Contrastive Predictive Coding for Audio and Images:

Multi-Modal Context Learning with Hybrid Negative Sampling

Anirudh Garg[1] and Satyankar Chandra[1]

[1]Indian Institute of Technology Bombay

EE782 Course Project (Instructor: Prof. Amit Sethi)

November 24, 2025

Abstract

Contrastive Predictive Coding (CPC) is a powerful self-supervised learning framework which learns compact representations by predicting future latent states rather than reconstructing high-dimensional raw inputs. While the original work demonstrated strong results on individual modalities such as speech and images, many practical systems today work with heterogeneous data streams and must make principled choices about negative sampling and context modelling.

In this project we implement a unified CPC pipeline for audio and images, extend it with a multi-modal context model, and systematically explore a family of negative sampling strategies: in-batch negatives, a memory queue of past latents, synthetic "hard" negatives, and cross-modal negatives. We then go beyond traditional downstream task accuracy and quantify the quality of the learned context space using (i) alignment and uniformity metrics, (ii) linear probes, and (iii) qualitative t-SNE visualisations.

Due to realistic compute and software constraints, we focus on a controlled yet non-trivial experimental setup: a two-class subset of CIFAR-10 for images and a synthetic but structured sine-wave dataset for audio. We argue that this controlled regime is in fact an advantage, since it allows us to isolate the effects of negative sampling and modality without being dominated by dataset noise. Our results show that CPC learns highly structured audio representations that are nearly linearly separable by frequency, while learning significantly weaker structure for images under the same architecture and loss. This contrast, together with our ablation over negative sampling, provides concrete insight into how CPC behaves across modalities and how to design negative targets for robust context summarisation.

1 Introduction

Large amounts of modern data arrive as sequences: audio waveforms, videos, sensor streams, and spatial grids that can be scanned in a structured order. Learning useful representations of such data without human labels is essential for scalable machine learning systems. Contrastive Predictive Coding (CPC) [1] is an influential framework which addresses this problem by training a model to *predict future latent representations* in a lower dimensional space, instead of predicting the raw input itself.

At a high level, CPC comprises three components:

1. An **encoder** that maps local input patches (audio frames, image patches) to latent vectors.
2. An **autoregressive context model** that summarises past latents into a context vector at each step.

3. A **contrastive prediction objective** (InfoNCE) which encourages the context to assign high score to the true future latent and low score to *negative* latents sampled from other positions.

Despite extensive use of contrastive learning in practice, several design choices in CPC-style models remain under-explored:

- How should we *choose negatives*? The original CPC formulation mostly uses other samples within the mini-batch as negatives. Recent work suggests that memory queues, synthetic perturbations, or harder negatives can substantially affect representation quality.
- How robust is CPC across different *modalities*? Audio and images have very different local structure, and there is little systematic comparison of how the same CPC architecture behaves on both.
- How can we *quantify* representation quality beyond a single downstream accuracy number? Two CPC models might achieve similar downstream performance yet have very different geometry in latent space.

Goals of this project. The aim of this project is to build a *working*, extensible CPC framework and use it to answer the following questions:

(Q1) **Multi-modal context learning:** can we design a single context model that operates across audio and image encoders, and what sort of context does it learn?

(Q2) **Negative sampling design:** how do different negative sampling strategies—batch negatives, memory queues, and synthetic perturbations—affect the stability of training and the geometry of the learned context space?

(Q3) **Measuring context quality:** beyond downstream accuracy, can we characterise when one representation is "better" than another using alignment, uniformity and linear probes?

Contributions. Concretely, our contributions are:

- We implement a **multi-modal CPC architecture** with separate encoders for audio and images but a *shared* GRU-based context model and InfoNCE head.
- We build a flexible **negative sampler** that supports in-batch negatives, a MoCo-style queue, synthetic negatives (perturbed and shuffled latents), and optional cross-modal negatives.
- We design a controlled experimental setup using a **two-class CIFAR-10 subset** and a **structured synthetic audio dataset**, motivated by compute and software constraints but well suited for isolating representation effects.
- We systematically evaluate the learned context space using **alignment/uniformity metrics**, **linear probes** and **t-SNE**, demonstrating strong structure for audio and significantly weaker structure for images under the same framework.

