上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

HoliAntiSpoof

Audio LLM for Holistic Speech Anti-Spoofing

Xuenan Xu
Shanghai Artificial Intelligence Laboratory

Outline

Outline

Background

- Advances in speech generation methods
  - Text-to-speech synthesis (voice cloning)
  - Voice conversion
  - Seamless speech editing
  - Waveform manipulation
  - ……
- Ethical risks are posed: generation / manipulation of sensitive information
- Speech anti-spoofing becomes critical

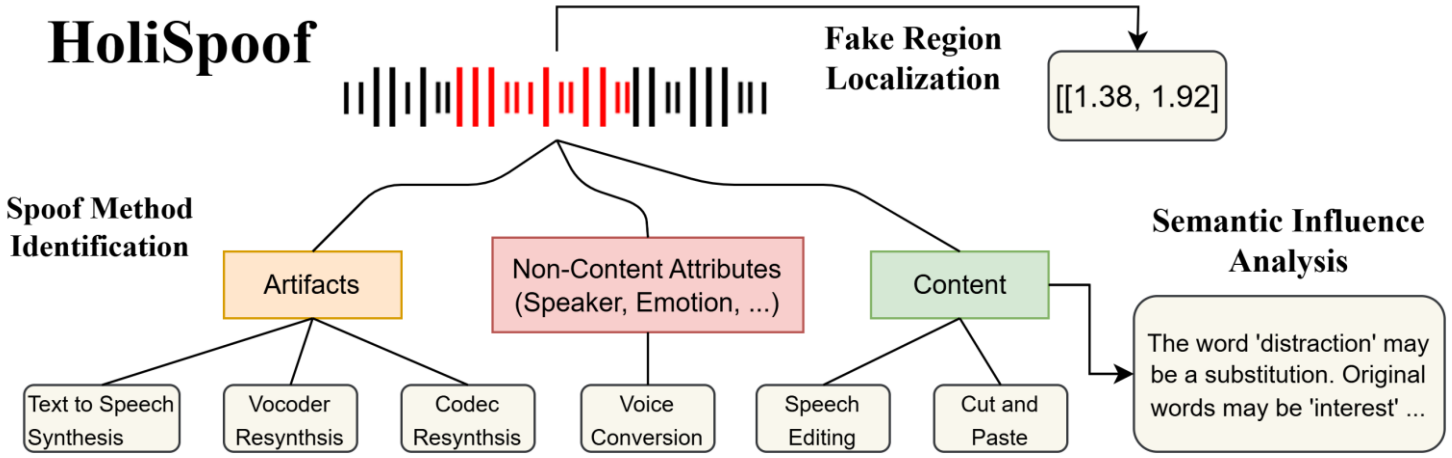
Motivation

- Conventional speech anti-spoofing methods
  - Real / fake binary classification paradigm
  - Front-end (mostly SSL) features + backend classifiers

- Repurposing LLM as the backend classifier: ALLM4ADD
- What is overlooked in binary classification:
  - Speech understanding from multiple levels
    - Low-level signal -> paralinguistic attributes -> linguistic content
  - Semantic influence
  - Interpretability: how the utterance is edited

