task of Mask Prediction, that is, we mask a portion of both the input video and audiospectrogram and try to predict the missing pixels. The evaluation on a single model is purely due to time and compute constraints. Based on previous results (previous iterations of AVJPEA) we expected extreme difficulty in this task and pursued this object as a kind of "proof of concept". To this end, we trained an attention probe on top of the outputs of the *encoder* and *predictor*, mapping the combined masked (predicted) and unmasked tokens to pixel space. We should point out a few critical points about this approach: (1) Firstly, the attention probe we used was the "best we could do" given the compute we have and time-frame, there remains to be a huge amount of experimentation and editing available to optimize structure and performance of the attentive probe. (2) The training process had to be conducted stochastically, training on only a single input and 2 masks at a time. This is due to the immense engineering that would be required to accommodate higher batch counts (we learned this during avjepa training implementation) as well as computational constraints.

Figure 5 describes our observed training process and Figure 7 the pipeline for the downstream attention probe training. We observed once again, a double descent phenomena, and generally high variance during training. This can be accounted for by the stochastic training however we suspect that either our attention probe or our baseline encoder/predictor structure is creating some level of noise/variance that is heavily impacting training and performance. In Figure 6, we provide an example prediction, including the *masked* input and the resulting reconstructed/predicted output. Generally the example provided is representative of pipeline performance across the dataset. The resulting reconstruction is always blurry, lacking details however captures generally the color shading of the image and critically generally correctly predicting the shading of the masked region correctly. This, compared to our previous results is a massive improvement. We also verified that a constant output was not being created. We noticed that especially for audio, our pipeline was performing remarkably well, predicting masked areas very well. We attribute this to the more simple nature and dimensionality (B/W) of the audiospectrogram input.

Our results in mask prediction are intriguing, With essential minimal model capacity, minimal compute and data for training, we were able to reconstruct multi-modal outputs at a low-level of fidelity. Especially in audio, our reconstruction was more effective than we envisioned. We would expect that through scaling of training data quantity and diversity, encoder/predictor size, implementing a more all encompassing and dynamic masking strategy for both training and probe fine-tuning, that we can achieve a far higher level of detail in the reconstructions.

6 Future Work

(1) Scaling: A critical missing piece in this paper is the scale of our operations. Our encoder and decoder models are operating at a fraction of size of SOTA approaches. Further, the quantity and diversity of training/fine-tuning data is also a fraction of previous works. We suspect that given the increased complexity of a multimodal embedding space, scaling model capacity and data quality/size would thus proportional result in respective increases in model performance.

(2) Masking: In both training of JEPA variations, we adopted a relatively simple approach to masking. Masking is however, a essential component of model exposure to information, we expect that a more controlled masking approach, such as increasing or decreasing masking amount throughout the training/fine-tuning process can result in faster and better learning.

(3) Multi-modality: Here we investigate a dual-modal embedding space. An obvious future direction is to incorporate more modalities and exploring the difficulties and complexities associated with each.

7 Conclusion

This paper explores the AV-JEPA framework for learning multimodal representations through self-supervised feature prediction. By extending the JEPA approach to include both visual and audio modalities, we demonstrated the multimodal integration under constrained computating resources. Despite challenges in engineering and evaluation, our results highlight the promise of compact, predictive architectures for multimodal tasks such as video classification and mask prediction (both pixel space and audio spectrogram). Future work can try to scale to larger datasets and larger models. Also, refining masking strategy during training and downstream evaluation methods can further extend these findings.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15619–15629.

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39th International Conference on Machine Learning*. 1298–1312. https://proceedings.mlr.press/v162/baevski22a.html

[3] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471* (2024).

[4] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Converence on Learning Representations*.

[5] Yann LeCun. [n. d.]. A Path Towards Autonomous Machine Intelligence Version. ([n. d.]).

[6] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Jiao Qiao. 2022. UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning. *ArXiv* abs/2201.04676 (2022). https://api.semanticscholar.org/CorpusID:245906266

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[8] R Rao. 1999. Predictive coding in the visual cortex. *Nature Neuroscience* 2, 1 (1999), 9–10.

[9] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Z1Qlm11uOM

[10] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised Learning of Video Representations using LSTMs. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 843–852. https://proceedings.mlr.press/v37/srivastava15.html

[11] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2015. Anticipating Visual Representations from Unlabeled Video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 98–106. https://api.semanticscholar.org/CorpusID:10533233

