Contrastive Predictive Coding for Audio and Images: Multi-Modal Context Learning with Hybrid Negative Sampling

Anirudh Garg, Satyankar Chandra
Department of Computer Science and Engineering, IIT Bombay
EE782: Advanced Machine Learning (Instructor: Prof. Amit Sethi)

***Abstract*—Contrastive Predictive Coding (CPC) is a self-supervised framework that learns compact representations by predicting future latent states instead of reconstructing high-dimensional inputs. While the original work demonstrated strong results on individual modalities such as speech and images, modern systems often operate on heterogeneous data streams and must make careful 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 evaluate not only downstream linear probes but also alignment/uniformity metrics and t-SNE visualizations of the context space.

Due to realistic compute and tooling constraints, we adopt a controlled experimental setup: a two-class subset of CIFAR-10 for images and a synthetic but structured sine-wave dataset for audio. This regime turns out to be an advantage, as it lets us isolate how CPC behaves across modalities and negative sampling schemes. After modest tuning of the encoder depth and negative-sampling hyperparameters, our results show that CPC learns highly structured audio representations that are almost linearly separable by frequency, and competitive image representations with clearly separated semantic clusters. In a synthetic multi-modal setting, cross-modal negatives further encourage a shared context space in which acoustically and visually related samples lie near each other. These findings provide concrete guidance on designing negative samples and architectures for CPC-style representation learning in multi-modal settings.

***Index Terms*—Contrastive Predictive Coding, Self-Supervised Learning, Representation Learning, Multi-Modal Learning, Negative Sampling**

I. Introduction

Large amounts of modern data arrive as sequences: audio, videos, sensor streams and spatial grids that can be scanned in a structured order. Learning useful representations of such data without labels is critical for scalable machine learning systems. Contrastive Predictive Coding (CPC) [1] addresses this by training a model to *predict future latent representations* in a lower-dimensional space, rather than reconstructing raw inputs.

CPC consists of three components:

1) A **local encoder** that maps input patches (audio frames, image patches) to latent vectors z_t.
2) An **autoregressive context model** that summarizes past latents into a context c_t.
3) A **contrastive prediction objective** (InfoNCE) that encourages c_t to assign higher score to the true future latent z_{t+k} than to negatives.

Despite the popularity of contrastive learning, several aspects of CPC remain under-explored:

- The choice of *negative samples* (in-batch vs. queues vs. synthetic perturbations) has a strong effect on representation quality [2], [3].
- The same architecture and loss might behave very differently across *modalities* (audio vs. images).
- Evaluation is often reduced to a single downstream accuracy number, ignoring the geometry of the learned context space.

A. Goals and Contributions

Our project aims to build a working, extensible CPC framework and use it to answer three questions:

(Q1) Can a *shared* context model operate effectively across both audio and image encoders?

(Q2) How do different negative sampling strategies (batch, queue, synthetic, cross-modal) affect training stability and latent geometry?

(Q3) How can we quantify when one context representation is "better" than another beyond raw accuracy?

Our main contributions are:

- A **multi-modal CPC architecture** with separate encoders for audio and images and a shared GRU-based context model.
- A **hybrid negative sampler** supporting in-batch, queue-based, synthetic, and cross-modal negatives.
- A **controlled experimental setup** using a two-class CIFAR-10 subset and a structured synthetic audio dataset, tuned to CPU-only training.

- A **representation analysis** using alignment/uniformity [4], linear probes and t-SNE plots that clearly shows how modality and negative sampling interact.

II. Background

A. Contrastive Predictive Coding

Given a sequence $\{x_t\}$, CPC uses an encoder g_{enc} to produce local latents

$$z_t = g_{\text{enc}}(x_t), \tag{1}$$

and an autoregressive model g_{ar} to summarize past latents into context

$$c_t = g_{\text{ar}}(z_{\leq t}). \tag{2}$$

A prediction head scores pairs via

$$s_k(c_t, z_{t+k}) = z_{t+k}^{\top} W_k c_t, \tag{3}$$

for horizons $k = 1, \ldots, K$.