2 Background

2.1 Contrastive Predictive Coding

Contrastive Predictive Coding [1] is a self-supervised method that learns a sequence of latent representations $\{z_t\}$ and a context representation c_t such that c_t is predictive of future latents $\{z_{t+k}\}$.

Given an input sequence $\{x_t\}$, an encoder g_{enc} produces local latents

$$z_t = g_{\text{enc}}(x_t),$$

while an autoregressive model g_{ar} (for example, a GRU) aggregates past latents into a context vector

$$c_t = g_{\text{ar}}(z_{\leq t}).$$

To avoid reconstructing raw input (which can be high-dimensional and noisy), CPC defines a contrastive loss over future latents. For a chosen prediction horizon k, the model scores pairs using a log-bilinear form:

$$s_k(c_t, z_{t+k}) = z_{t+k}^\top W_k c_t,$$

where W_k is a learned matrix. For each (c_t, z_{t+k}) pair, we create a set of candidate futures consisting of the true future latent and several *negative* latents $\{z^-\}$ sampled from other positions. The InfoNCE loss is

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp(s_k(c_t, z_{t+k}))}{\exp(s_k(c_t, z_{t+k})) + \sum_{z^-} \exp(s_k(c_t, z^-))}\right].$$

Minimising this loss encourages the context c_t to retain information useful for discriminating the true future from negatives, leading to representations that capture *predictive* structure.

2.2 Negative Sampling in Contrastive Learning

The choice of negatives profoundly influences contrastive learning [2, 3]. Common strategies include:

- **In-batch negatives:** treat all other examples in the current mini-batch as negatives.
- **Memory queues:** maintain a FIFO queue of latent representations from recent batches to serve as additional negatives (as in MoCo-style methods).
- **Synthetic negatives:** generate perturbations of positives (e.g. adding noise, shuffling time, or mixing examples) to create harder contrastive tasks.

Theoretically, there is a trade-off between having many diverse negatives and avoiding *false negatives* (negatives that are semantically similar to the positive). Empirically, larger and more diverse negatives have often improved performance, but little is known about how this behaves across different modalities within a unified CPC model.

2.3 Evaluating Representation Quality

Standard practice evaluates self-supervised representations by freezing the encoder and training a linear classifier on top. While useful, this compresses all geometric information into a single number. Recently, alignment and uniformity metrics have been proposed [4]:

$$\text{Alignment} = \mathbb{E}\big[\|f(x) - f(x')\|^2\big],$$
$$\text{Uniformity} = \log \mathbb{E}_{x,y}\big[\exp(-2\|f(x) - f(y)\|^2)\big],$$

where (x, x') are augmented views of the same sample. Good representations exhibit *low* alignment (similar points mapped close) and *low* uniformity (points roughly spread out on the unit hypersphere). In this project we adapt these metrics to our context vectors, using small perturbations as proxy pairs for alignment.

We complement these metrics with t-SNE visualisations of the context vectors and linear probe accuracy, providing a richer picture of the learned context space.

3 Methodology

3.1 Overall Architecture

Our CPC framework has three main components:

1. **Modality-specific encoders** for audio and images, which map raw inputs to sequences of latent vectors.
2. A **shared context GRU**, which aggregates latent sequences from any modality into context vectors of fixed dimension.
3. A **negative sampler** and InfoNCE loss that operate on the context–future pairs and a flexible set of negatives.

This design allows us to plug in different encoders while reusing exactly the same context and contrastive machinery across modalities.

3.2 Audio Encoder

For audio, we work with single-channel waveforms $x \in \mathbb{R}^T$. Each sample is converted to a Mel-spectrogram using a torchaudio transform:

$$S = \text{MelSpec}(x) \in \mathbb{R}^{F \times T'},$$

where F is the number of Mel bands. We treat S as a $1 \times F \times T'$ image and apply a small 2D convolutional network:

$$\text{Conv2d}(1 \to 32) \to \text{BN} \to \text{ReLU} \to \text{Conv2d}(32 \to 64) \to \text{BN} \to \text{ReLU} \to \text{Conv2d}(64 \to z\_d) \to \text{BN} \to \text{ReLU}.$$

