Audio-Video Deepfake Detection

[Graduation Project Final Presentation]

Advisor: Prof. Jung Uk Kim

Student(Presenter): Jun Hee Lee

Date: 19, December 2025

Contents

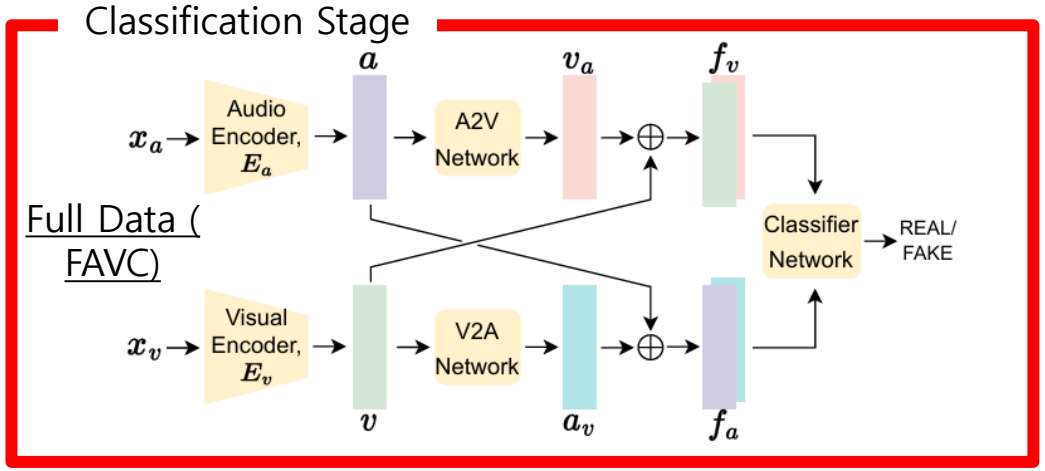
- Experiments
  - Audio Only
  - Audio + A2V (Entire)
- Wrap up
- Future Work & Research Roadmap

Date: 03, December 2025

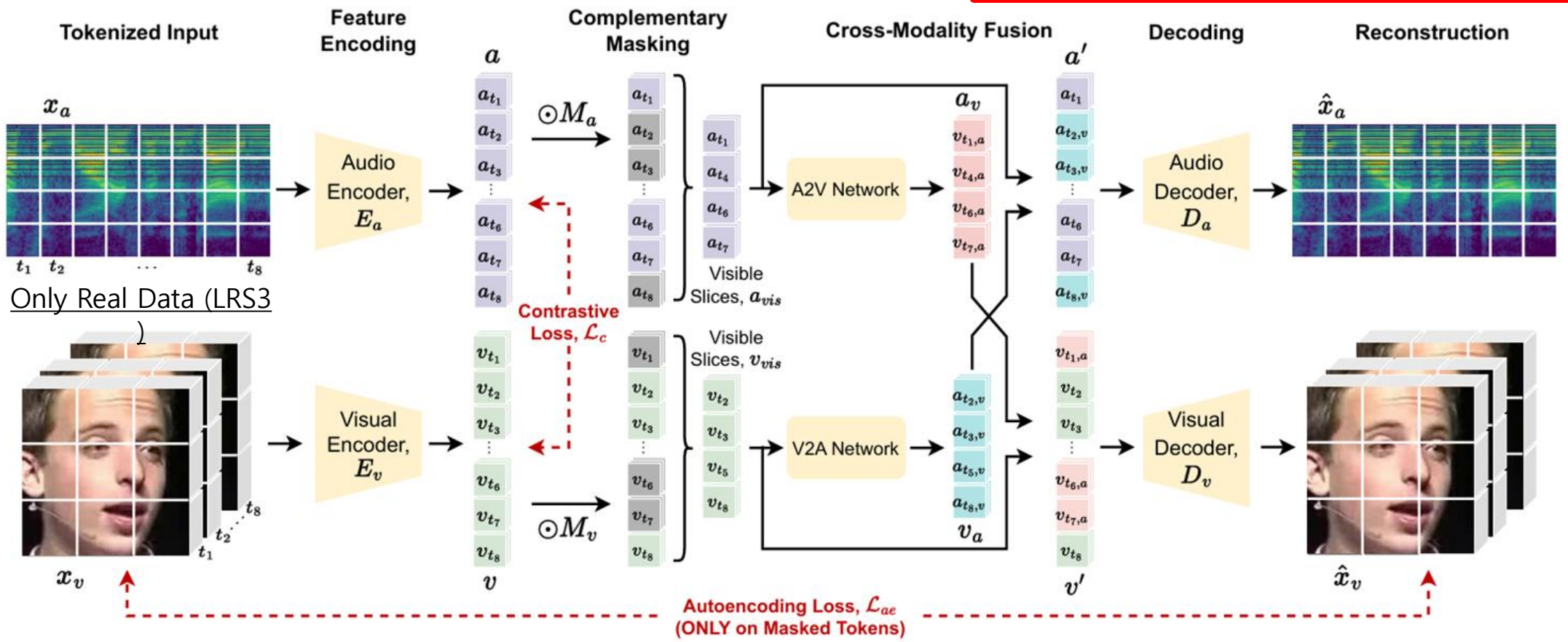
Recap: Experiment Settings

BaseLine Model : AVFF [CVPR'24]