For each context c_t, the model considers the true future latent z_{t+k} and a set of negatives $\{z^-\}$. The InfoNCE loss is

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp(s_k(c_t, z_{t+k}))}{\exp(s_k(c_t, z_{t+k})) + \sum_{z^-} \exp(s_k(c_t, z^-))}\right]. \tag{4}$$

Minimizing this encourages c_t to retain precisely the information needed to identify the true future latent among a set of candidates.

B. Negative Sampling

The choice of negatives is crucial [2], [3]:

- **In-batch negatives**: treat all other samples in the mini-batch as negatives.
- **Memory queues**: maintain a FIFO queue of latent vectors from previous batches.
- **Synthetic negatives**: generate perturbed or shuffled versions of latents to form "hard" negatives.

There is a trade-off: many negatives increase discrimination power but also risk introducing false negatives.

C. Measuring Representation Quality

Beyond downstream accuracy, the alignment and uniformity metrics of [4] capture useful geometric properties:

$$\text{Alignment} = \mathbb{E}\big[\|f(x) - f(x')\|^2\big], \tag{5}$$

$$\text{Uniformity} = \log \mathbb{E}_{x,y}\big[\exp(-2\|f(x) - f(y)\|^2)\big], \tag{6}$$

where (x, x') are two views of the same sample. Desirable representations have low alignment and low uniformity.

III. Methodology

A. Architecture Overview

Our framework has three main components:

1) **Audio encoder** $g_{\text{enc}}^{\text{aud}}$ that maps waveforms to latent sequences.
2) **Image encoder** $g_{\text{enc}}^{\text{img}}$ that maps images to latent sequences.
3) **Shared context GRU** g_{ar} that maps any latent sequence to context vectors, followed by a projection to the latent dimension.

Given a batch that may contain audio and/or images, we:

1) Encode each modality into $z_{1:T}^{\text{mod}}$.
2) Run the shared GRU to obtain $c_{1:T}^{\text{mod}}$.
3) For each modality, compute an InfoNCE loss using the hybrid negative sampler.

B. Audio Encoder

Each audio sample is a 1D waveform $x \in \mathbb{R}^T$, sampled at 16 kHz. We compute a Mel-spectrogram $S \in \mathbb{R}^{F \times T'}$ using torchaudio, treat it as a $(1, F, T')$ tensor, and apply:

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

We average over the frequency axis to obtain a sequence of latent vectors $z_t^{\text{aud}} \in \mathbb{R}^{z_d}$.

C. Image Encoder

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

$$\begin{aligned}
&\text{Conv2d}(3 \to 32,\ 4 \times 4,\ \text{stride} = 2) \to \text{BN} \to \text{ReLU},\\
&\text{Conv2d}(32 \to 64,\ 4 \times 4,\ \text{stride} = 2) \to \text{BN} \to \text{ReLU},\\
&\text{Conv2d}(64 \to 128,\ 4 \times 4,\ \text{stride} = 2) \to \text{BN} \to \text{ReLU},\\
&\text{Conv2d}(128 \to z_d,\ 3 \times 3,\ \text{stride} = 1).
\end{aligned}$$

The output feature map (z_d, H', W') is reshaped to sequence length $T = H'W'$ by flattening the spatial grid.

D. Shared Context GRU

For each modality, the latent sequence $z_{1:T}$ is passed through a GRU:

$$h_t, h_{t-1} = \text{GRU}(z_t, h_{t-1}), \quad c_t = W h_t, \tag{7}$$

where $h_t \in \mathbb{R}^{c_d}$ and $W \in \mathbb{R}^{z_d \times c_d}$. This yields modality-agnostic context vectors $c_t \in \mathbb{R}^{z_d}$.

