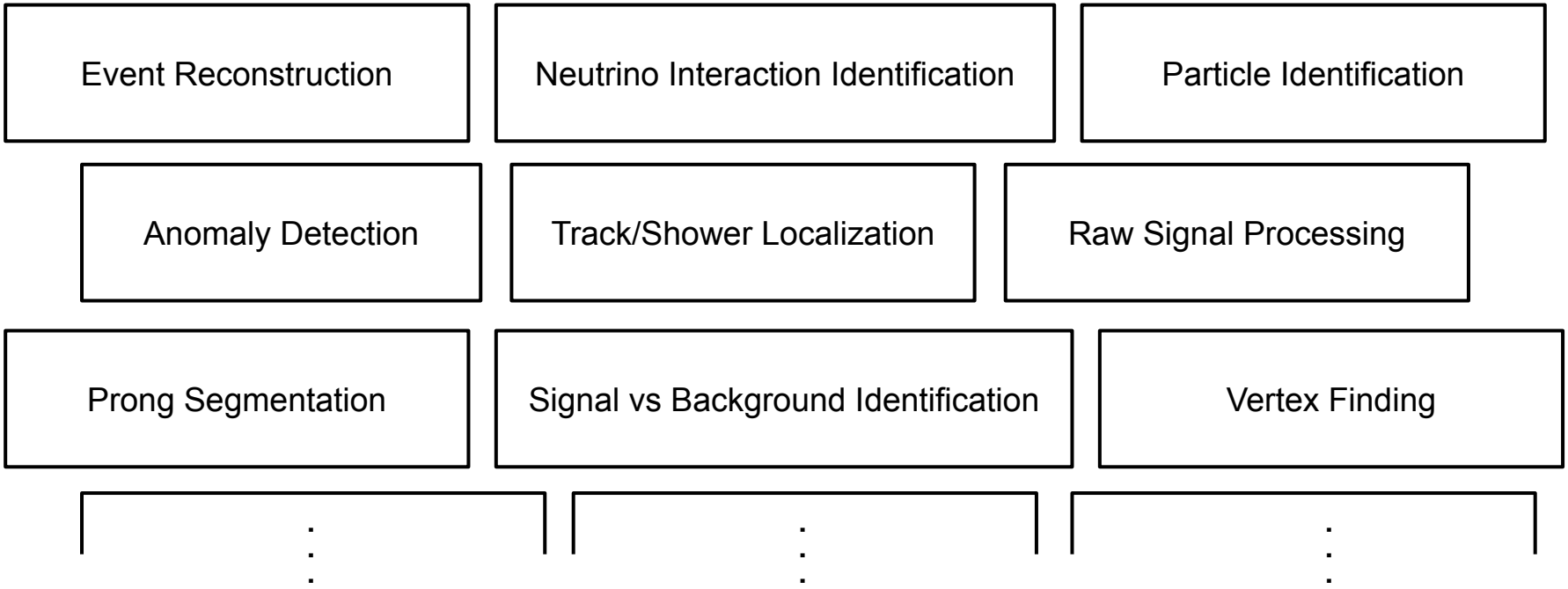
Adapting Vision-Language Models for Neutrino Interactions in High-Energy Physics

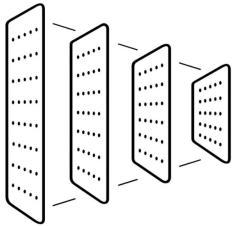
Dikshant Sagar, Kaiwen Yu, Alejandro Yankelevich, Jianming Bian, Pierre Baldi

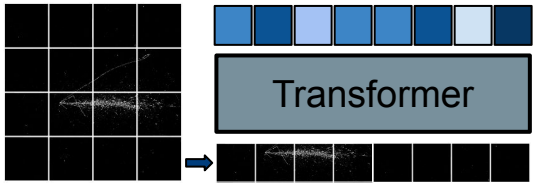
NPML 2026

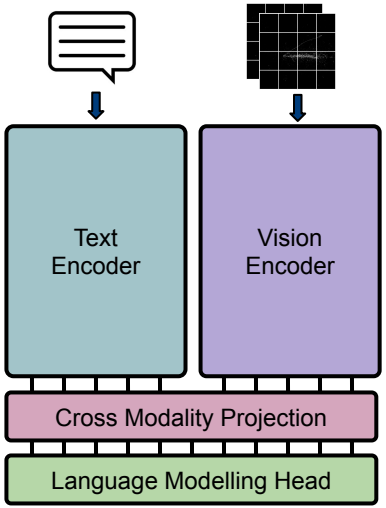
Deep Learning Tasks in Neutrino Physics



Evolution of Architectures in HEP







CNNs

- Strong local feature extraction
- Widely used in HEP
- Limited global reasoning

Vision Transformers (ViTs)

- Global self-attention
- Better long-range dependency modeling
- More robust to topology variation

Vision-Language Models (VLMs)

- Joint visual + semantic reasoning
- Natural-language outputs
- Foundation-model potential

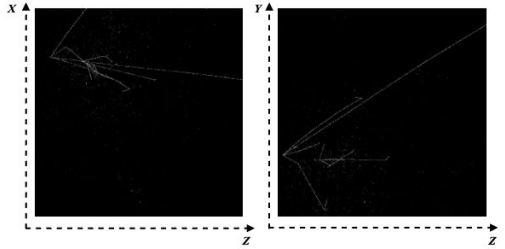
Why Vision-Language Models (VLMs) in HEP?

Unifying **image-like detector data + semantic reasoning**

Natural interface for physicists (prompts, captions, explanations)

Potential roles in:

- Event classification (CC vs NC)
- Topology recognition
- Anomaly detection
- Human-in-the-loop analysis



Classify the attached pixel maps as ν_e CC, ν_μ CC or Neutral Current.

In the given pixel maps, the muon track is longer and narrow, which suggests that the event is ν_μ CC.

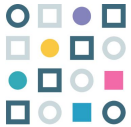
https://www.nature.com/articles/s42005-026-02688-3

But Why Vision Language Models ?

Moving beyond traditional ML toward foundation models

Data at Unprecedented Scale

DUNE Far Detector will generate ~ **30 PB/year of raw LArTPC data**. Traditional reconstruction pipelines require rethinking. Foundation models offer a single-model solution across multiple tasks



Complex Visual Patterns

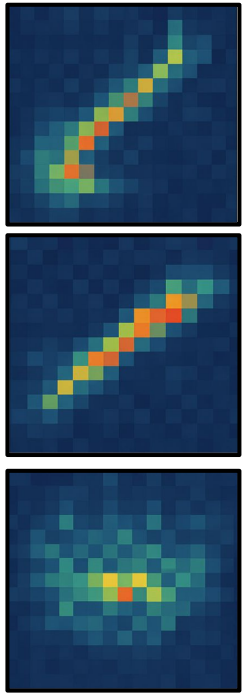
3D LArTPC wire images contain **intricate topology**: track vs. shower separation, neutrino flavor ID, vertex reconstruction, exactly the type of tasks where vision models excel.

Generalization Across Tasks

Foundation models **adapt** across simulation, calibration, and trigger tasks via fine-tuning, reducing duplicated ML development across working groups.

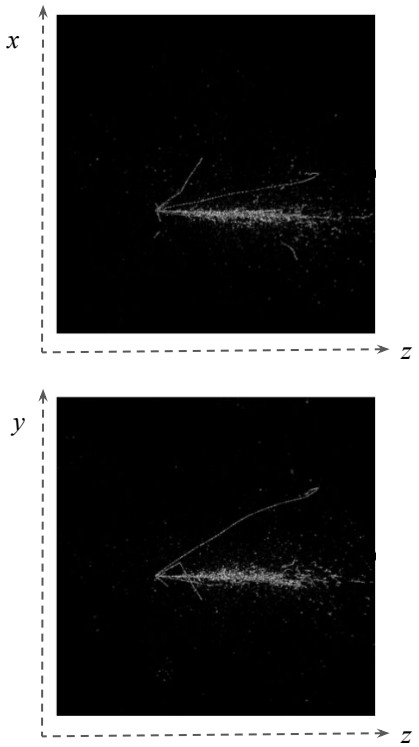
Problem Definition

- **Task:** Classify neutrino interactions in pixelated LArTPC detector data
  - **Interactions:** ν_e CC, ν_μ CC, Neutral Current (NC)
- **Importance:**
  - Distinguishing neutrino flavors → crucial for oscillation studies
  - Need interpretable, generalizable ML models.