Train Here →



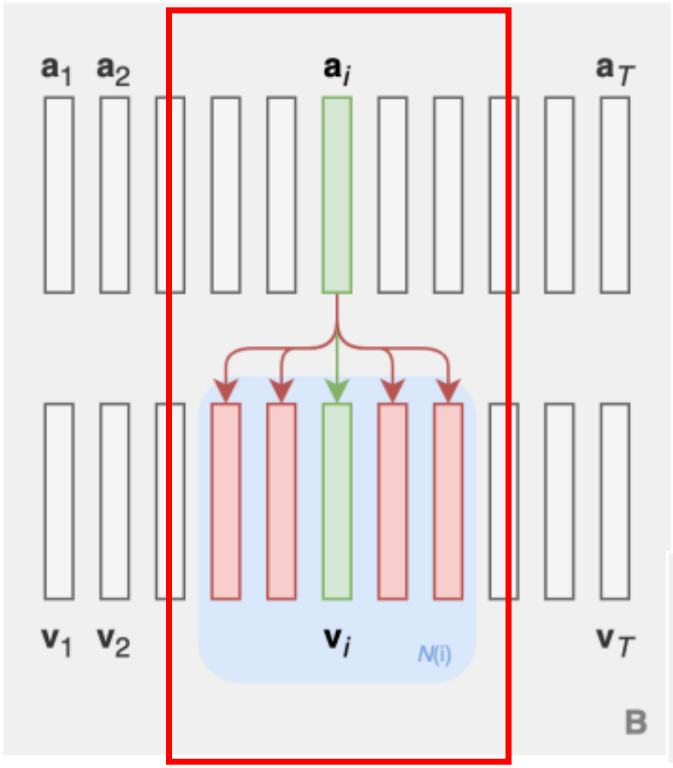
Audio-Visual Rep. Learning Stage

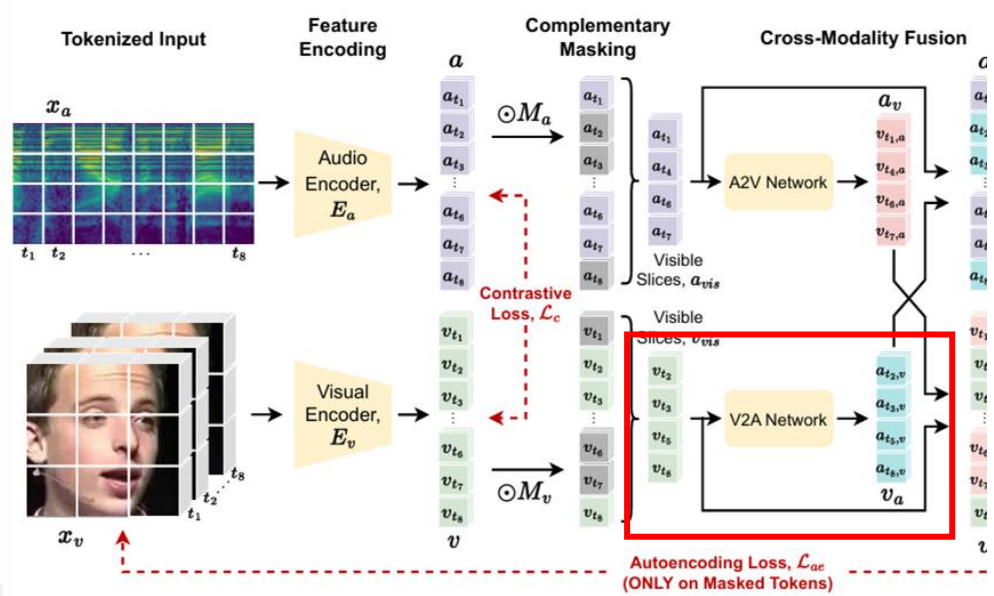


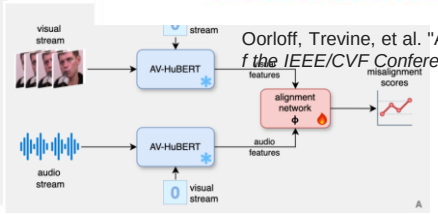
Oorloff, Trevine, et al. "Avff: Audio-visual feature fusion for video deepfake detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Date: 21, October 2025

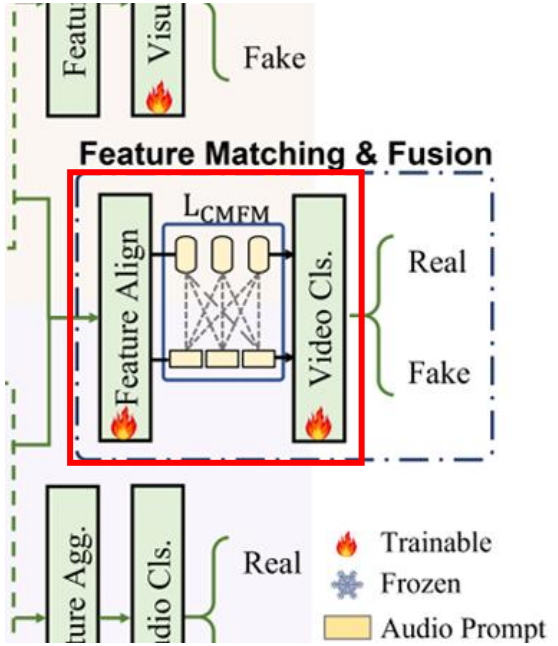
Ideation: Limits of Previous Methods







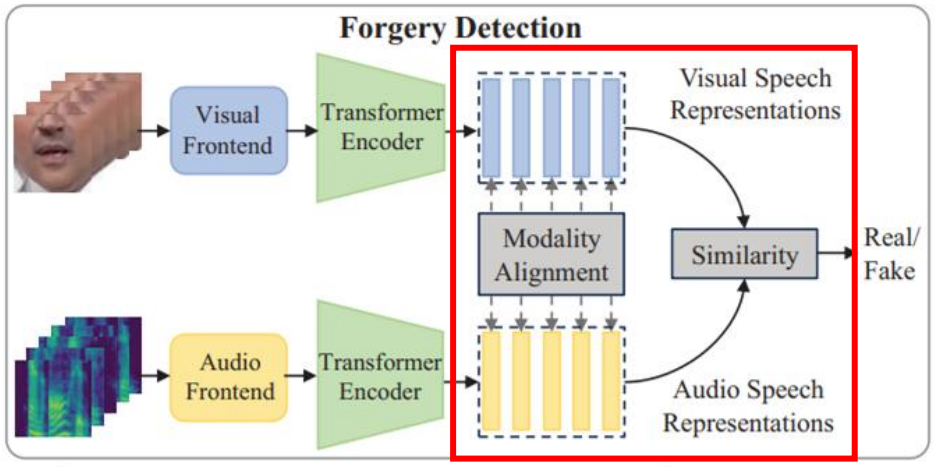
Oorloff, Trevine, et al. "Avff: Audio-visual feature fusion for video deepfake detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.