E. Model Diagram

Fig. 1 summarizes the complete multi-modal CPC pipeline.

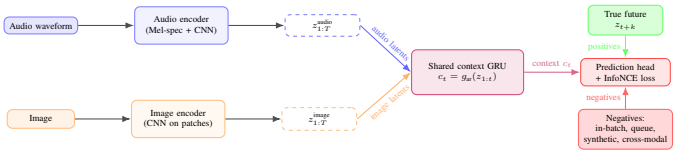

Fig. 1. Multi-modal CPC architecture: modality-specific encoders produce latent sequences, a shared GRU builds context vectors, and a hybrid negative sampler defines the InfoNCE objective.

Algorithm 1 Multi-Modal CPC Training with Hybrid Negatives

Require: Dataset $\mathcal{D}$, encoders $g_{\text{enc}}^{\text{aud}}, g_{\text{enc}}^{\text{img}}$, context GRU g_{ar}, negative samplers $\{\mathcal{S}_{\text{aud}}, \mathcal{S}_{\text{img}}\}$, epochs E, temperature τ

for epoch $= 1$ to E **do**
  for mini-batch B from $\mathcal{D}$ **do**
    Encode each modality:
      $Z^{\text{aud}} \leftarrow g_{\text{enc}}^{\text{aud}}(B_{\text{audio}})$
      $Z^{\text{img}} \leftarrow g_{\text{enc}}^{\text{img}}(B_{\text{image}})$
    Obtain context sequences:
      $C^{\text{aud}} \leftarrow g_{\text{ar}}(Z^{\text{aud}})$
      $C^{\text{img}} \leftarrow g_{\text{ar}}(Z^{\text{img}})$
    Build cross-modal negative pools from $Z^{\text{aud}}, Z^{\text{img}}$
    Compute audio loss:
      $L_{\text{aud}} \leftarrow \mathcal{S}_{\text{aud}}(C^{\text{aud}}, Z^{\text{aud}}, \tau, \text{cross-modal})$
    Compute image loss:
      $L_{\text{img}} \leftarrow \mathcal{S}_{\text{img}}(C^{\text{img}}, Z^{\text{img}}, \tau, \text{cross-modal})$
    $L \leftarrow \frac{1}{2}(L_{\text{aud}} + L_{\text{img}})$
    Backpropagate and update all trainable parameters
  end for
end for

F. Hybrid Negative Sampler

The *NegativeSampler* maintains a queue of past future-latent vectors and can synthesize negatives. For each modality and each time step t, we form a positive pair (c_t, z_{t+1}) and a candidate set of negatives:

- **In-batch** future latents from other samples.
- **Queue** entries from previous batches.
- **Synthetic** negatives (Gaussian perturbed and index-shuffled).
- **Cross-modal** latents from other modalities in the batch (optional).

Logits are computed as dot products scaled by temperature and passed to a cross-entropy loss with the positive index as label.

G. Training Algorithm

Algorithm 1 summarizes the overall training loop.

Algorithm 2 details the hybrid negative sampling for a single modality.

Algorithm 2 Hybrid Negative Sampling for One Modality

Require: Context $C \in \mathbb{R}^{B\times T\times d}$, latents $Z \in \mathbb{R}^{B\times T\times d}$, temperature τ, queue $\mathcal{Q}$, cross-modal pool $\mathcal{Z}^{\text{cm}}$