Dataset

- Custom simulated **LArTPC** detector (2m × 2m × 7m)
- Generated using **GENIE** (v3.0.6) and **GEANT4** (v11.2.0)
- 190,000 simulated events (ν_e and ν_μ) with uniform energy flux in 0-10 GeV with beam direction along z-axis.
  - 74% charged current, 26% neutral current
- Limited detector realization with smearing based on drift electron diffusion.
- Smeared energy deposition pooled into **512 X 512** grayscale pixel maps.
  - XZ and YZ views



https://www.nature.com/articles/s42005-026-02688-3

Few-Shot In-Context Evaluation

```
{
   "role": "user",
   "content": [
       {"type": "image", "image": ex["zx"]},
       {"type": "image", "image": ex["zy"]},
       {
           "type": "text",
           "text": "Classify the attached pixel maps as ν_e CC, ν_μ CC or Neutral Current."
       }
   ]
}
```

```
{
   "role": "assistant",
   "content": [
       {
           "type": "text",
           "text": f"I classify the pixel maps as {ex['label']}."
       }
   ]
}
```

All events predicted as ν_e**CC**. Accuracy: **36.78%**

Reason: visual features learned during LlaMa's pre-training are **insufficiently aligned** with the domain-specific semantics of sparse detector images.

Proposed Approach with LLaMA 3.2 Vision

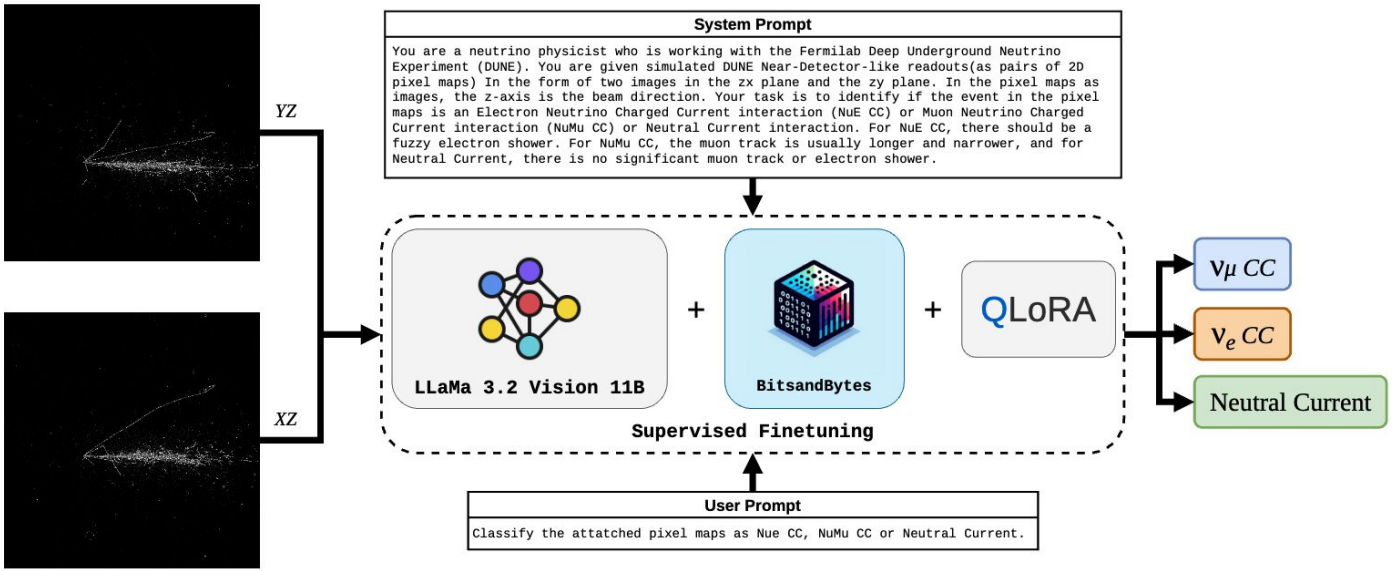


Figure 1: **LLaMA 3.2 Vision Model Finetuning Overview:** Fine-tuning overview of the LLaMA 3.2 Vision Model for neutrino event classification. Pixel map projections (YZ and XZ) are provided as input, combined with a physics-informed system prompt, and used in a supervised fine-tuning pipeline with BitsAndBytes and QLoRA to classify events into ν_μ charged current, ν_e charged current, or neutral current categories.

https://www.nature.com/articles/s42005-026-02688-3

Memory & Compute Bottlenecks

Where the Compute Goes

- Vision encoder (image resolution sensitive)
- Cross-attention layers
- Token generation (for text-heavy tasks)

Practical Constraints

- Image resolution ↑ → quadratic memory growth
- Batch size often limited to 1-4
- Multi-GPU communication overhead dominates at scale. (NVLink ! or HGX)

Example (Illustrative)

- 11B VLM + QLoRA (r=8):
    - ~43 GB VRAM per GPU for fine tuning.

https://www.nature.com/articles/s42005-026-02688-3

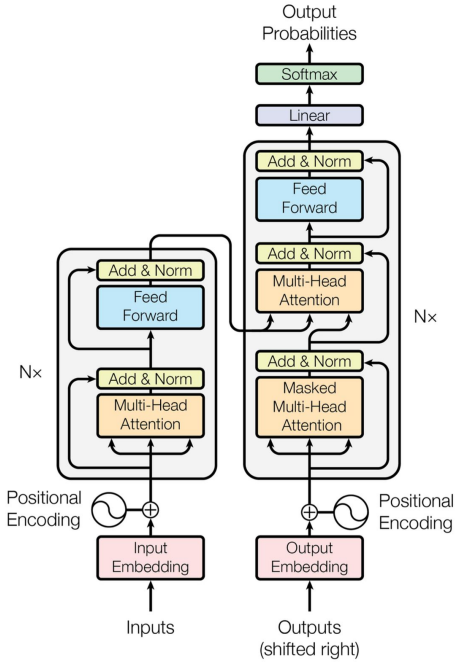
Training Time Memory Breakdown (Per GPU)

For an ~11B VLM:

- **Weights (BF16):** ~22 GB
- **Optimizer states (Adam):** ~44 GB
- **Activations (batch dependent):** 10–30 GB
- Total (full fine-tune): **80-100+ GB**

With QLoRA:

- Frozen base weights quantized.
- Trainable adapter params <1% (rank dependent)
- Memory drops to **~43 GB**



https://www.nature.com/articles/s42005-026-02688-3

Why QLoRA ?

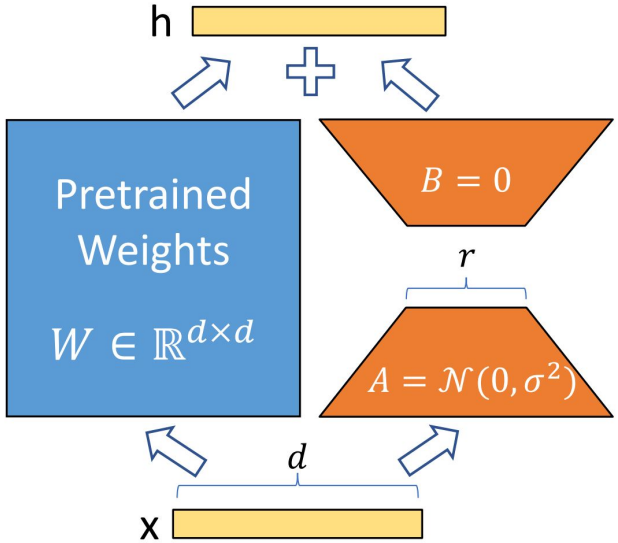
Parameter-Efficient Fine-Tuning

Benefits

- Only 29.5M trainable parameters
- Preserve pretrained knowledge
- Trainable on 4× A6000 GPUs