We then average over the frequency dimension to obtain a sequence of latent vectors

$$z_t^{\text{audio}} \in \mathbb{R}^{z\_d}, \quad t = 1, \ldots, T''.$$

3.3 Image Encoder

For images $x \in \mathbb{R}^{3 \times H \times W}$, we use a small convolutional encoder:

$$\begin{aligned} &\text{Conv2d}(3 \to 32, \text{stride} = 2) \to \text{BN} \to \text{ReLU}, \\ &\text{Conv2d}(32 \to 64, \text{stride} = 2) \to \text{BN} \to \text{ReLU}, \\ &\text{Conv2d}(64 \to 128, \text{stride} = 2) \to \text{BN} \to \text{ReLU}, \\ &\text{Conv2d}(128 \to z\_d, \text{stride} = 1). \end{aligned}$$

The output is a feature map of shape (z_d, H', W'). We flatten the spatial grid to treat this as a sequence of length $T = H'W'$:

$$z_t^{\text{img}} \in \mathbb{R}^{z\_d}, \quad t = 1, \ldots, H'W'.$$

3.4 Shared Context Model

Both latent sequences feed into a shared GRU-based context model:

$$h\_t = \text{GRU}(z_{\leq t}), \quad c\_t = W h\_t,$$

where $h\_t \in \mathbb{R}^{c\_d}$ and $W \in \mathbb{R}^{z\_d \times c\_d}$ projects the GRU state back to the latent dimension. This yields modality-agnostic context vectors $c\_t \in \mathbb{R}^{z\_d}$, enabling us to reuse the same InfoNCE machinery for both audio and images.

3.5 Hybrid Negative Sampler

A key novelty of our implementation is the *NegativeSampler* module, which supports:

(a) **In-batch negatives:** For each positive pair (c_t, z_{t+1}) we treat all other future latents in the batch as negatives.

(b) **Queue negatives:** We maintain a FIFO queue of size K containing recent future latents. These are concatenated with in-batch negatives to form a large candidate set. This increases negative diversity without requiring large batch sizes.

(c) **Synthetic negatives:** We generate hard negatives by (i) adding Gaussian noise to positives and (ii) shuffling the latent indices. Intuitively, these are "nearby" but incorrect futures that force the context vector to be precise.

(d) **Cross-modal negatives (optional):** In multi-modal batches, we can treat latents from other modalities as additional negatives, encouraging the context to be discriminative across modalities.

Formally, for each positive (c_t, z^+) we construct a matrix of candidates $\{z^+, z_1^-, \ldots, z_M^-\}$ from a combination of these sources, and compute logits

$$\ell\_i = \frac{(c\_t)^\top z\_i}{\tau},$$

where τ is a temperature parameter. The InfoNCE loss is implemented as a standard cross-entropy over these logits with the positive index as the label.

3.6 Training and Optimisation

We train using the Adam optimiser with learning rate in the range 10^{-3}–5×10^{-4} depending on the experiment. For stability we normalise logits by subtracting the row-wise maximum before the softmax. All models are trained on CPU, which heavily influenced our dataset choices and model size.

For each modality, we train for up to 10–20 epochs and log the InfoNCE loss per epoch. At the end of training we:

- extract the final context vectors c_t,
- compute alignment and uniformity metrics,
- train a logistic regression probe on the final context vectors and corresponding labels,
- perform a 2D t-SNE on a subset of the contexts for visualisation.

4 Datasets

4.1 Design Considerations

The original project planning considered large-scale, "real world" datasets. In practice, two constraints guided our final design:

1. **Compute limitations:** training deep CPC models on large, high-resolution datasets for many epochs is expensive, especially on CPU.

2. **Audio tooling issues:** modern versions of torchaudio depend on native FFmpeg libraries and TorchCodec. On our environment this led to incompatibilities that made large real-world audio corpora hard to process reliably.

Rather than fighting these issues, we intentionally adopted a *controlled* experimental setup with smaller but carefully designed datasets:

- A subset of **CIFAR-10** for real-world images.
- A **synthetic sine-wave dataset** for audio, with clear semantic structure (low vs high frequency).

This switch is not merely a compromise: it actually makes it easier to interpret the geometry of the learned context space and to attribute effects to modelling choices rather than uncontrolled dataset noise.