Miao, Hui, et al. "Multi-modal Deepfake Detection via Multi-task Audio-Visual Prompt Learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39 . No. 1. 2025.

Smeu, Stefan, et al. "Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

Ignoring <u>Audio–Visual Asymmetry</u>!



Liang, Yachao, et al. "SpeechForensics: Audio-visual speech representation learning for face forgery detection." *Advances in Neural Information Processing Systems* 37 (2024): 86124-86144.

Date: 04, November 2025

Ideation: Core Idea

What is '**Audio–Visual Asymmetry**' ?

- **Unidirectional Information Flow**

  Audio carries dense temporal cues (pitch, rhythm, articulation) that can reliably predict mouth motion, while visual frames cannot reconstruct audio with comparable precision.

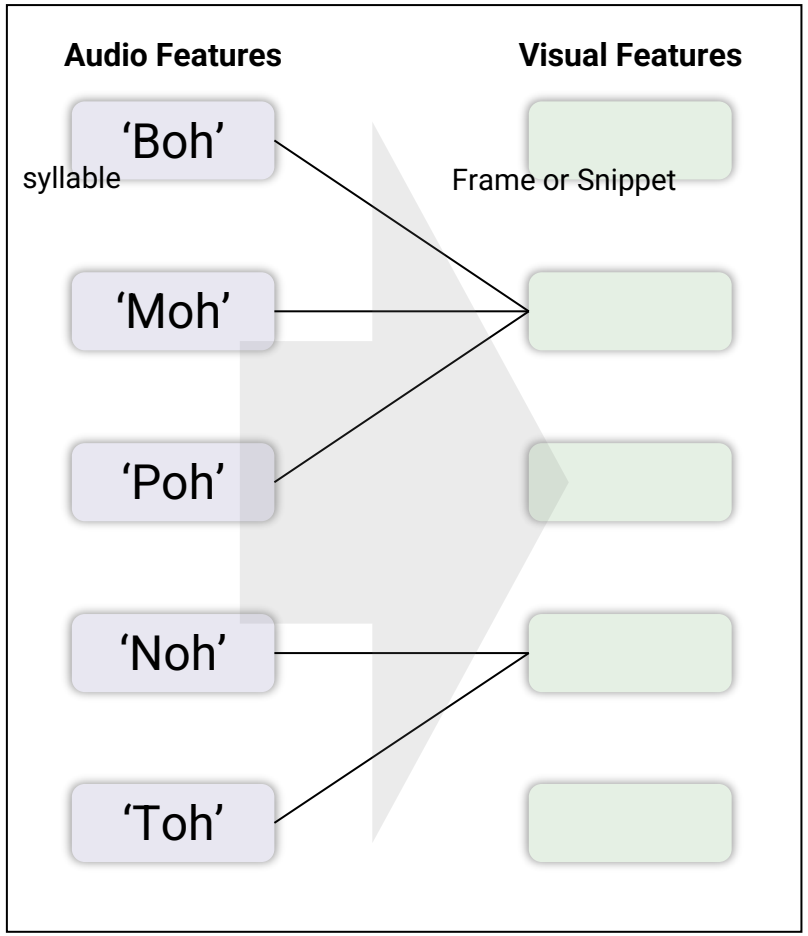
- **One-Way Mapping (Audio → Visual)**

  The relationship is inherently asymmetric — multiple valid audios may match a single lip pattern, but one audio typically implies a unique visual articulation.

- **Impact on Detection Frameworks**

  Prior works assume symmetric A↔V alignment, enforcing mutual sync that leads to unstable learning. Exploiting this asymmetry allows models to capture subtle inconsistencies where audio drives articulation but the generated video fails to follow naturally.

Directed Mapping, Not a Bijective Mapping



Date: 04, November 2025

Experimental Verification of the Limits

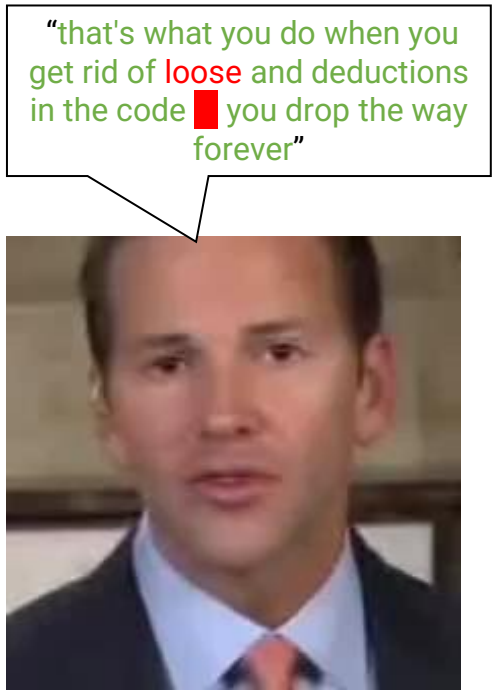
Evaluating the Model's Ability to **Infer Audio Information from Visual** Cues

- Conducted an experiment to examine whether the model has a weak ability to infer audio information fro m visual input.
- AV-HuBERT was used as the backbone, as it has been adopted in several recent audio-visual deepfake d etection studies (CVPR 2025, ICCV 2025).
- AV-HuBERT demonstrates strong multi-modal representation learning capabilities and has been propose d for downstream tasks such as Automatic Speech Recognition (ASR) and Visual Speech Recognition (V SR).
- Therefore, we evaluated how well the model performs lip reading (VSR) using the real subset of the Fake AVCeleb dataset.

Shi, Bowen, et al. "Learning audio-visual speech representation by masked multimodal cluster prediction." *arXiv preprint arXiv:2201.02184* (2022).