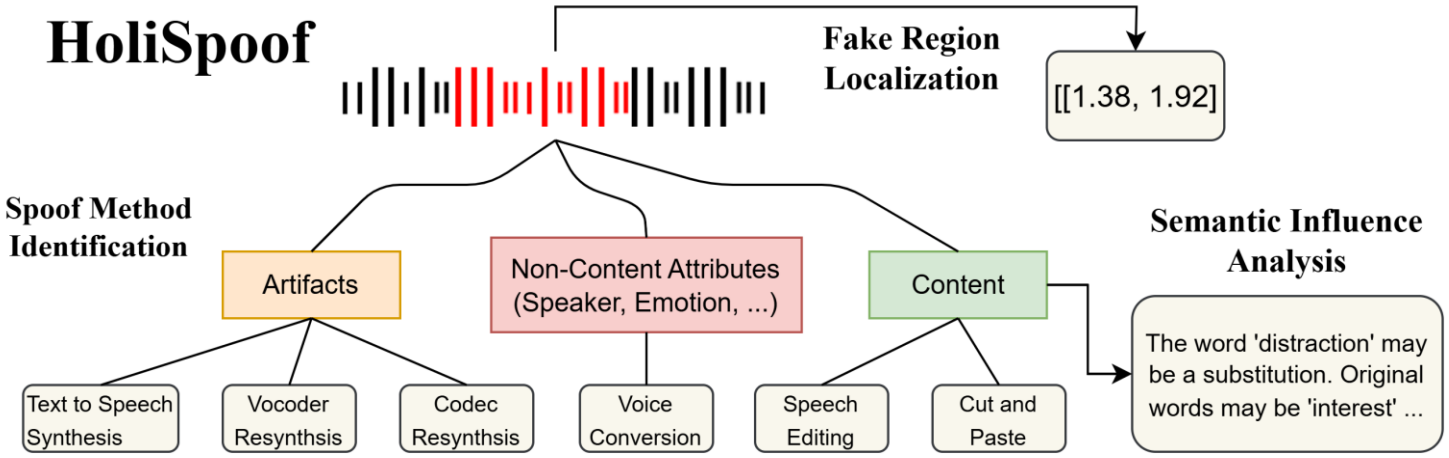
Motivation

- Multiple levels of speech spoofing utterances
  - TTS: introduce artifacts
  - VC: alter paralinguistics, unnatural expressions
  - Cut-and-Paste & seamless speech editing: artifacts + abnormal content



Motivation

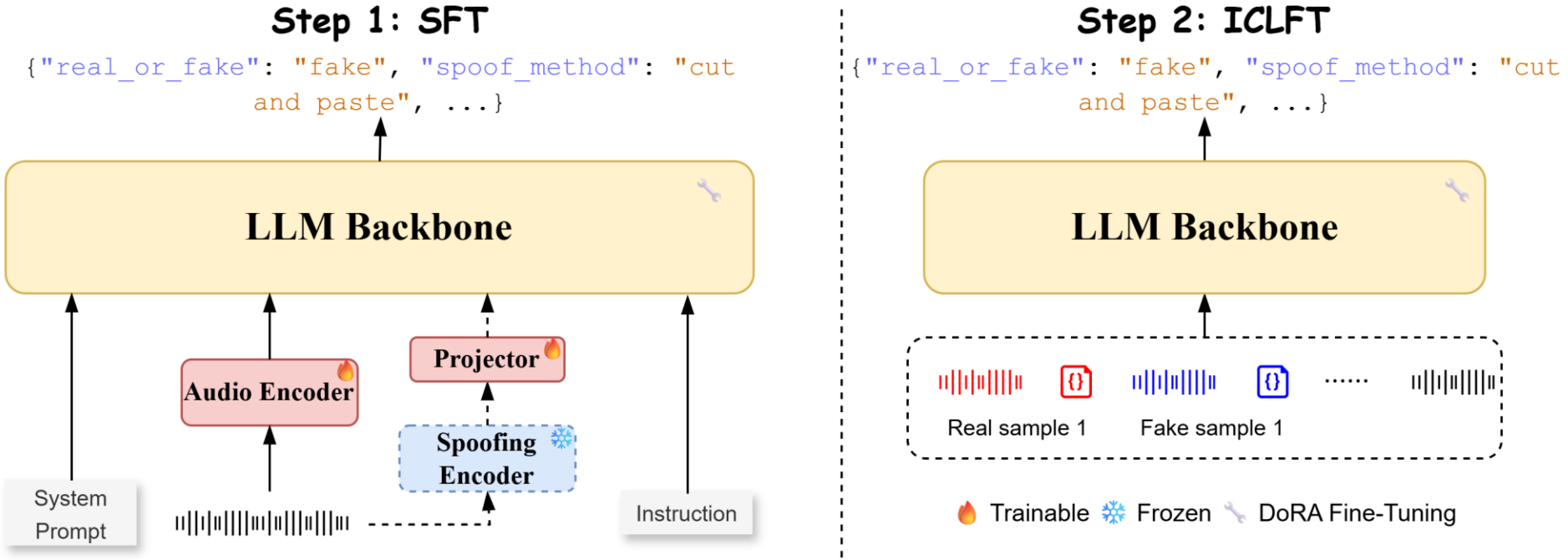
- LLM as a powerful holistic spoofing analyzer
  - Jointly detects ***spoofing method***, ***spoofing region***, and analyzes ***semantic impact***
  - Leverage in-context-learning (ICL) capabilities of LLM