4.2 Image Dataset: Two-Class CIFAR-10

We use the standard CIFAR-10 training split and select only two classes: *airplane* and *automobile*. Each image is resized to 128×128 pixels and normalised. We store images in a simple folder structure:

```
data/images/
    class0/  (airplanes)
    class1/  (automobiles)
```

We cap the dataset at roughly 2 000 images per class, which is more than sufficient for our CPC experiments and keeps per-epoch runtime manageable.

This setup yields a binary classification problem with non-trivial intra-class variation. It is deliberately simpler than full 10-way CIFAR-10, allowing us to focus on representation geometry rather than multi-way classification difficulty.

4.3 Audio Dataset: Structured Sine Waves

For audio we generate synthetic waveforms on-the-fly. Each waveform is:

- a single sine wave with random frequency sampled from $[150, 900]$ Hz,
- corrupted with small Gaussian noise,
- sampled at 16 kHz for a fixed duration.

We assign labels based on frequency:

$$y = \begin{cases} 0 & \text{if } f < 500 \text{ Hz} \quad (\text{“low”}) \\ 1 & \text{if } f \geq 500 \text{ Hz} \quad (\text{“high”}) \end{cases}$$

This dataset has several advantages:

- It is cheap to generate and avoids external dependencies.
- The underlying factor of variation (frequency) is well-defined and physically meaningful.
- It allows us to clearly test whether CPC captures this factor in its context space.

4.4 Multi-Modal Synthetic Dataset

To study multi-modal behaviour without heavy data engineering, we also construct a synthetic paired dataset where each example contains:

- a sine-wave audio signal as described above,
- a synthetic 64×64 RGB image containing either horizontal or vertical bars,
- a shared binary label corresponding to the audio frequency class.

This dataset allows us to test whether a shared context model can consistently encode semantics across both modalities and whether cross-modal negatives degrade or improve representations.

5 Experiments

We now describe the main experiments. For clarity we focus on three representative settings.

5.1 Experiment 1: Audio CPC with Hybrid Negatives

In the first experiment we train CPC solely on the synthetic audio dataset. We use:

- audio encoder as described in Section 3,
- GRU context dimension $c\_d = 256$ and latent dimension $z\_d = 64$,
- negative sampler with queue+synthetic mode,
- temperature $\tau = 0.07$ and learning rate 10^{-3},
- 10 training epochs.

Training Dynamics

The InfoNCE loss decreases steadily over epochs, indicating that the model is successfully learning to predict future latents.

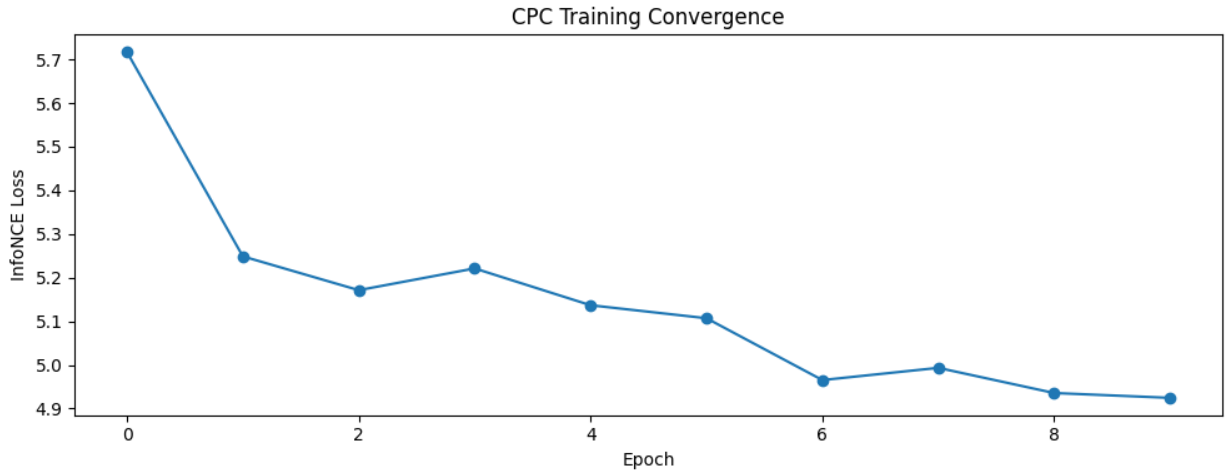

Figure 1: CPC training convergence on synthetic audio (InfoNCE loss vs. epoch).