Date: 04, November 2025

Wrap up

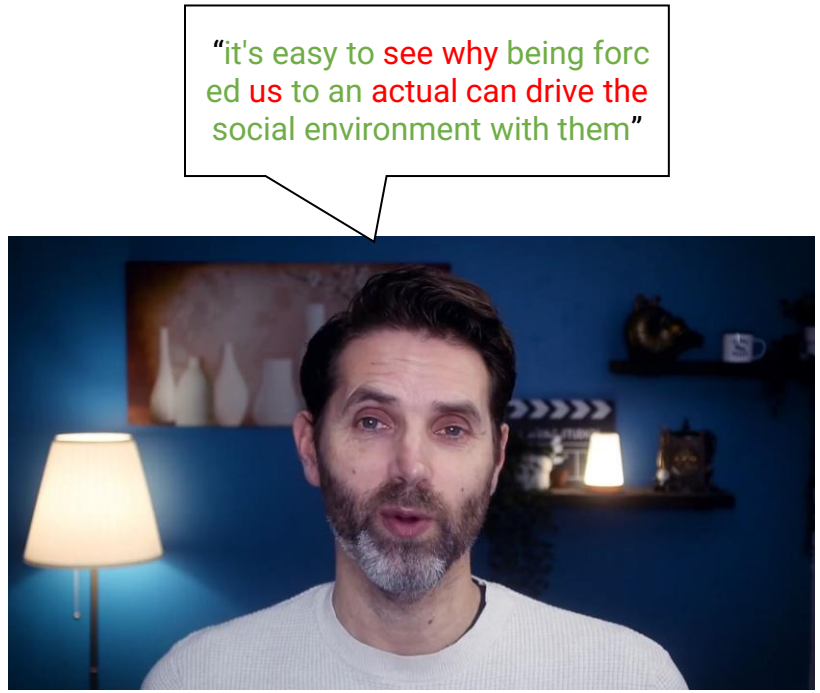


Seen Domain (VoxCeleb2)

Partially Success

Seen Domain (VoxCeleb2)

Fail



Uneen Domain (YouTube)

Fail

- The results indicate that **V2A cannot perform complete inference from visual information**.
- Since deepfake detection involves identifying subtle and fine-grained artifacts, **even a minor syllabic error may cause sub stantial distortion** of information.

Date: 03, December 2025

Recap: Research Direction

The Problem of Dataset (RVFA in FakeAVCeleb)

RVRA

① **Erroneous STT-generated sentences**

I went to the Australia for the commercial (Real, my listening)

I went to the Australia for the promotion (STT, in dataset)

Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." *arXiv preprint arXiv:2108.05080* (2021).

Date: 30, September 2025

Recap: Research Direction

The Problem of Dataset (RVFA in FakeAVCeleb)

RVRA

RVFA

② **Lip-sync naturally fails**

Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." *arXiv preprint arXiv:2108.05080* (2021).

Date: 30, September 2025

Recap: Research Direction

The Reason of the Problem

Fake Audio

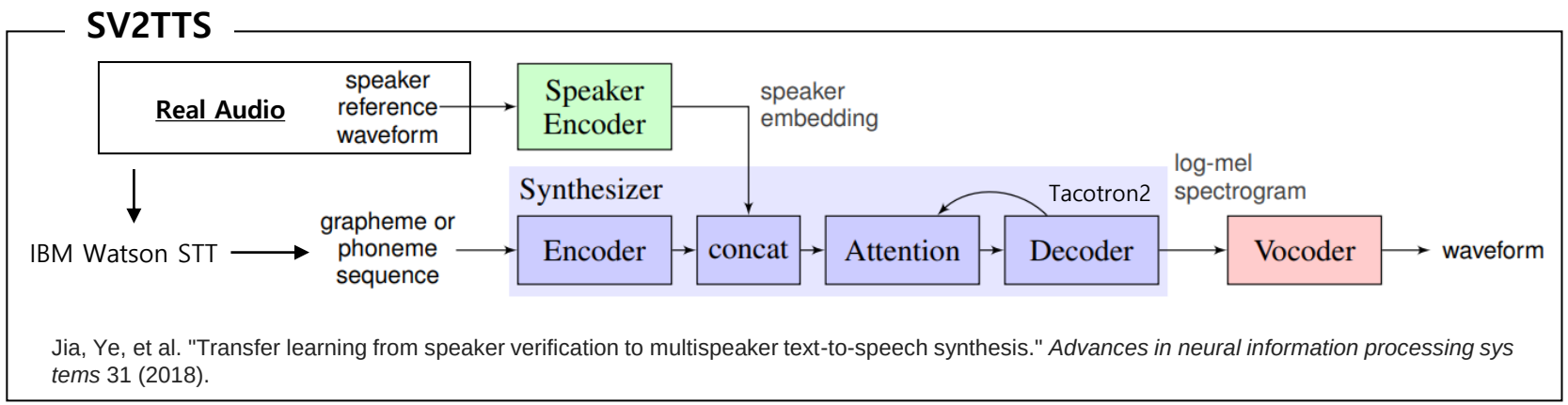
All sentences re-synthesized by the model; absent visual conditioning, **lip-sync naturally fails (∴ RVFA must be wrong)**

Else (FVRA, FVFA)

Replace only the lip region using Wav2Lip to match the given audio; swap into the original frames → generate a **natural lip-synced video**

The paper explicitly acknowledges out-of-sync cases within this category

Fake-Audio and Real-Video ($\mathcal{A}_{\mathcal{F}}\mathcal{V}_{\mathcal{R}}$). This deepfake type contains the cloned fake audio of a person along with the real video. We generate cloned fake audio using a transfer learning-based real-time voice cloning tool (SV2TTS), which takes real audio and text as an input, and outputs synthesized audio (with voice matching) of the same person, as shown in Table 2 ($\mathcal{A}_{\mathcal{F}}\mathcal{V}_{\mathcal{R}}$). Please refer to the top-right block in Figure 1 for some sample results, and Figure 2 for spectrograms from real and fake audio. Since we do not have the text spoken in the video, we used IBM Watson speech-to-text service [64] that converts audio into text. This text, along with the corresponding audio, is then passed to the SV2TTS. Later, we merge the synthesized audio with the original video, resulting in the $\mathcal{A}_{\mathcal{F}}\mathcal{V}_{\mathcal{R}}$ pair (see Figure 3). As a result, we were able to generate 500 fake videos. Note that our pipeline is not dependent on IBM Watson speech-to-text service; any speech-to-text service can be used as we are just extracting text from our audios. Since it is impossible to generate fake audio with the same timestamp as the original audio, this type of deepfake is not lip-synced. The possible use-case of this type of deepfake is a person performing identity fraud by impersonating a person or speaker recognition system. This dataset type can be also used to defend against anti-voice spoofing attacks, since we have real-fake audio pairs with similar text.