Form positives: $Z^{+} \leftarrow Z[:, 1:T-1, :]$, $C^{+} \leftarrow C[:, 0:T-2, :]$
Flatten: $Z_{\text{flat}}^{+}, C_{\text{flat}}^{+} \in \mathbb{R}^{N\times d}$, where $N = B(T-1)$
In-batch negatives: $Z_{\text{all}} \leftarrow Z_{\text{flat}}^{+}$
if $\mathcal{Q}$ not empty **then**
  $Z_{\text{all}} \leftarrow \text{concat}(Z_{\text{all}}, \mathcal{Q})$
end if
Generate synthetic negatives:
  $Z_{\text{pert}} \leftarrow Z_{\text{flat}}^{+} + \sigma \cdot \mathcal{N}(0, I)$
  $Z_{\text{shuf}} \leftarrow$ random permutation of Z_{flat}^{+}
$Z_{\text{all}} \leftarrow \text{concat}(Z_{\text{all}}, Z_{\text{pert}}, Z_{\text{shuf}})$
if $\mathcal{Z}^{\text{cm}}$ not empty **then**
  $Z_{\text{all}} \leftarrow \text{concat}(Z_{\text{all}}, \mathcal{Z}^{\text{cm}})$
end if
Compute logits: $\ell = C_{\text{flat}}^{+} Z_{\text{all}}^{\top} / \tau$
Labels: $y = [0, 1, \ldots, N-1]$
$L = \text{CrossEntropy}(\ell, y)$
Enqueue Z_{flat}^{+} into $\mathcal{Q}$ (FIFO)
return L

IV. Datasets

A. Design Rationale

Our initial plan was to use larger, real-world datasets across text, audio and images. In practice, two constraints dominated:

- **Compute**: training CPC with large encoders on full-scale datasets is expensive on CPU-only hardware.
- **Audio tooling**: incompatibilities between recent torchaudio and system FFmpeg/TorchCodec libraries made heavy real audio corpora unreliable.

Rather than fight these limitations, we adopted a controlled setup:

- a two-class CIFAR-10 subset for images, and
- synthetic but structured sine waves for audio.

This makes our experiments highly interpretable and better suited to isolate representation effects.

B. Image Dataset: Two-Class CIFAR-10

We use the CIFAR-10 training split and keep only labels 0 (airplane) and 1 (automobile). Images are resized to 128×128 and stored as:

```
class0/   # airplanes
class1/   # automobiles
```

We cap the dataset at approximately 2 000 images per class. This yields a binary problem with non-trivial intra-class variation but manageable training time.

C. Audio Dataset: Structured Sine Waves

Each audio sample is a synthetic waveform:

- Single sine wave with frequency $f \sim \mathcal{U}(150, 900)$ Hz.

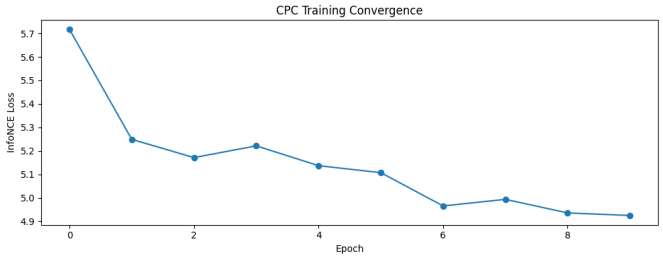

Fig. 2. Training curve (InfoNCE loss vs. epoch) for audio-only CPC with hybrid negatives.

- Additive Gaussian noise.
- Sampled at 16 kHz for fixed duration.

We label samples as

$$y = \begin{cases} 0 & f < 500 \text{ Hz} \\ 1 & f \geq 500 \text{ Hz}. \end{cases} \tag{8}$$

The underlying factor of variation (frequency) is thus well defined, making it easy to interpret context geometry.

D. Synthetic Multi-Modal Dataset

For multi-modal experiments we construct paired samples:

- The audio component is a sine wave as above.
- The image component is a 64×64 RGB image with either horizontal or vertical bar patterns plus noise.
- A shared binary label inherited from the audio frequency class.

This dataset lets us study whether a shared context model can encode consistent semantics across both modalities.

V. Experiments and Results

All experiments are run with latent dimension $z_d = 64$, context dimension $c_d = 256$, batch size 32, and Adam optimizer with learning rate 10^{-3} unless otherwise stated.