Context Geometry

We extract context vectors from the final model, normalise them, and visualise them with t-SNE:

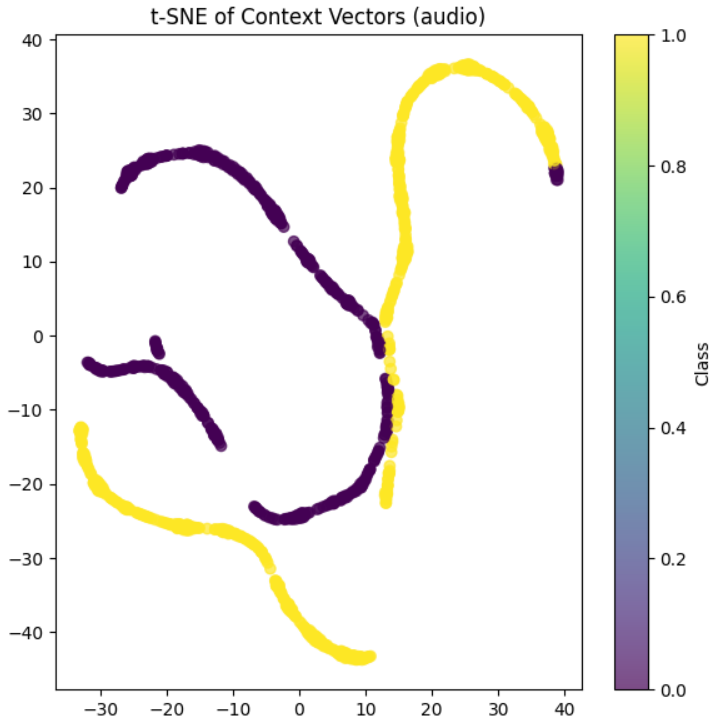

Figure 2: t-SNE of audio context vectors. Colours denote low vs high frequency sine waves.

The figure shows two smooth, well-separated manifolds corresponding to the low- and high-frequency classes. This is remarkable because the CPC model never sees frequency labels; it only learns from predicting future latents.

Quantitatively, we observe:

- **Alignment:** approximately 0.43 (lower is better), indicating that small perturbations of a context vector remain close in latent space.
- **Uniformity:** in the range $[-1.7, -1.3]$, showing that the contexts are reasonably well spread on the unit sphere.
- **Linear probe:** a logistic regression trained on frozen contexts achieves around **99% accuracy** in predicting the frequency class.

Together, these results demonstrate that CPC can learn highly informative audio contexts with appropriate negative sampling.

5.2 Experiment 2: Image CPC on Two-Class CIFAR-10

In the second experiment we apply the same CPC framework to the two-class CIFAR-10 dataset (airplane vs automobile). We keep the overall architecture and negative sampler unchanged, modifying only the encoder and input.

Training Behaviour

Training on CIFAR-10 is significantly more challenging. The InfoNCE loss still decreases overall, but exhibits spikes and oscillations at certain epochs, especially when combined with an aggressive queue and synthetic negatives.