Jia, Ye, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." *Advances in neural information processing systems* 31 (2018).

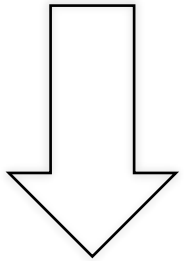
Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." *arXiv preprint arXiv:2108.05080* (2021).

Date: 30, September 2025

Recap: Research Direction

Solution(Research Direction) : Voice Conversion

- Even latest datasets, including FakeAVCeleb, synthesize audio via TTS
- Erroneous STT-generated sentences yield corrupted audio data
- Lip-sync must be enforced, and mis-reenactment yields low-quality, easily detectable data



VC instead of (STT → TTS)

If fake audio **preserves rhythm/timing while altering only pitch-related attributes**, it remains **closer to the original**

- Enabling RVFA
- Facilitating paired training for enhancing subtle traces of generative models

Date: 30, September 2025

Recap: Research Direction

Solution(Research Direction) : Voice Conversion

Example

RVRA (Source)

RVRA (Reference)

RVFA (Target)

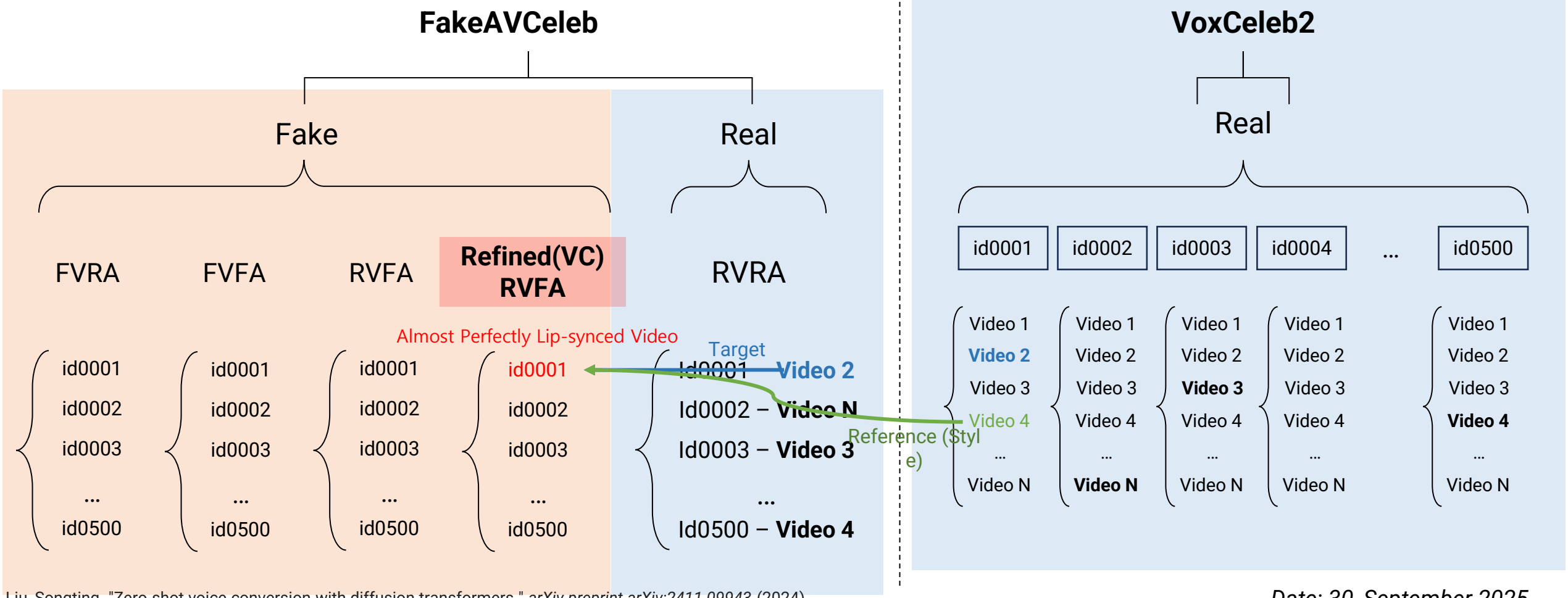
Audio–video sync verified to be well maintained

Liu, Songting. "Zero-shot voice conversion with diffusion transformers." *arXiv preprint arXiv:2411.09943* (2024), https://huggingface.co/spaces/Plachta/Seed-VC