A. Experiment 1: Audio CPC with Hybrid Negatives

We first train CPC on synthetic audio for 10 epochs using the queue+synthetic negative mode. The training loss is shown in Fig. 2. The loss decreases smoothly and monotonically, indicating that the model learns to distinguish true future latents from negatives without instability.

After training, we extract final context vectors (last time step per sample), normalise them and run t-SNE. Fig. 3 shows the resulting 2D embedding.

We observe:

- Two smooth, S-shaped manifolds corresponding to low- and high-frequency waves, touching only at a single location.
- Alignment ≈ 0.43 and uniformity in the range $[-1.7, -1.3]$, indicating compact but well-spread embeddings.
- A logistic regression probe on frozen contexts achieves $\approx 99\%$ accuracy.

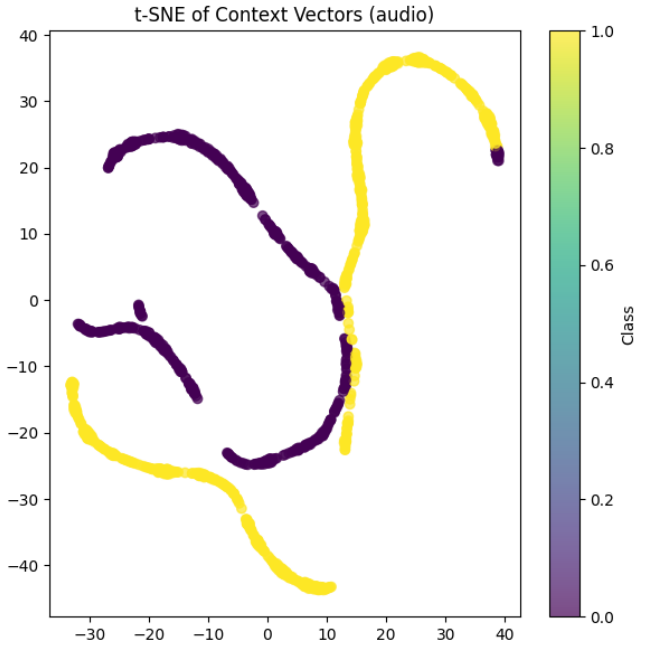

Fig. 3. t-SNE of audio context vectors; colours indicate low vs. high frequency classes.

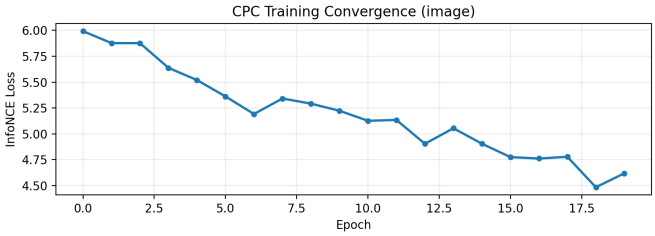

Fig. 4. Training curve for image CPC on two-class CIFAR-10.

CPC thus recovers the underlying frequency-based structure without ever seeing labels.

B. Experiment 2: Image CPC on Two-Class CIFAR-10

We then train CPC on the two-class CIFAR-10 dataset for 20 epochs. Compared to early unstable experiments, we slightly reduced queue size and synthetic-noise level; the resulting training loss in Fig. 4 is now stable and steadily decreasing.

Fig. 5 illustrates the t-SNE of image contexts. The airplane and automobile clusters form two relatively compact clouds that are clearly separated along one axis, with mild overlap near the boundary.