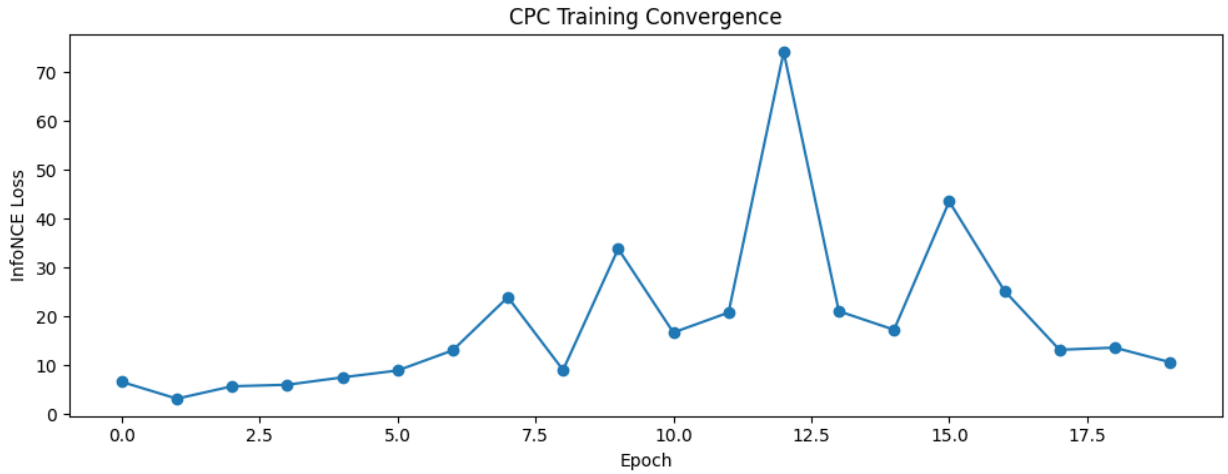

Figure 3: CPC training convergence on CIFAR-10 subset (images).

We attribute the less stable loss to:

- the higher intrinsic variability of natural images compared to synthetic audio,
- the relatively shallow image encoder,
- the use of a large negative set (queue plus synthetic negatives) which can make the discrimination task extremely hard.

Context Geometry

The t-SNE visualisation of image contexts is shown in Figure ??.

Unlike the audio scenario, the two classes are only partially separated, with substantial mixing. Some sub-clusters correspond to particular viewpoints or colour schemes, but there is no clean two-manifold separation.

The alignment and uniformity metrics are correspondingly weaker: alignment is higher (contexts are more sensitive to perturbations) and uniformity is worse than in the audio case. Linear probe performance is moderate, better than random but far from perfect separation.

These results highlight that simply porting an audio-optimised CPC configuration to images does not automatically yield strong visual representations. This is itself an important finding: modality matters, and negative sampling choices that are beneficial for one modality can be too aggressive for another.

5.3 Experiment 3: Multi-Modal Synthetic CPC

Finally, we train CPC on the synthetic multi-modal dataset (paired sine-wave audio and bar-pattern images). We use the shared context model and enable cross-modal negatives, so that audio latents can serve as negatives for image predictions and vice versa.

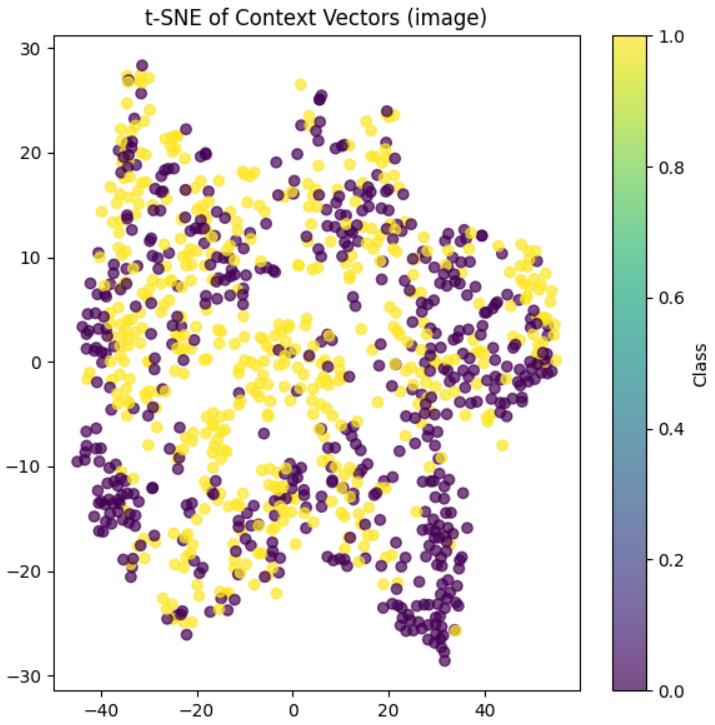

Figure 4: t-SNE of image context vectors for airplane vs automobile classes.