Date: 30, September 2025

Current Progress

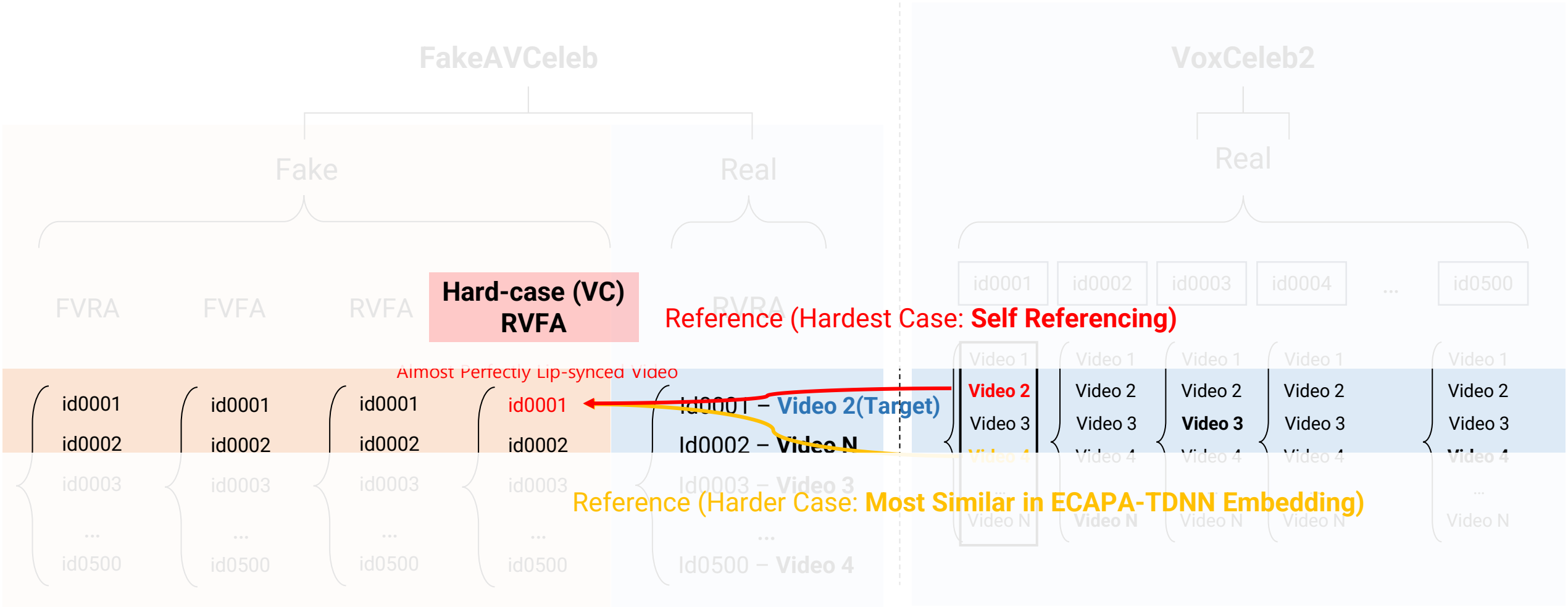
Refined Dataset



Liu, Songting. "Zero-shot voice conversion with diffusion transformers." *arXiv preprint arXiv:2411.09943* (2024)

Date: 30, September 2025

Current Progress: H-RVFA(Hard-case RVFA) Dataset



Desplanques, Brecht, Jenthe Thienpondt, and Kris Demuynck. "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification." arXiv preprint arXiv:2005.07143 (2020).

Date: 21, October 2025

Current Progress

Dataset Annotations & Settings

- **Identity-disjoint split**
  - *Each identity must appear in only one split* – either **train** or **test**, never both.
  - ***Race and gender distributions must remain balanced*** in both train and test sets.

- **Paired real–fake selection**
  - *Whenever a real sample (RVRA) of an identity is selected*, all required **fake versions of th e same identity** must also be included (e.g., RVFA, FVRA, FVFA as needed).

Date: 30, September 2025

Ideation for a Model Framework

OOD(Out of Distribution) in Audio Features

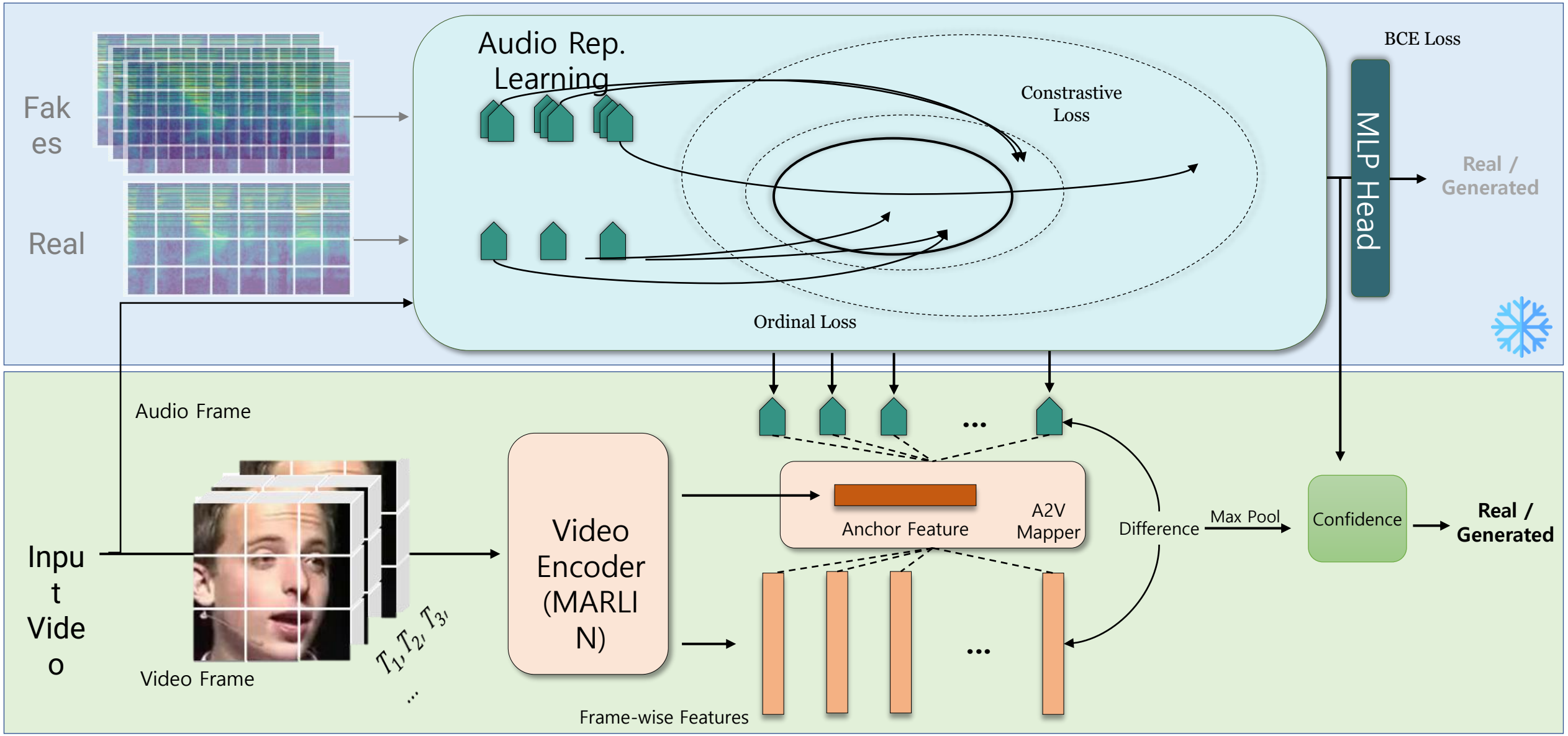
- Problem: Voice-converted audio sounds real but has subtle distributional shifts.
- Solution: Detect Out-of-Distribution (OOD) audio compared to real speaker distribution.
- How:
  - Use pretrained audio encoder (e.g., WavLM) → extract features.
  - Train multi-decoder autoencoder to model real voice patterns.
  - High reconstruction error = OOD (potentially fake).
  - **Input Paired Audio to enhance difference**