Outline

Method

- HoliAntiSpoof framework
  - Unified text generation objective
  - Structured JSON output covering everything
  - (Optional) spoofing encoder to enhance audio features



Method

- HoliAntiSpoof framework
  - Qwen2.5-Omni backbone
  - Two training stages
    - Supervised Fine-Tuning (SFT): audio encoder fully trained + LLM DoRA, on all data

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, \mathcal{E}_a),$$

    - In-Context Learning Fine-Tuning (ICLFT): only LLM DoRA, on small-scale ICL data

$$\mathcal{L}_{\mathrm{ICLFT}} = -\sum_{t=1}^{T} \log p\left(y_t \,\middle|\, y_{<t}, \underbrace{(\mathcal{E}_{a_1}, y_{a_1}), \ldots, (\mathcal{E}_{a_K}, y_{a_K})}_{\text{in-context examples}}, \mathcal{E}_a \right).$$

Outline

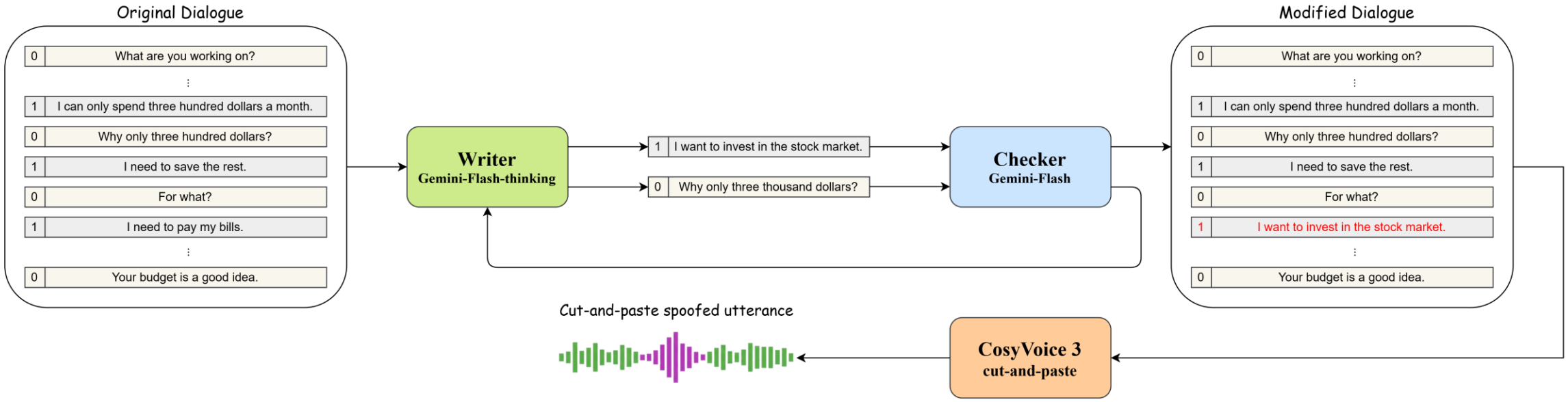
Data Construction

- Speech editing datasets
  - PartialEdit[1]
    - Modify one or two words in a sentence
    - Example: “Please *call* Stella” -> “Please *ignore* Stella”
    - Automatically annotate semantic influence of the spoofed word by GPT-5
  - Lack dialogue-level speech editing datasets

[1] Zhang, Y., Tian, B., Zhang, L., and Duan, Z. PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing. In Proc. Interspeech, pp. 5353–5357, 2025.