Fig. 4. Video Classification Evaluation

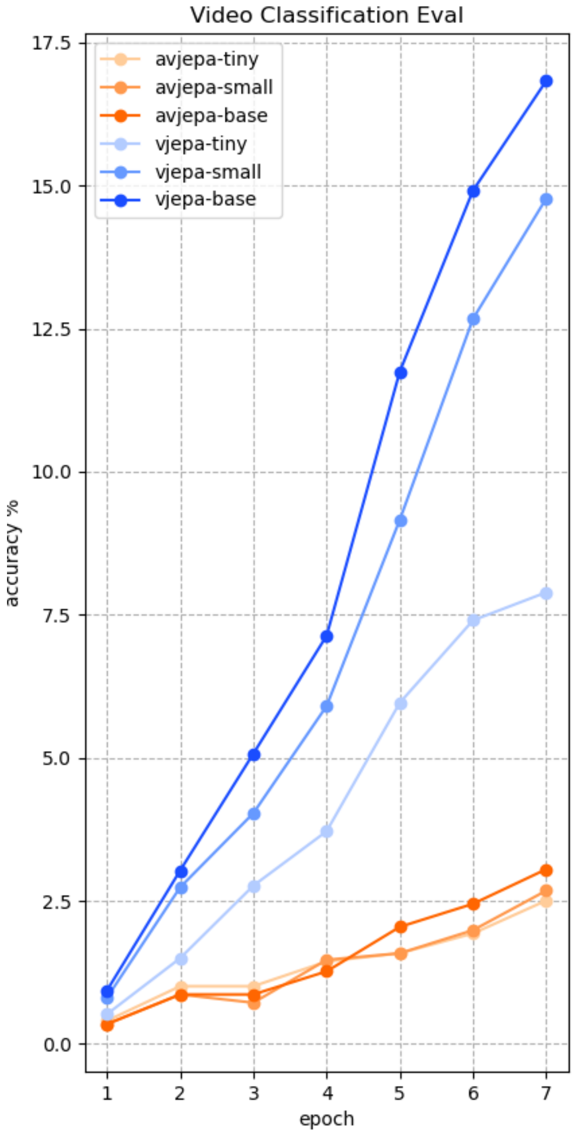


Fig. 5. Mask Prediction Probe Training

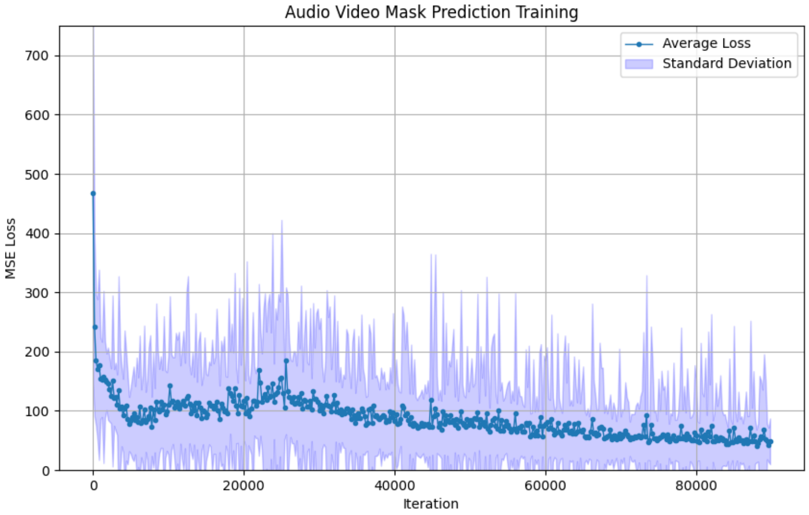


Fig. 6. Mask Prediction Example

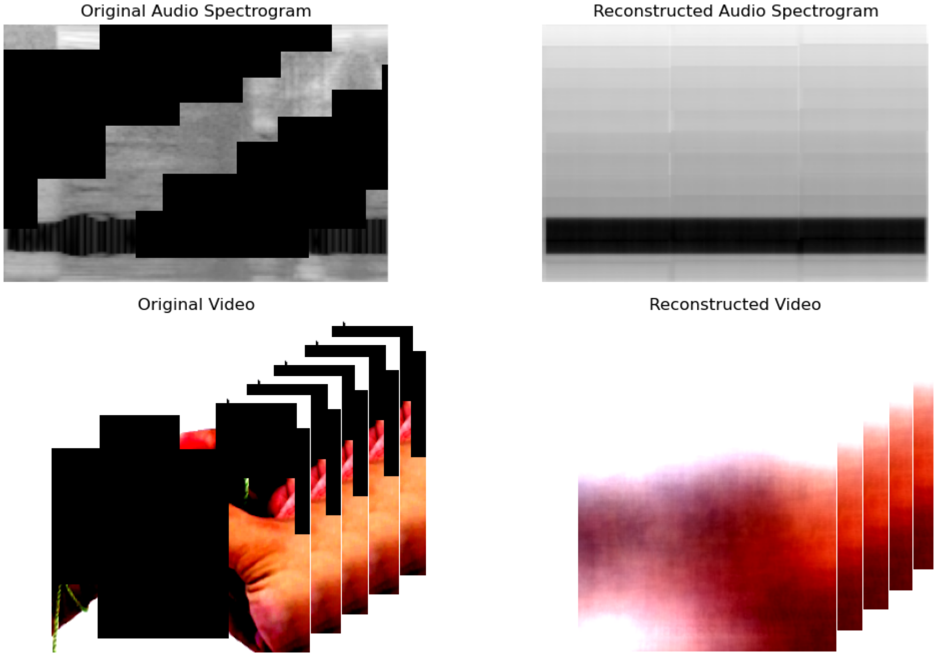
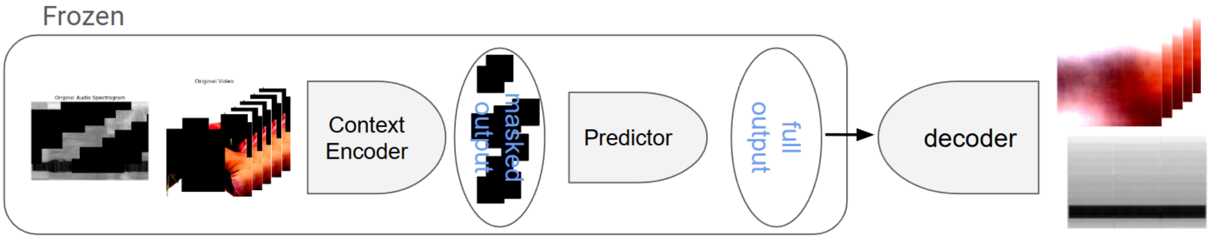


Fig. 7. Mask Prediction Pipeline


Frozen
Original Audio Spectrogram
Original Video
Context Encoder
masked output
Predictor
full output
decoder