Training Setup

- 1 epoch
- Effective batch size = 8
- ~1 week training time



https://www.nature.com/articles/s42005-026-02688-3

Model Inference

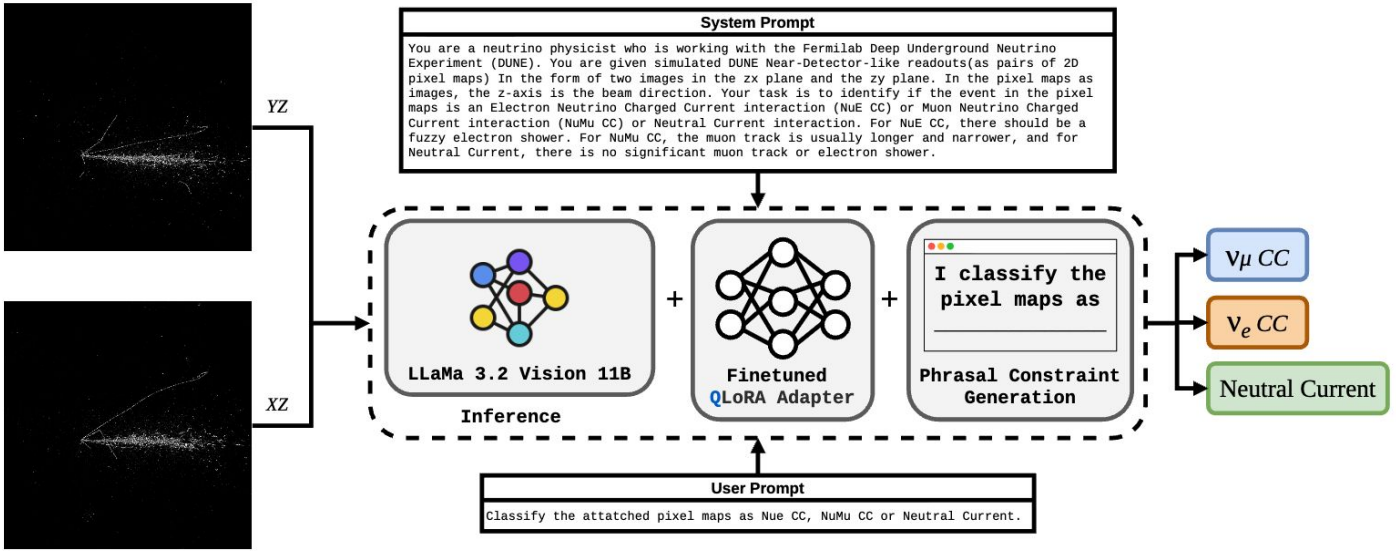


Figure 2: **LLaMA 3.2 Vision Model Inference Overview:** Inference pipeline for the fine-tuned LLaMA 3.2 Vision Model. YZ and XZ pixel map projections from the detector are processed with a physics-informed system prompt, passed through the base model with a fine-tuned QLoRA adapter, and decoded using constrained generation to produce classifications of ν_μ charged current, ν_e charged current, or neutral current events.

https://www.nature.com/articles/s42005-026-02688-3

Other Baselines

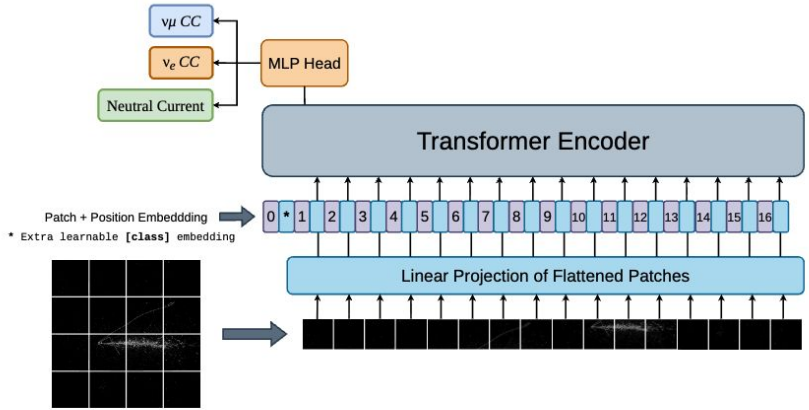


Figure 4: **ViT-h/14 Architecture:** ViT-h/14 splits an image into 14x14 patches, linearly embeds them, adds positional embeddings, and feeds the resulting sequence of vectors to a standard Transformer encoder, which in turn feeds into a classification MLP head.

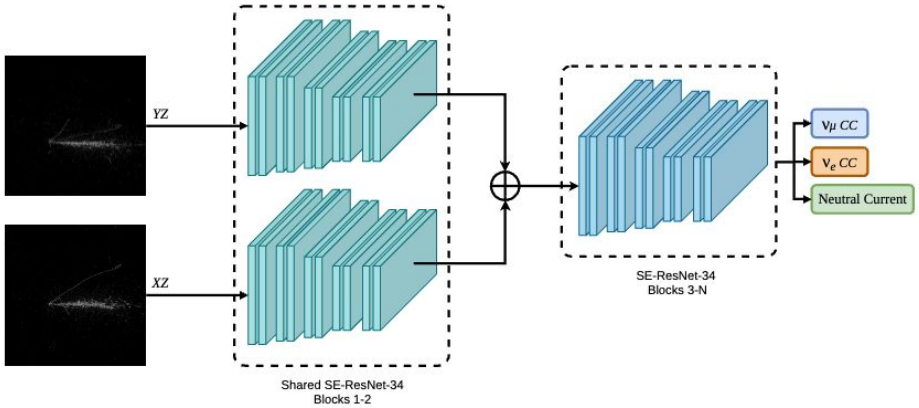


Figure 5: **CNN Architecture:** Simplified diagram of the CNN architecture based on [5]. The model takes in pixel maps in the x–z and y–z projections for a simulated LArTPC event and produces an event class output.

Results

Table 1: Event classification aggregated metrics.

Metric	LLaMA 3.2 Vision	ViT-h/14	CNN
Accuracy	0.87	0.86	0.80
Precision	0.87	0.86	0.80
Recall	0.87	0.85	0.79
AUC-ROC	0.96	0.96	0.94
# of Trainable Parameters	29.5M (QLoRA)	632M	21.7M
Training Regime	PEFT, 1 epoch	Full, 10 epochs	Full, 300 epochs
Inference Memory Usage (GB)	12.91	2.56	2.44
Time per Sample (ms)	3412	299.1	23.90

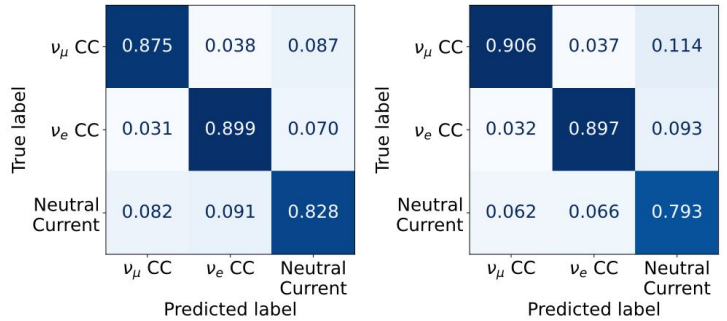


Figure 8: Finetuned LLaMA 3.2 Vision's (a) recall matrix (truth normalized) and (b) precision matrix (prediction normalized).