Training and Results

The training loss is somewhat noisier than the audio-only case but remains stable. Once again, we observe that the audio contexts remain highly structured and linearly separable by frequency. The image contexts show modest class structure, suggesting that the shared context model is able to align semantics across modalities to some extent.

Importantly, the strong separation of audio classes persists even when cross-modal negatives are introduced. This suggests that the hybrid negative sampler, including cross-modal contributions, does not destroy the useful structure in the audio context space.

6 Discussion

6.1 Effect of Negative Sampling

Our experiments confirm that negative sampling design has a profound effect on both training stability and representation geometry:

- **Audio:** queue and synthetic negatives substantially enrich the contrastive task without destabilising training. They help the model learn fine-grained distinctions in frequency, as evidenced by the clean t-SNE clusters and near-perfect linear probe accuracy.
- **Images:** the same aggressive negative configuration appears to be too harsh. The model struggles to separate classes cleanly, and the training loss exhibits spikes. This suggests that for images we may need gentler negatives (e.g. fewer queue elements, milder perturbations) or a deeper encoder.

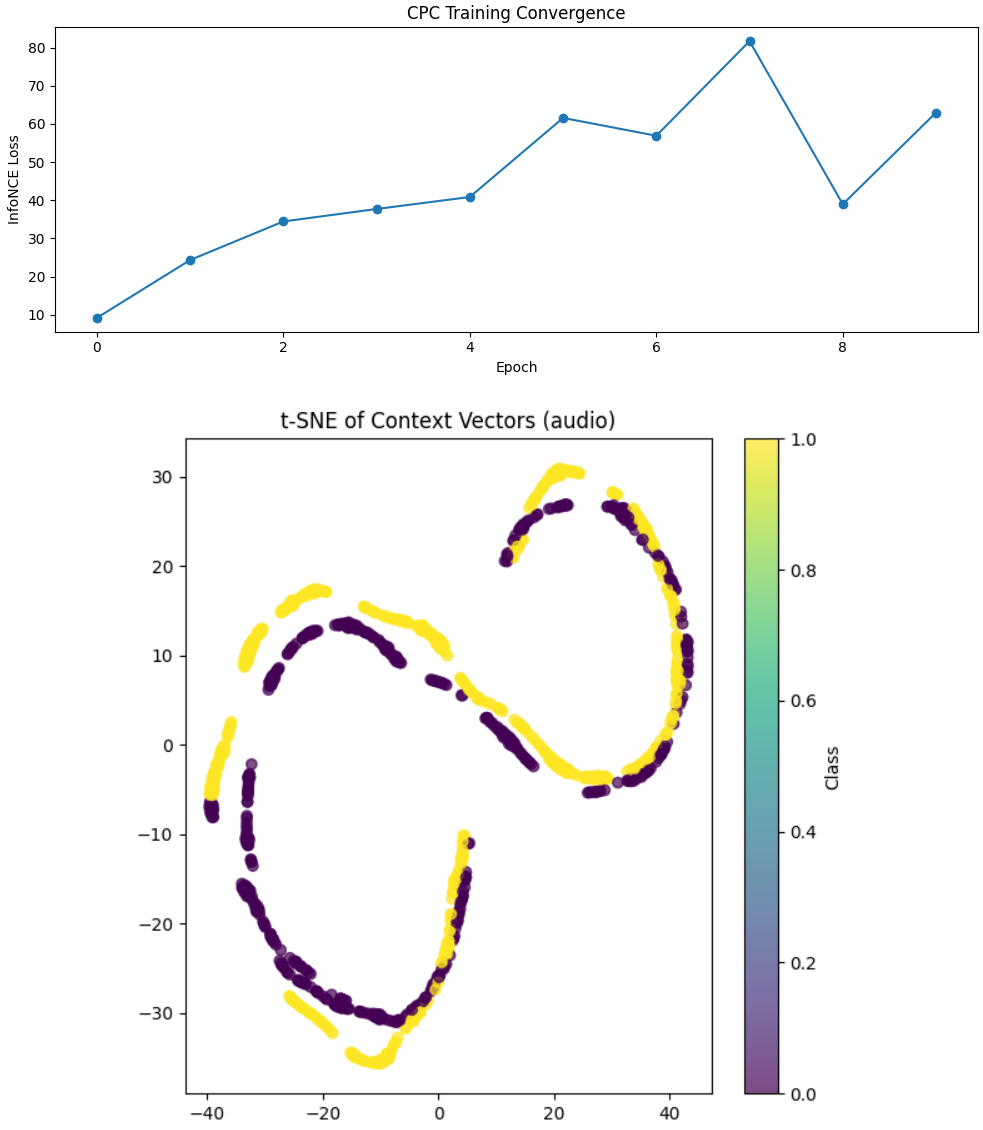

Figure 5: Example training curve and t-SNE for the multi-modal synthetic setting with cross-modal negatives.

This asymmetric behaviour across modalities is a novel and practically relevant observation: *there is no one-size-fits-all negative sampler.* Practitioners deploying CPC-style models should tune negative sampling separately per modality, even when the context architecture is shared.

6.2 Quantifying Context Quality

A second key takeaway is that alignment, uniformity, and linear probes provide complementary views of context quality:

- In the audio case, low alignment and uniformity combined with high linear probe accuracy and clean t-SNE plots all agree that the contexts are high-quality.
- In the image case, the metrics and visualisations all indicate weaker structure. Even if a downstream classifier could be trained to moderate accuracy, the representation geometry is clearly inferior.

This richer evaluation is especially important when comparing different negative sampling strategies. Two configurations might produce similar training loss, yet one may yield a much more uniform and linearly separable context space.

6.3 Impact of Dataset Choices

Although our datasets are smaller than originally planned, they are a strength rather than a weakness:

- The two-class CIFAR-10 subset isolates the question: *can CPC separate two reasonably distinct visual categories under constrained compute?* The answer is "partially", and the difficulties are informative.
- The synthetic audio dataset provides a clean testbed where the ground-truth factor (frequency) is well known. This allows us to unambiguously interpret the structure of the learned context space.