DailyTalkEdit: Dialogue-Level Spoofing Dataset

- Dual-agent text manipulation and synthesis
  - Base: DailyTalk[1], two-person conversation
  - Writer proposes utterance manipulation until Checker approves
  - Synthesized utterances replaces the original -> “Cut-and-Paste” spoofing

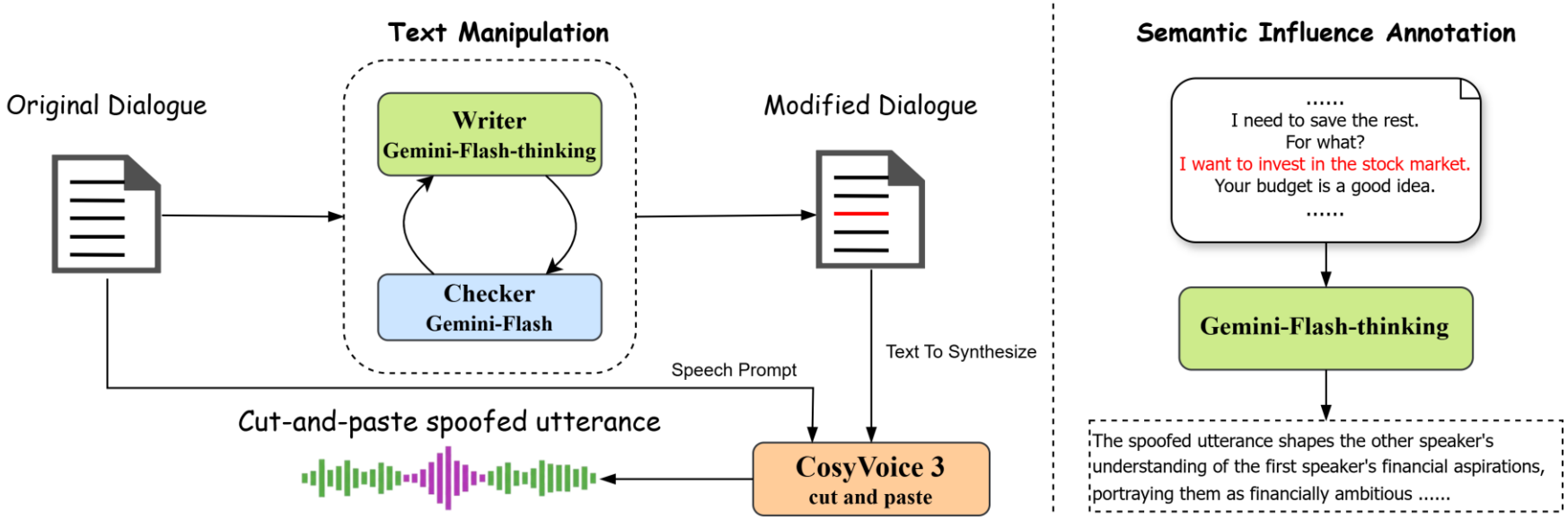


[1] Lee, K., Park, K., and Kim, D. DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech. In Proc. ICASSP, pp. 1–5, 2023.

DailyTalkEdit: Dialogue-Level Spoofing Dataset

- Semantic Influence Annotation
  - Only the original dialogue and the spoofed part are given
  - LLM considers potential original sentences and gives a concise textual analysis

- Result: 2,541 DailyTalk dialogues -> 2,475 spoofed DailyTalkEdit



Outline

Experimental Setup

- Training datasets
  - Mixture of existing English anti-spoof datasets + DailyTalkEdit
  - 900K samples in total
  - ICLFT: 6 reference samples
- Evaluation
- Baseline

Experimental Setup

- Training datasets

Dataset	Spoof Method	# Samples
ASVSpoof2019 (ASV19.) (Nautsch et al., 2021)	TTS, VC	25,380
PartialSpoof (Zhang et al., 2022)	CaP	25,380
SINE (Huang et al., 2024)	CaP, SE	104,942
WaveFake (Frank & Schönherr, 2021)	VR	71,883
CodecFake (Wu et al., 2024)	CR	17,267
PartialEdit (Zhang et al., 2025)	CaP, SE	63,204
VCTK (Veaux et al., 2013)	/	7,730
LJSpeech (Ito & Johnson, 2017)	/	10,269
EmoFake (Yan et al., 2024)	VC	27,300
SpeechFake (Huang et al., 2025)-en	TTS, VC	458,245
DailyTalkEdit	CaP	3,600
FakeOrReal (Reimao & Tzerpos, 2019)	Unknown	53,868
InTheWild (Müller et al., 2022)	Unknown	31,779