Output: Confidence score on whether audio is in-distribution.

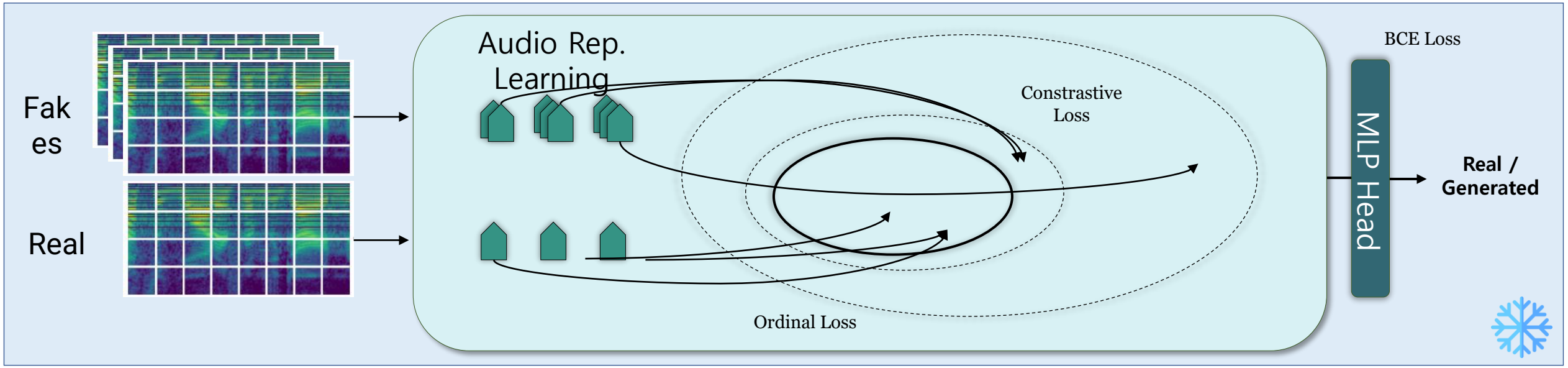
Advantage: Catches subtle fake patterns that humans can't hear.

Date: 28, October 2025

Overall Framework



Audio-Only



Setup: Trained with **RVFA** → Tested on **both RVFA and RVFA-VC**

	Baseline (AVFF)	Ours
Modality	Audio + Visual	Audio
Performance (AP)	87.56	85.32

Reasonable performance despite relying on a single modality

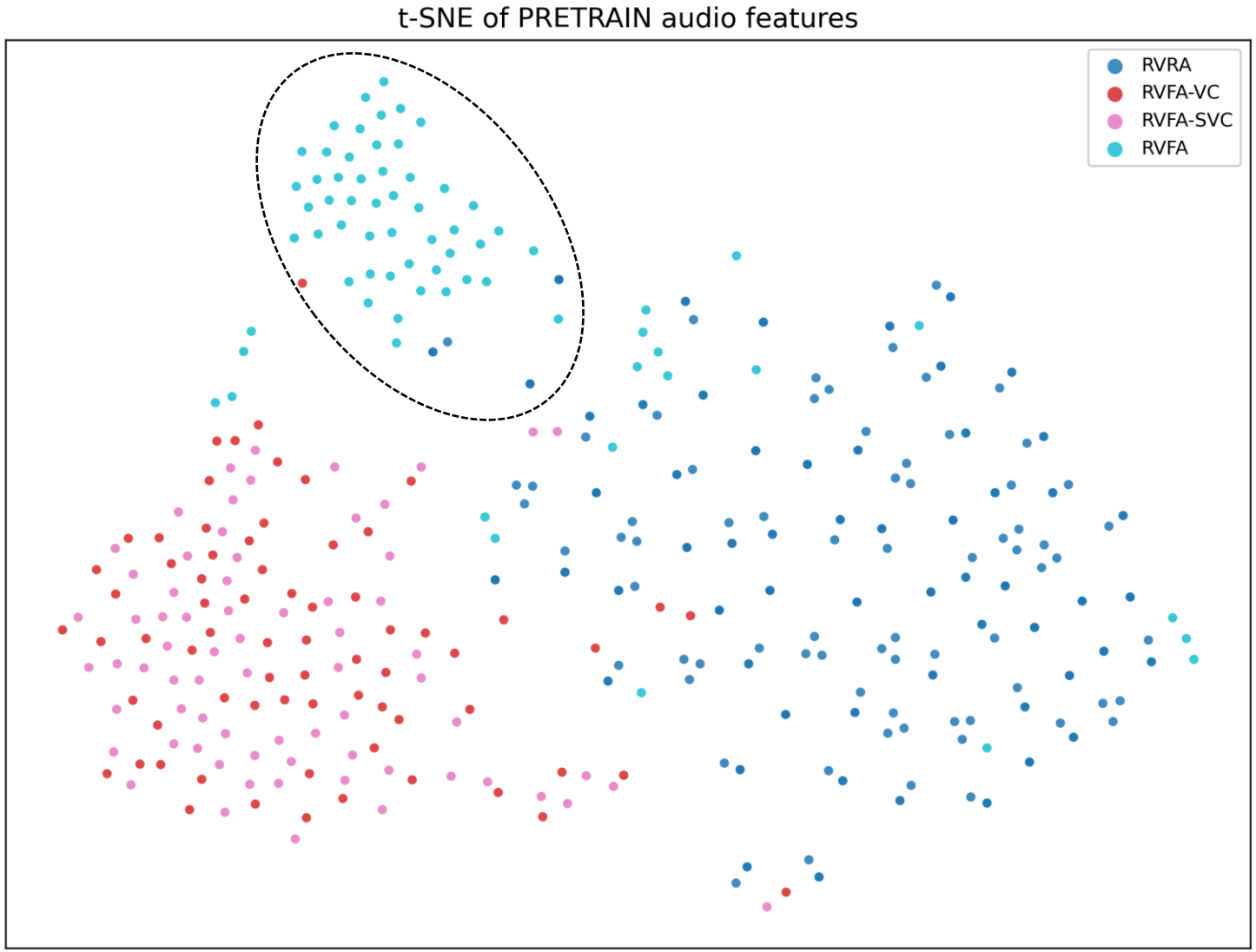
Focused evaluation on the discriminative capability of the Audio-only branch