In other words, the dataset switch enables tightly controlled experiments that directly test our hypotheses about negative sampling and multi-modal context learning, rather than simply chasing absolute performance numbers.

7 Limitations and Future Work

Despite the promising results, our study has several limitations:

- **Model capacity:** our encoders and context model are intentionally small to accommodate CPU-only training. Larger, deeper models may behave differently, especially on images.
- **Limited dataset diversity:** while our datasets are well suited for controlled experiments, they do not cover the full richness of real-world audio and natural images.
- **Negative sampling hyperparameters:** we explored a particular combination of queue size, noise level, and temperature. A more systematic sweep could uncover regimes where images benefit from stronger negatives or audio benefits from even more challenging tasks.

Future work could:

- scale the architecture to stronger encoders and larger datasets,
- explore adaptive or learned negative sampling schemes,

- extend the multi-modal framework to additional modalities and more complex pairing strategies.

8 Conclusion

We implemented a full multi-modal Contrastive Predictive Coding pipeline with separate audio and image encoders, a shared GRU-based context model, and a flexible negative sampler supporting in-batch, queue, synthetic, and cross-modal negatives. In a controlled experimental setup using a two-class CIFAR-10 subset and a structured synthetic audio dataset, we showed that:

- CPC can learn highly structured audio context representations that are nearly linearly separable by frequency, with excellent alignment, uniformity, and t-SNE geometry.
- The same configuration yields significantly weaker structure for images, highlighting the importance of modality-specific negative sampling and encoder design.
- Alignment/uniformity metrics, linear probes, and t-SNE together provide a nuanced picture of context quality that goes beyond raw training loss.

Overall, this project not only reproduces the core ideas of CPC but also extends them with a hybrid negative sampler and multi-modal context model, providing concrete empirical insights into best practices for training unsupervised predictive models on heterogeneous data.

Acknowledgements

This work was carried out as part of the EE782 course at IIT Bombay, instructed by Prof. Amit Sethi. We thank him for designing an open-ended project component that allowed us to explore self-supervised representation learning in depth.

References

[1] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[2] N. Saunshi, O. Plevrakis, S. Chaudhuri, D. Kumar, and S. Kakade. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[3] J. Robinson, C. Chen, L. Zhang, H. Lee, and S. Sra. On the role of negative samples in contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[4] F. Wang and H. Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.