Experimental Setup

- Training datasets

- Evaluation
  - In-domain: standard ASVSpoof2019 LA + mixed test set
  - Out-of-domain: SpeechFake-MD, SpoofCeleb, HAD
  - Metrics: accuracy & EER (authenticity), F1 (methodology), Seg-F1 (localization), LLM-as-judge (semantic)

- Baseline: Conventional + ALLM4ADD + Closed-source LLM (Gemini)

Results

- Comprehensive spoofing analysis performance
  - In-domain: outperforms conventional baselines
  - Out-of-domain: spoofing method is the dominant factor
  - General LLM (Gemini-3) cannot detect spoofing reliably

Models	In Domain				Out Domain			
	ASV19.	Mixed			SF-MD.	SpoofCeleb	HAD	
	Auth.	Auth.	Meth.	Loc.	Auth.	Auth.	Auth.	Loc.
Conventional Models								
RawNet2	77.69	91.24	90.16	53.57	88.42	68.06	5.56	37.84
RawGAT-ST	96.35	90.42	82.50	-	93.69	85.70	31.86	-
AASIST	96.29	94.29	**96.44**	-	92.11	75.91	10.49	-
Wav2Vec2 (F) - RT	79.08	70.68	65.86	41.10	69.46	41.46	47.65	37.12
Wav2Vec2 (T) - RT	79.48	91.49	90.44	53.41	91.78	41.09	22.31	<u>42.03</u>
HuBERT (F) - RT	94.53	88.28	81.74	55.15	88.39	72.40	39.29	40.73
HuBERT (T) - RT	85.32	91.14	90.99	<u>56.37</u>	89.30	61.63	25.80	37.36
WavLM (F) - RT	94.72	86.29	78.23	44.19	91.05	58.06	29.56	38.19
WavLM (T) - RT	74.56	88.77	90.96	54.90	91.49	63.89	14.31	**42.96**
ALLM Methods								
Gemini-3-Flash	-	63.39	16.10	33.26	69.49	60.73	78.59	29.55
HoliAntiSpoof	**96.59**	**96.16**	<u>95.12</u>	**91.33**	**97.89**	**90.36**	**90.19**	41.27

Results

- Comprehensive spoofing analysis performance
  - EER shows the same trend as accuracy

Models	In Domain		Out Domain	
	ASV19.	Mixed	SF-MD.	SpoofCeleb
Conventional Models				
RawNet2	6.58	8.67	6.36	31.21
RawGAT-ST	2.86	6.65	6.08	21.20
AASIST	2.65	4.60	7.41	16.80
Wav2Vec2 (F) - RT	13.51	24.89	19.57	44.19
Wav2Vec2 (T) - RT	13.96	8.04	9.79	34.00
HuBERT (F) - RT	5.56	10.61	7.71	28.00
HuBERT (T) - RT	12.48	8.70	12.43	27.00
WavLM (F) - RT	5.97	12.70	6.95	32.59
WavLM (T) - RT	14.47	9.62	5.77	25.40
ALLM Methods				
HoliAntiSpoof-Auth.-Only	**1.10**	**3.06**	0.68	10.03
HoliAntiSpoof	1.11	3.48	**0.39**	**8.52**

Results

- Comprehensive spoofing analysis performance
  - Unified target augments output without drop in authenticity detection
  - ICLFT enables few-shot OOD adaptation
  - Spoofing features are complementary

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Results

- Semantic Influence Analysis
    - 3 evaluation aspects: fluency + correctness + plausibility
    - Increasing trainable parameter number and ICLFT improves semantic influence analysis
    - Difference between model variants is subtle

[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

Outline

Conclusion

- First ALLM framework for holistic speech anti-spoofing
- Semantic influence reasoning is proposed into anti-spoofing, with DailyTalkEdit as the supporting dataset
- HoliAntiSpoof substantially outperforms conventional baselines
- Preliminary exploration of ICL shows its potential for zero-shot adaptation

Future Directions

- Extension to multilingual and more diverse spoofing scenarios
- More realistic dialogue-level spoofing scenarios
- Unlock full ICL potential

Q & A

Thank you!