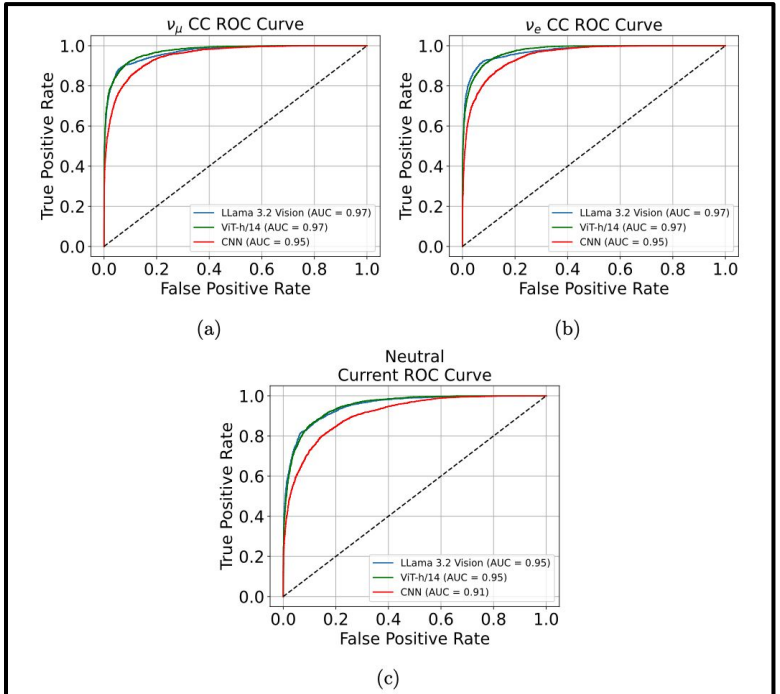


Figure 11: ROC curves for each class (a) ν_μ CC, (b) ν_e CC, and (c) NC comparing performance between the finetuned LLaMA 3.2 Vision and the CNN.

Results

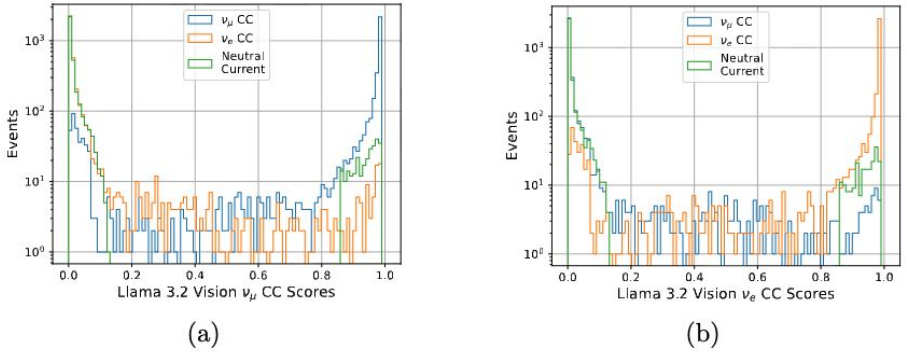
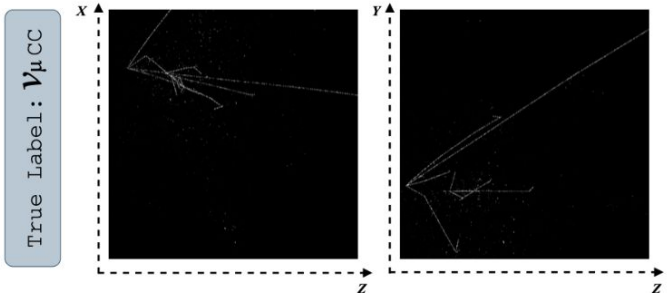


Figure 6: **LLaMA 3.2 Vision Model Event Classification Rejection Curves:** LLaMA 3.2 Vision model's (a) ν_μ CC event signal-background rejection curves, (b) ν_e CC confidence plot, and (c) NC confidence plot. Blue curves belong to ν_μ CC events, Orange curves belong to ν_e CC events, and Green curves belong to Neutral Current events.

https://www.nature.com/articles/s42005-026-02688-3

Interpretability and Explainability



Fine-tuned LLaMa 3.2 Vision's Explanation

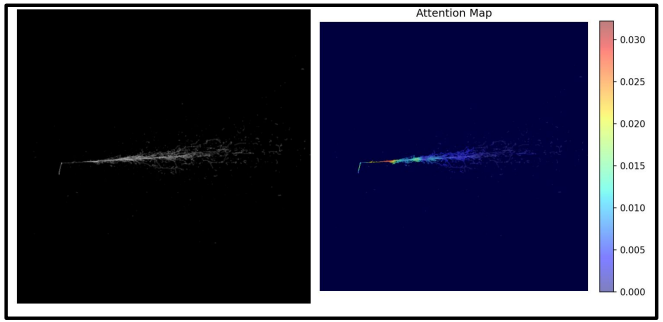
The pixel maps attached are the z-x and z-y maps of a DUNE Near-Detector-like readouts.
The z-axis is the beam direction. To classify the event as NuE CC, NuMu CC or Neutral Current,
we need to analyze the pixel maps and look for the following features:

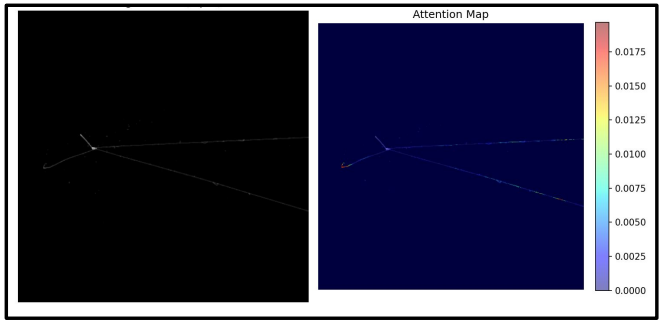
For NuE CC, there should be a fuzzy electron shower.
For NuMu CC, the muon track is usually longer and narrow.
For Neutral Current, there is no significant muon track or electron shower.

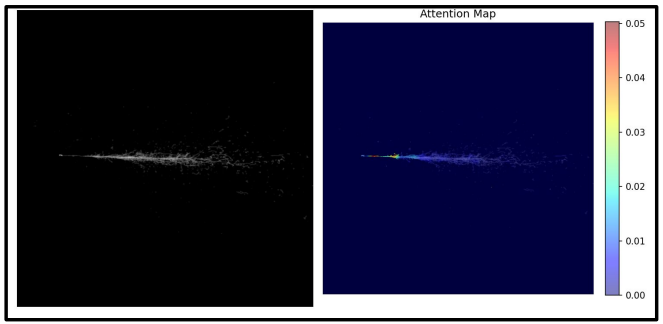
In the given pixel maps, the muon track is longer and narrow, which suggests that the event is NuMu CC.
Therefore, the classification of the event is NuMu CC.

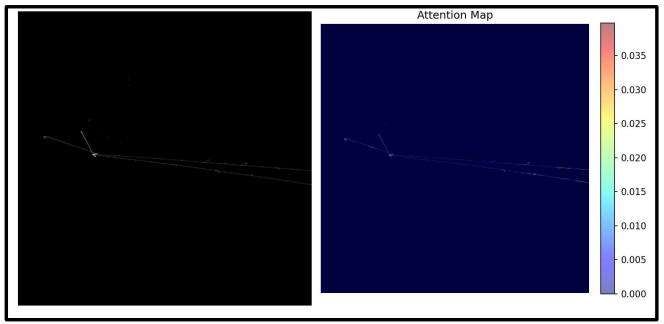
https://www.nature.com/articles/s42005-026-02688-3

Looking at what the Vision Encoder is looking at









ν_e CC

ν_μ CC

Generalization Testing

Downsampled Detector Images

Test:

- 512×512 → 256×256

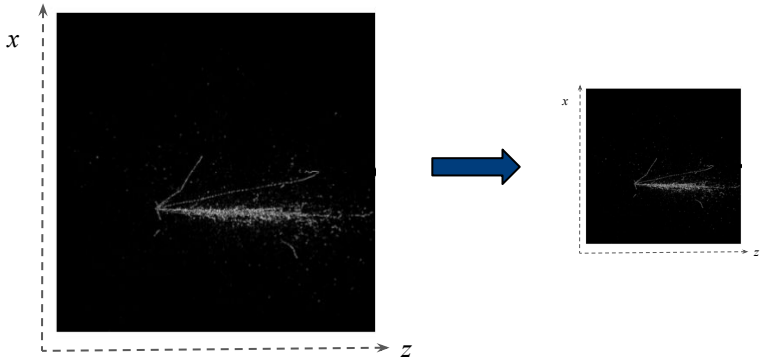
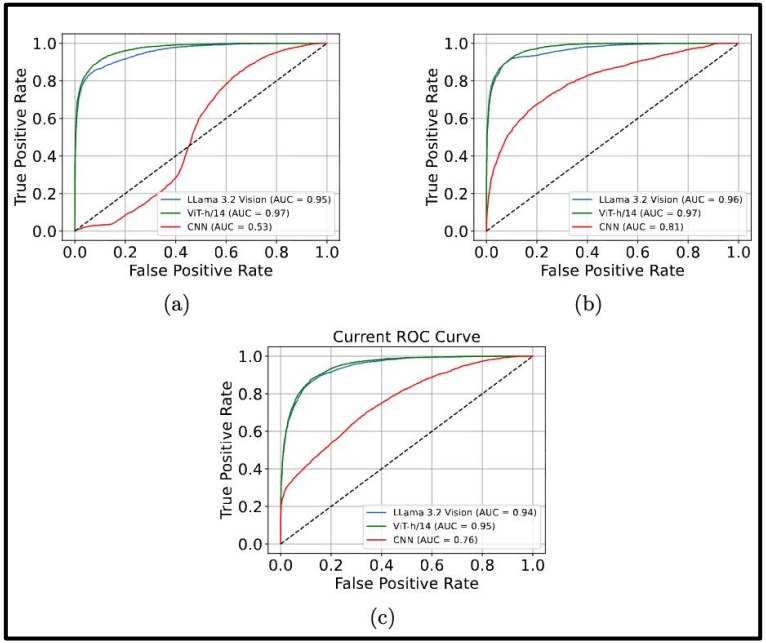


Table 2: Event classification aggregated metrics for generalization testing.

Metric	LLaMA 3.2 Vision	ViT-h/14	CNN
Accuracy	0.85	0.85	0.43
Precision	0.85	0.85	0.4
Recall	0.85	0.85	0.41
AUC-ROC	0.95	0.96	0.70



https://www.nature.com/articles/s42005-026-02688-3

Ablation Study: Role of Physics Definitions in the System Prompt

System Prompt

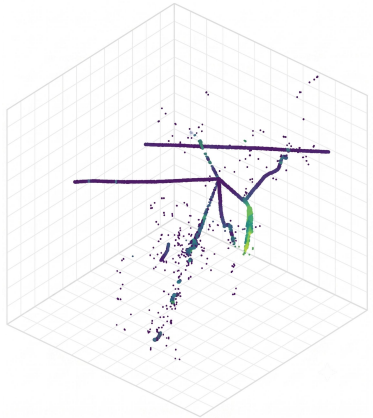
```
You are a neutrino physicist who is working with the Fermilab Deep Underground Neutrino
Experiment (DUNE). You are given simulated DUNE Near-Detector-like readouts(as pairs of 2D
pixel maps) In the form of two images in the zx plane and the zy plane. In the pixel maps as
images, the z-axis is the beam direction. Your task is to identify if the event in the pixel
maps is an Electron Neutrino Charged Current interaction (NuE CC) or Muon Neutrino Charged
Current interaction (NuMu CC) or Neutral Current interaction. For NuE CC, there should be a
fuzzy electron shower. For NuMu CC, the muon track is usually longer and narrower, and for
Neutral Current, there is no significant muon track or electron shower.
```

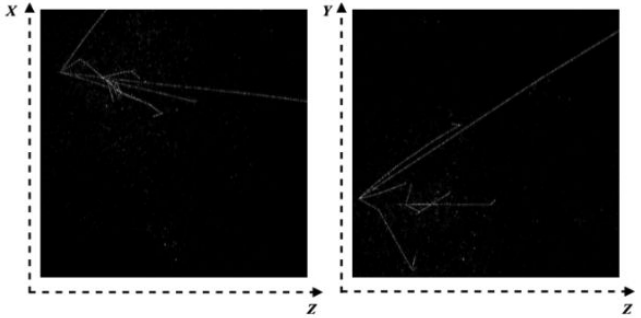
Qualitatively, the model continues to produce explanations that reference salient visual features of the detector images, such as long track-like structures, localized electromagnetic activity, or the absence of visible charged-lepton signatures.

Quantitatively, the model achieved an accuracy of **0.86 (↓0.01)**, precision of **0.86 (↓0.01)**, recall of **0.86 (↓0.01)**, and an AUC-ROC of **0.96** under the ablated prompt condition.