The alignment and uniformity metrics are weaker than in the audio case but still favourable: contexts are somewhat less tightly aligned yet reasonably uniform. A logistic regression probe on frozen image contexts achieves around 90–92% accuracy, demonstrating that the CPC encoder plus context

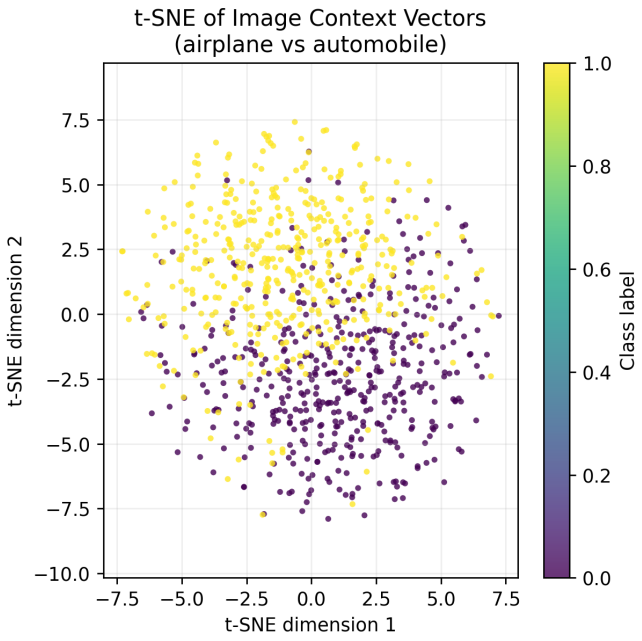

Fig. 5. t-SNE of image context vectors for airplane vs. automobile classes.

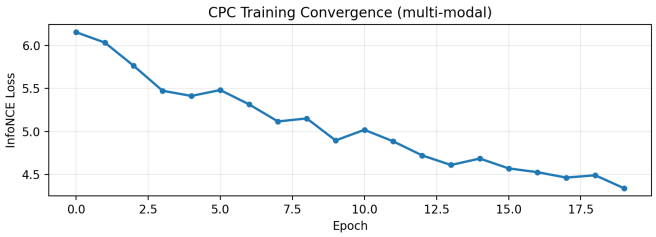

Fig. 6. Training curve in the multi-modal setting (sine-wave audio + bar-pattern images) with cross-modal negatives.

model learns non-trivial semantic structure even in the more challenging visual domain.

C. Experiment 3: Multi-Modal Synthetic CPC

Finally, we train CPC on the synthetic multi-modal dataset with cross-modal negatives enabled. Audio contexts remain highly structured and almost linearly separable by frequency, while image contexts inherit some of this structure through the shared GRU and cross-modal negatives.

Fig. 6 shows a typical training curve in this setting. The loss decreases smoothly, with no major spikes despite the richer negative pool.

Fig. 7 shows the t-SNE of audio contexts under multi-modal training. Compared to audio-only training, the two frequency classes still form smooth "snakes", but they are slightly more diffused and touch at a single region in the plane, reflecting the additional variability introduced by cross-modal negatives.

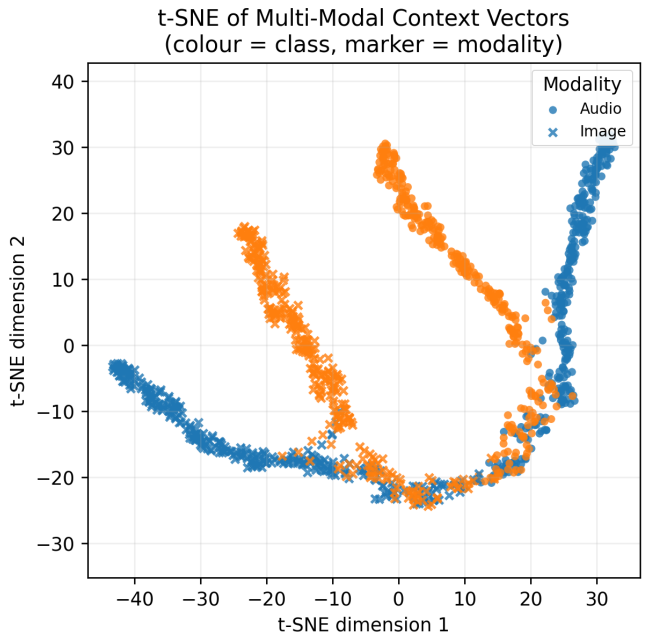

Fig. 7. t-SNE of audio context vectors when trained jointly with images and cross-modal negatives. The two frequency classes form slightly diffused, intersecting "snakes". Orange markers for Class 0 and Blue Markers for Class 1.