Category Splits		Score	
Train	Test	AUC	AP
RVFA	RVFA	96.70	97.35
RVFA-VC	RVFA-VC	98.21	98.51
RVFA	RVFA-VC	75.10	75.64
RVFA-VC	RVFA	79.44	80.68
RVFA+RVFA-VC	RVFA	94.77	95.51
RVFA+RVFA-VC	RVFA-VC	96.38	96.84
RVFA+RVFA-VC	RVFA+RVFA-VC	94.72	95.37

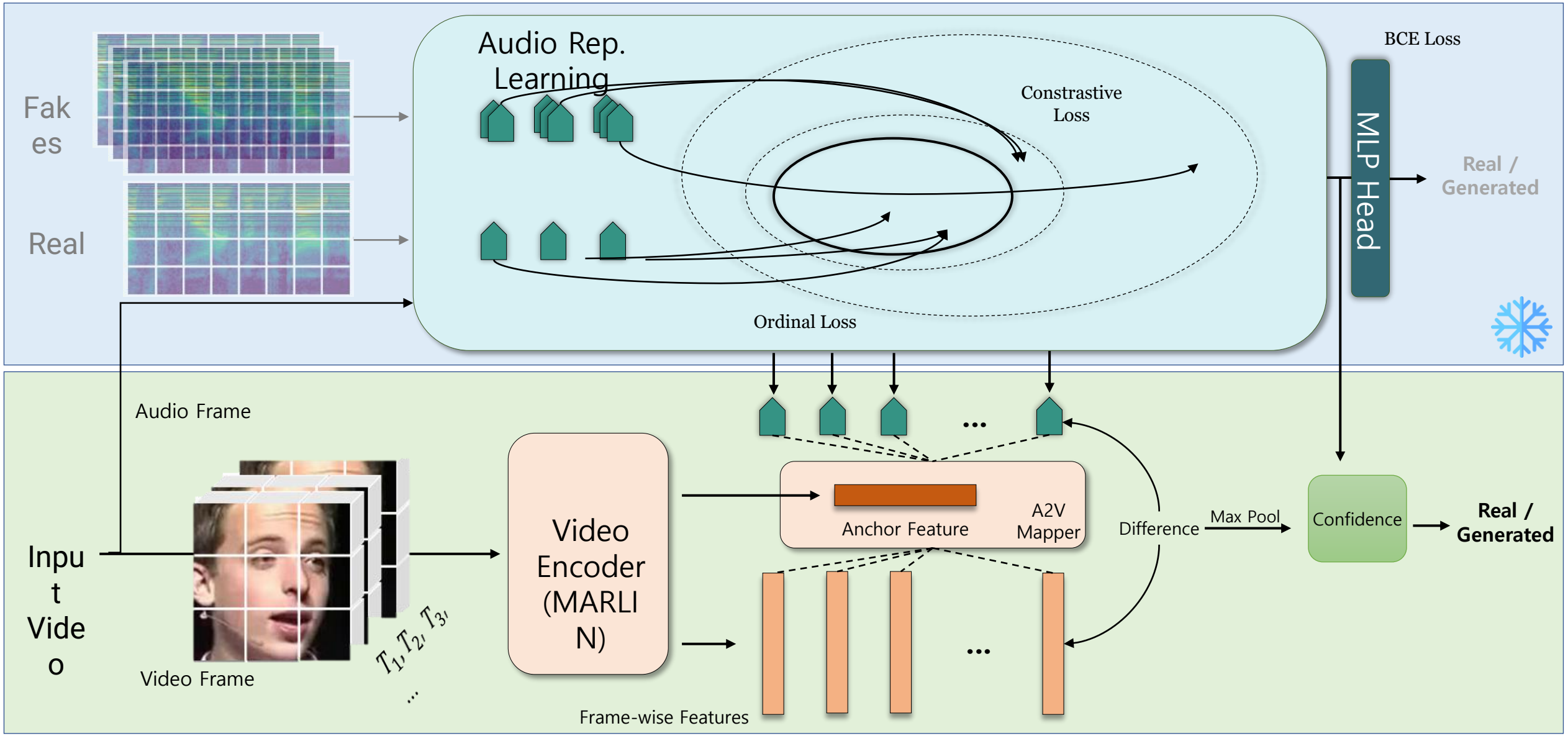
Audio-Only

- RVFA-VC / RVFA-SVC naturally cluster
- Clear separation into 3 groups

- "Is a gradient-style spread a better representation?"



Audio + A2V (Entire)



Date: 03, December 2025

Experiments: Audio + A2V (Entire)

Results

Setup: Trained with **RVFA** → Tested on **both RVFA and RVFA-VC**

	Baseline (AVFF)	Ours	Ours
Modality	Audio + Visual	Audio	Audio + Visual
Performance (AP)	87.56	85.32	91.96