Limitations / Failure Modes

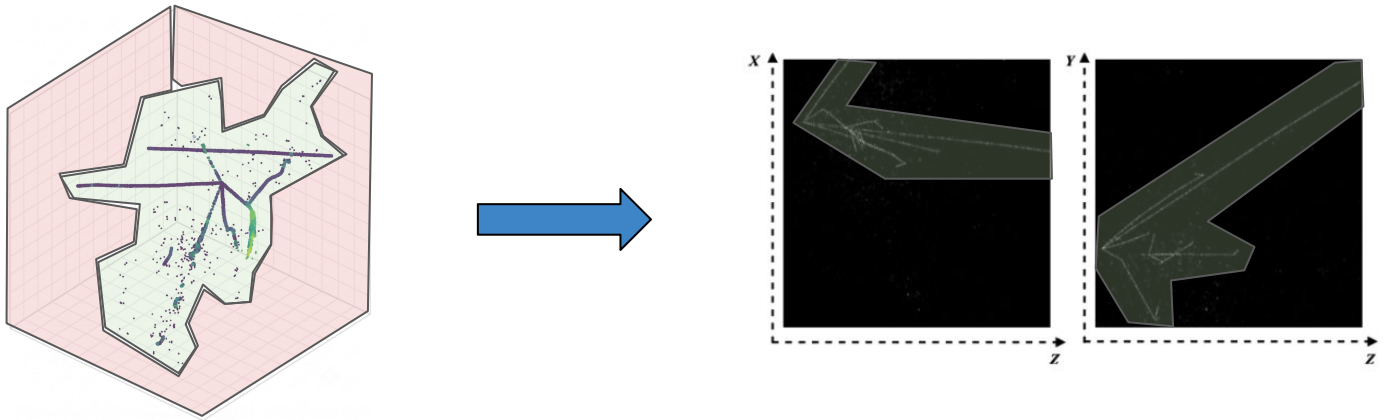
Sparse Detector Data





Limitations / Failure Modes

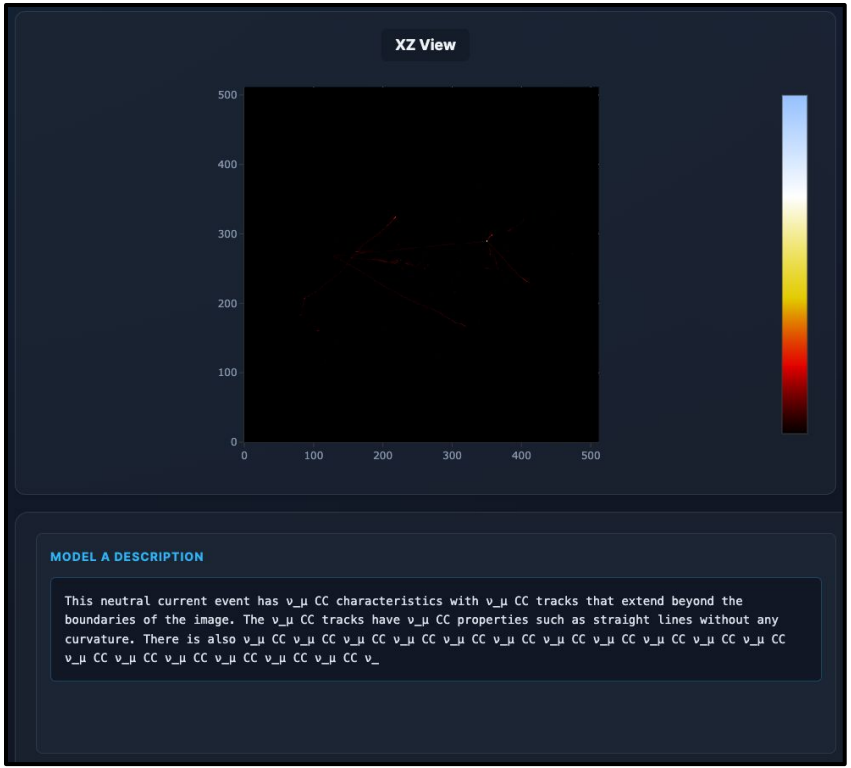
Sparse Detector Data



Limitations / Failure Modes

Ad-Hoc Explanations

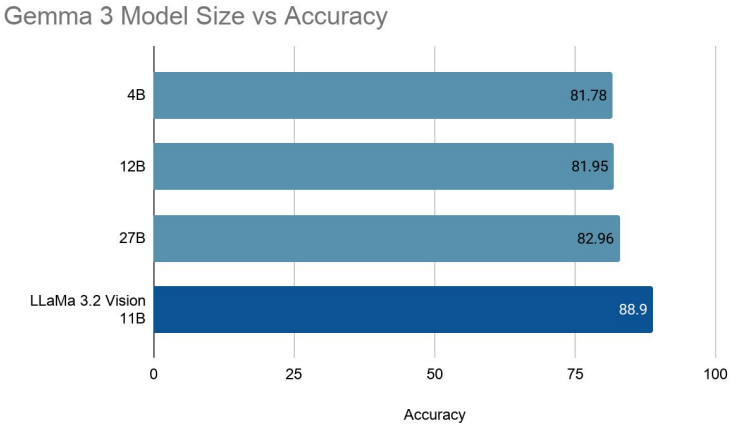
- Without expert validations
- Computationally expensive (autoregressive)



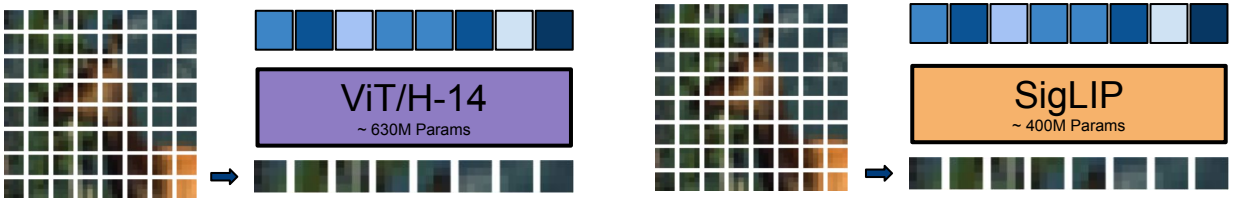
Ongoing Work

Scaling Law Check

Gemma 3



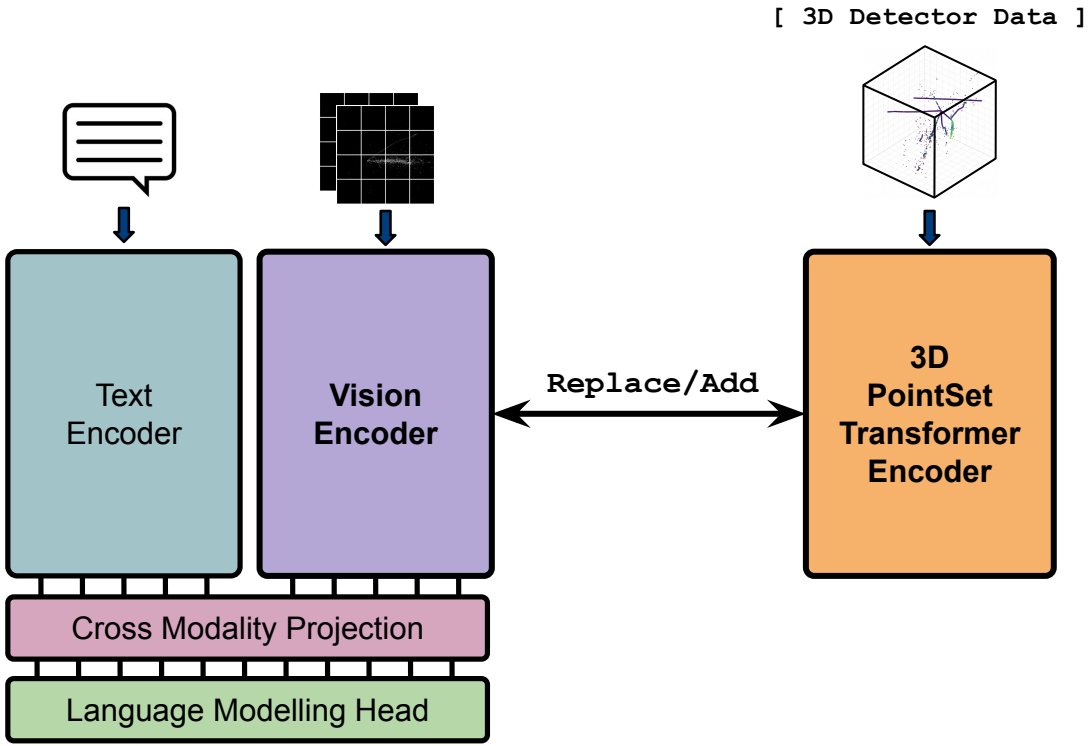
Encoder Size, Pre-Training and Features Matter



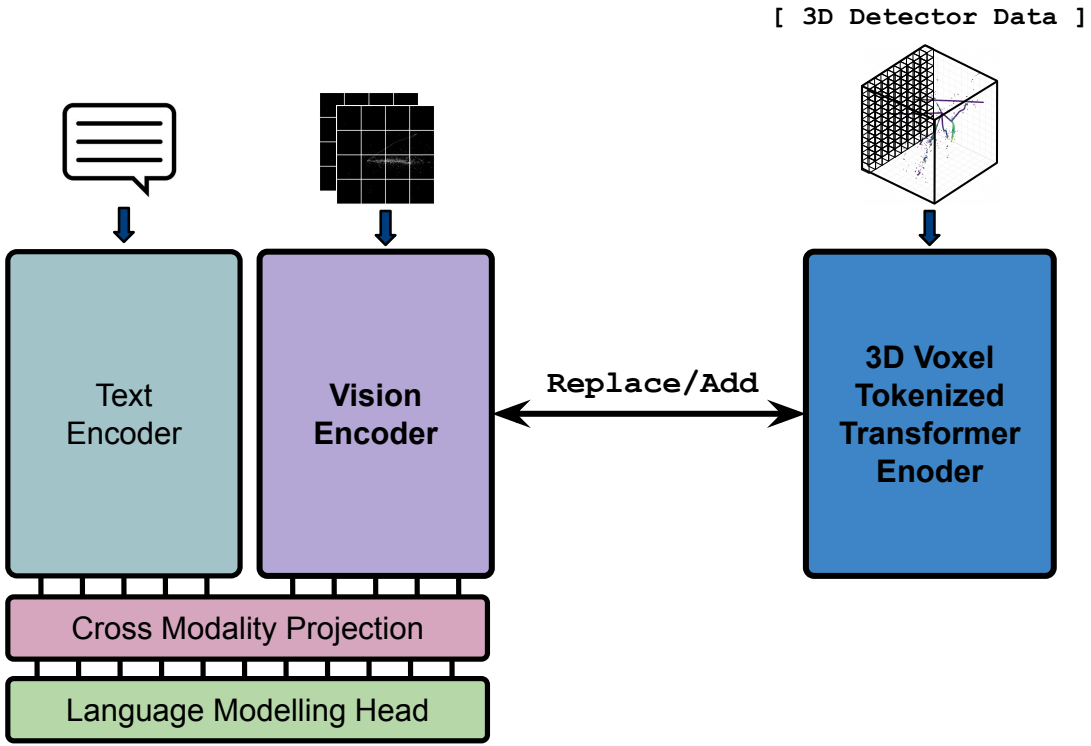
Loss Function	Softmax Contrastive Loss (Standard CLIP)	Sigmoid Loss (Pairwise, avoids global normalization)
Pre-training Scale	6 Billion image-text pairs	Heavily curated Web data & Visual Assistant tasks
Encoder Tuning	Co-trained / Unfrozen during main stages	Frozen completely during training
Image Compression	Keeps large token counts via cross-attention ~ 1,600 tokens	Compresses images into exactly 256 soft tokens

https://arxiv.org/html/2503.19786v1 https://huggingface.co/docs/transformers/model_doc/mllama https://github.com/meta-pytorch/torchtune/discussions/2462