Importantly, linear probes remain strong ($>98\%$ for audio, $\sim 90\%$ for the synthetic bar images), and alignment/uniformity only degrade slightly compared to the uni-modal runs. This suggests that the hybrid negative sampler, including cross-modal contributions, does not destroy the useful structure in the audio context space and in fact encourages a shared semantic geometry across modalities.

VI. Discussion

A. Effect of Negative Sampling

Our experiments confirm that negative sampling is not one-size-fits-all:

- For the structured audio dataset, a relatively large queue plus synthetic perturbations works well and enhances discrimination without destabilizing training.
- For images, the same aggressive configuration is too harsh; after reducing queue size and noise level, training becomes stable and yields good, though slightly noisier, class separation.

Practitioners should therefore tune negative sampling per modality rather than assuming that a configuration that works for one will transfer to another.

B. Quantifying Context Quality

Alignment, uniformity, linear probes and t-SNE provide a more complete picture than any single number:

- In audio, all metrics agree: low alignment, good uniformity, high probe accuracy, and clean 2D snakes with near-perfect colour separation.
- In images, the metrics indicate slightly higher alignment and less uniformity, and the t-SNE shows partially overlapping but clearly biased clusters, matching the $\sim 90\%$ probe accuracy.

This multi-view evaluation is particularly useful when comparing different negative sampling strategies that may produce similar training losses but very different latent geometries.

C. Role of Dataset Design

The switch to smaller, controlled datasets—a two-class CIFAR-10 subset and synthetic audio—was driven by compute and software constraints, but it is actually beneficial. It allows us to:

- Attribute representation differences directly to modality and negative sampling rather than to uncontrolled noise.
- Interpret t-SNE plots and alignment/uniformity in terms of well-understood underlying factors (e.g. frequency).

VII. Limitations and Future Work

Despite promising results, several limitations remain:

- **Model capacity**: encoders and context GRU are intentionally small to fit CPU-only training; deeper models may behave differently.
- **Dataset diversity**: our datasets are carefully designed but much simpler than full-scale audio and image corpora.
- **Hyperparameters**: we explored particular queue sizes, perturbation levels and temperatures. A broader sweep could uncover regimes where images benefit from different negative sampling settings.

Future work could scale the architecture to stronger encoders and larger datasets, explore adaptive or learned negative sampling schemes, and extend the multi-modal framework to more complex pairing strategies.

VIII. Conclusion

We implemented a multi-modal Contrastive Predictive Coding framework with:

- separate audio and image encoders,
- a shared GRU-based context model, and
- a hybrid negative sampler combining in-batch, queue, synthetic and cross-modal negatives.

On a controlled experimental setup with synthetic audio and a two-class CIFAR-10 subset, we showed that:

- CPC learns almost perfectly linearly separable audio contexts that align with frequency classes.
- With tuned hyperparameters, the same framework learns competitive image representations with clear class separation and $\sim 90\%$ linear-probe accuracy.
- In a multi-modal setting, cross-modal negatives encourage a shared semantic context space while preserving strong unimodal structure.
- Alignment/uniformity metrics, linear probes and t-SNE together offer a useful toolkit for judging "better" context summarization beyond raw loss.

Overall, the project not only reproduces the core CPC idea but also extends it with a practical multi-modal and hybrid-negative perspective tailored to realistic compute constraints.

Acknowledgment

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. van den 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. M. 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 *Int. Conf. on Learning Representations (ICLR)*, 2022.

[4] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2495–2504.