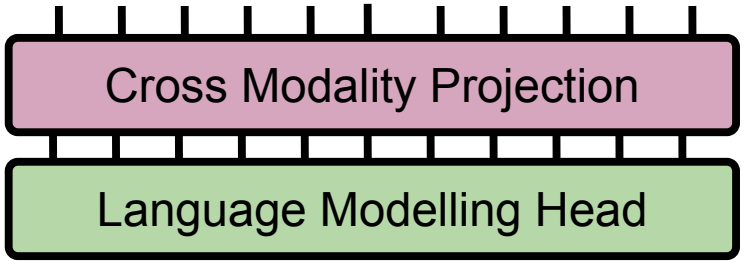
Data Native Encoder



Data Native Encoder



Things are never that easy…



Modality projector becomes invalid.

Data Native Encoder’s features come from a completely different manifold.

Catastrophic forgetting during alignment

Distribution shift in feature norms

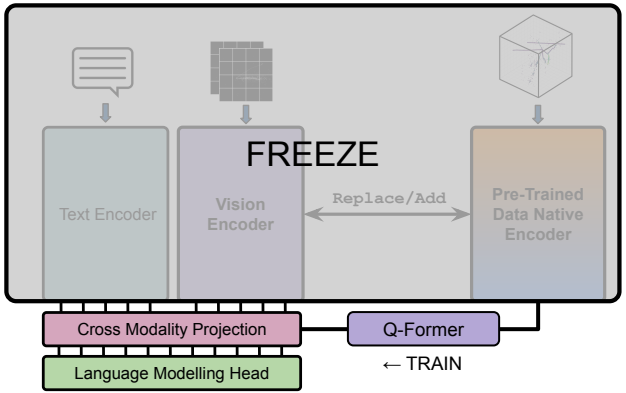
Attention Collapse

Context-window inefficiency

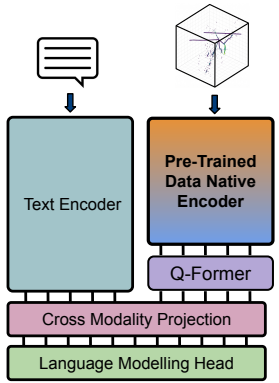
Insufficient detector-data pretraining

But maybe we can ease into it …

Stage 1

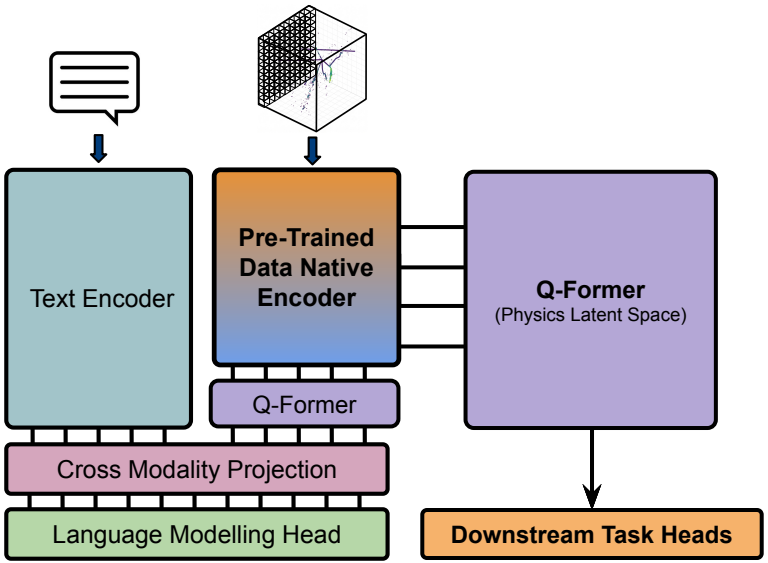


Stage 2

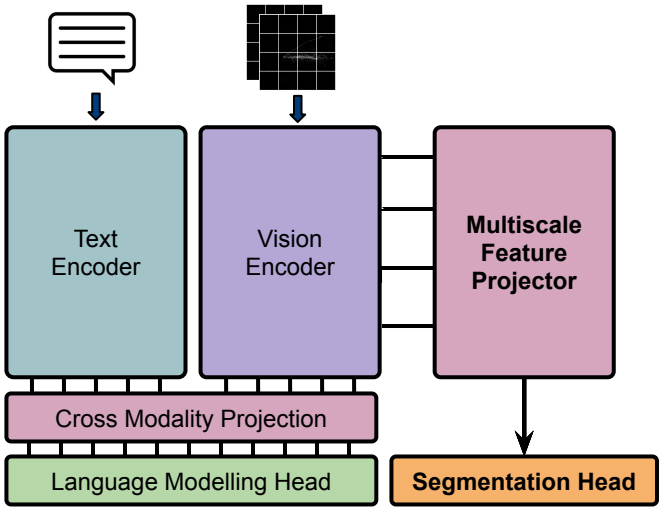


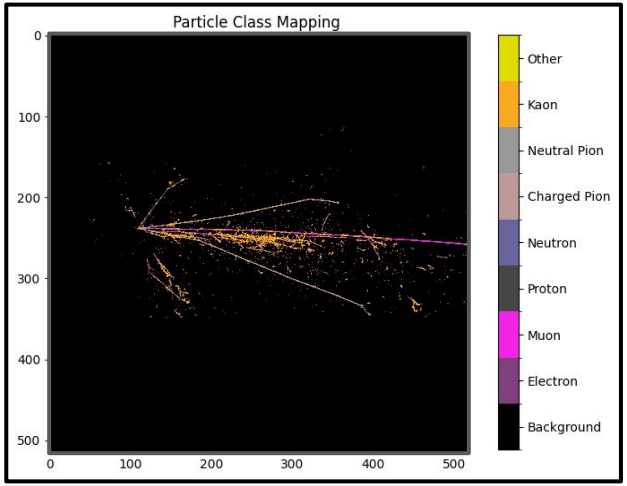
Full QLoRA Finetune

The Foundation Model



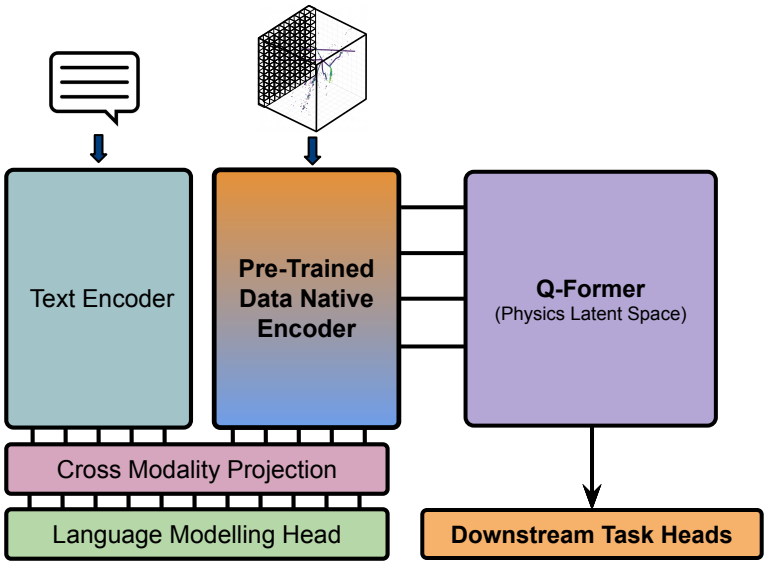
Capability Enhancements - Particle Segmentation



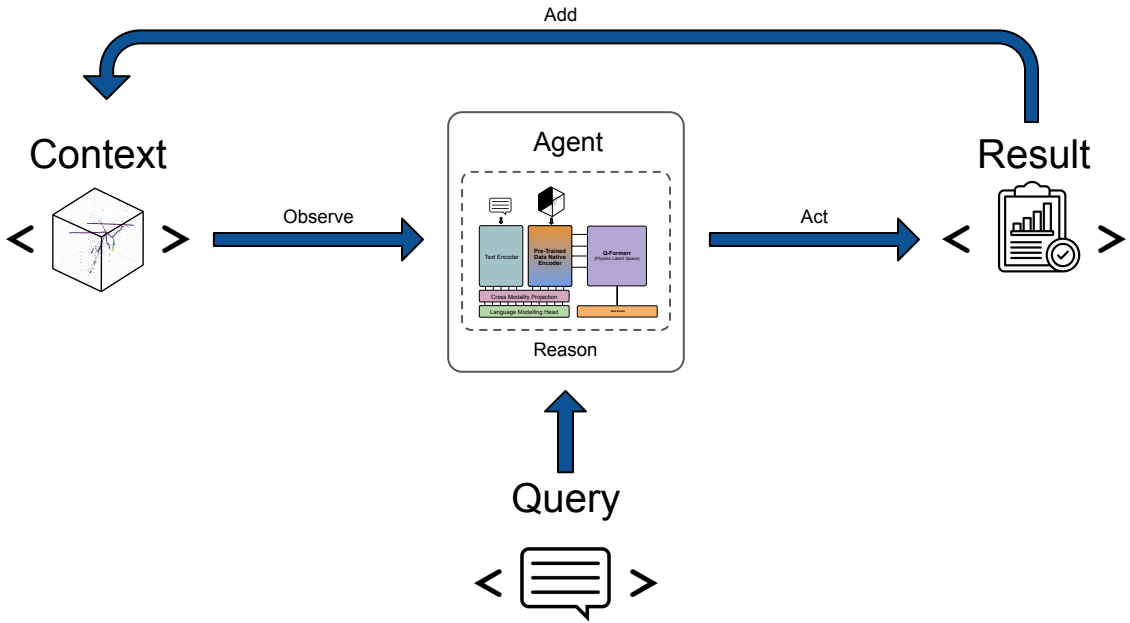


This Thing: **64.70%**	HPST: **72.40%**

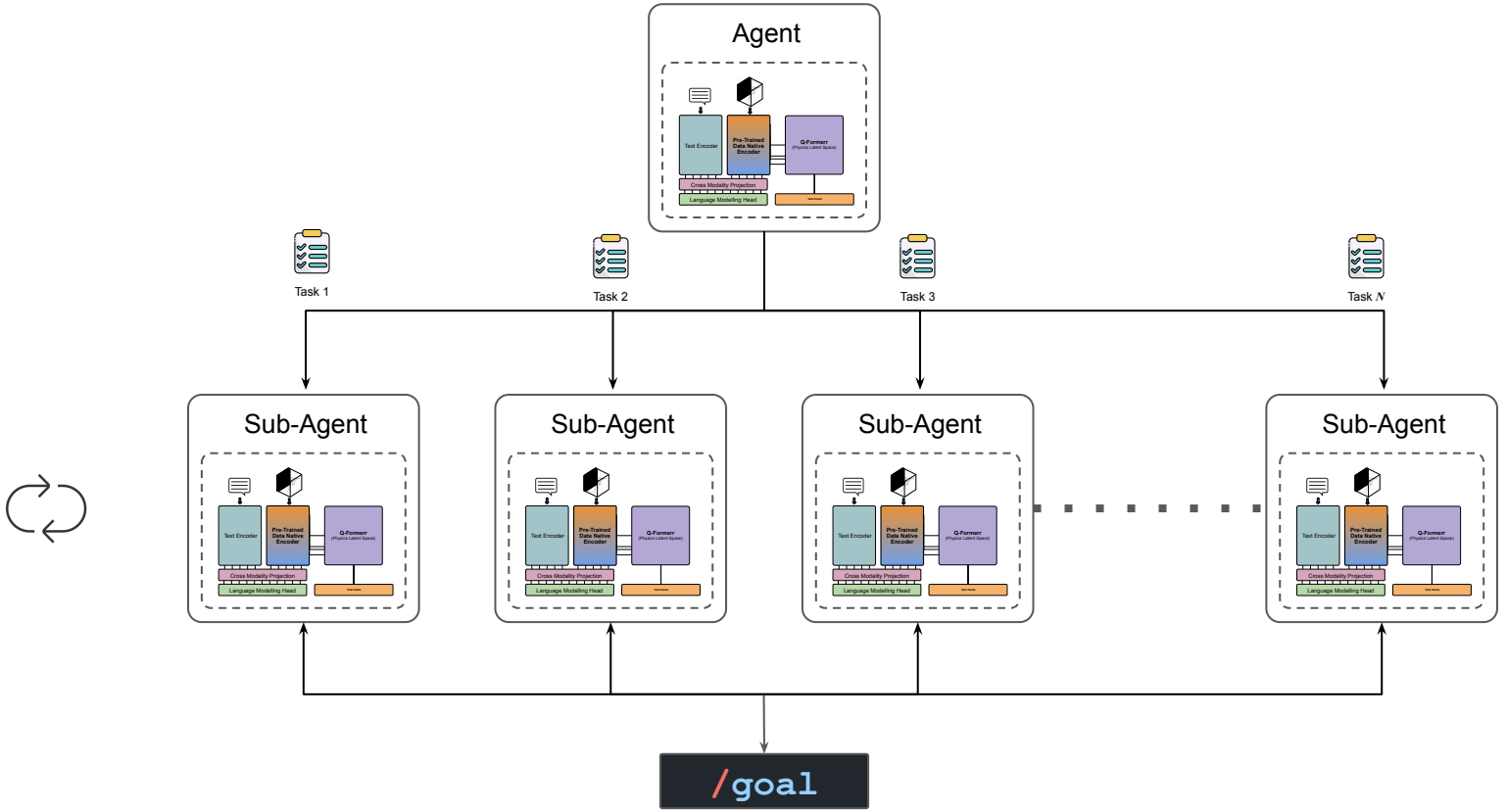
Why this Approach ?



The Agentic Loop



Multi-Agent Systems



Conclusions

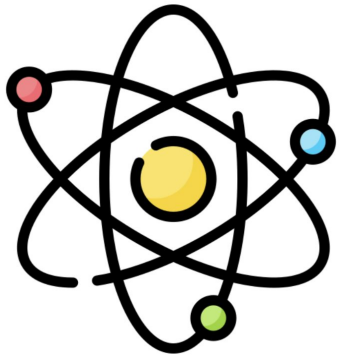
The Core Achievement: Demonstrated the feasibility of moving away from traditional task-specific ML pipelines toward multimodal foundation models by adapting LLaMA 3.2 Vision using QLoRA to successfully classify ν_e CC, ν_μ CC, and Neutral Current interactions from grayscale pixel maps.

Beyond Classification: Validated that Vision-Language Models can provide qualitative, human-in-the-loop explainability for high-energy physics experiments.

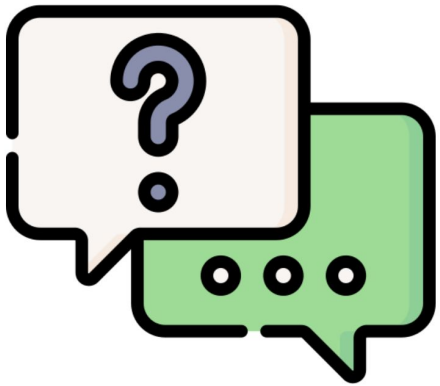
Addressing Current Limits: Mitigating failure modes such as ad-hoc explanations without expert validation

Easing into Data-Native Encoding Architectures: Strategizing a two-stage alignment strategy to handle distribution shifts and avoid attention collapse by using Q-Former based feature projections.

From Classification to Autonomy: The ultimate evolution of this project transitions the Vision-Language Model from a single-turn classifier into an autonomous agent executing iterative **Reasoning** → **Action** → **Observation** loops.



Thankyou.



Questions or Comments ?