

Beyond Co-existence:

Measuring Attribute Binding Hallucinations in Audio-Language Models

Mingi Kim, Minchol Kwon, Minuk Ma, Trung X. Pham, Junyeong Kim

Chung-Ang University · University of British Columbia · Van Lang University



Index

1. Related Work	VLM vs ALM Hallucination
2. Motivation	The evaluation gap & limits
3. Method	Synthetic dataset + ABEF Framework
4. Results	Binding is the dominant failure
5. Significance	Why models fail
6. Validation	Human alignment & metric comparison
7. Conclusion & Takeaways	From perceptions to relationships

Related Work

Vision-Language Models

- Object hallucination metric (CHAIR)
- Discriminative Y/N benchmarks (POPE)
- Atomic-fact frameworks: FaithScore, AMBER
- Rely on spatial visual grounding

Audio-Language Models

- Mostly ASR / captioning quality
- Focus on simple existence of sounds
- Heavy reliance on discriminative Y/N
- Audio is non-spatial, temporal, overlapping

Gap: VLM metrics don't transfer — no prior work quantifies attribute binding in ALMs.

Motivation: The Evaluation Gap

Audio: a man crying + a woman laughing



Discriminative prompt

"Is there a man crying and a woman laughing?"

✓ "Yes" → Passes (apparent accuracy)

Open-ended prompt

"Describe this sound briefly."

"...a man **laughing** and a woman **crying**..."

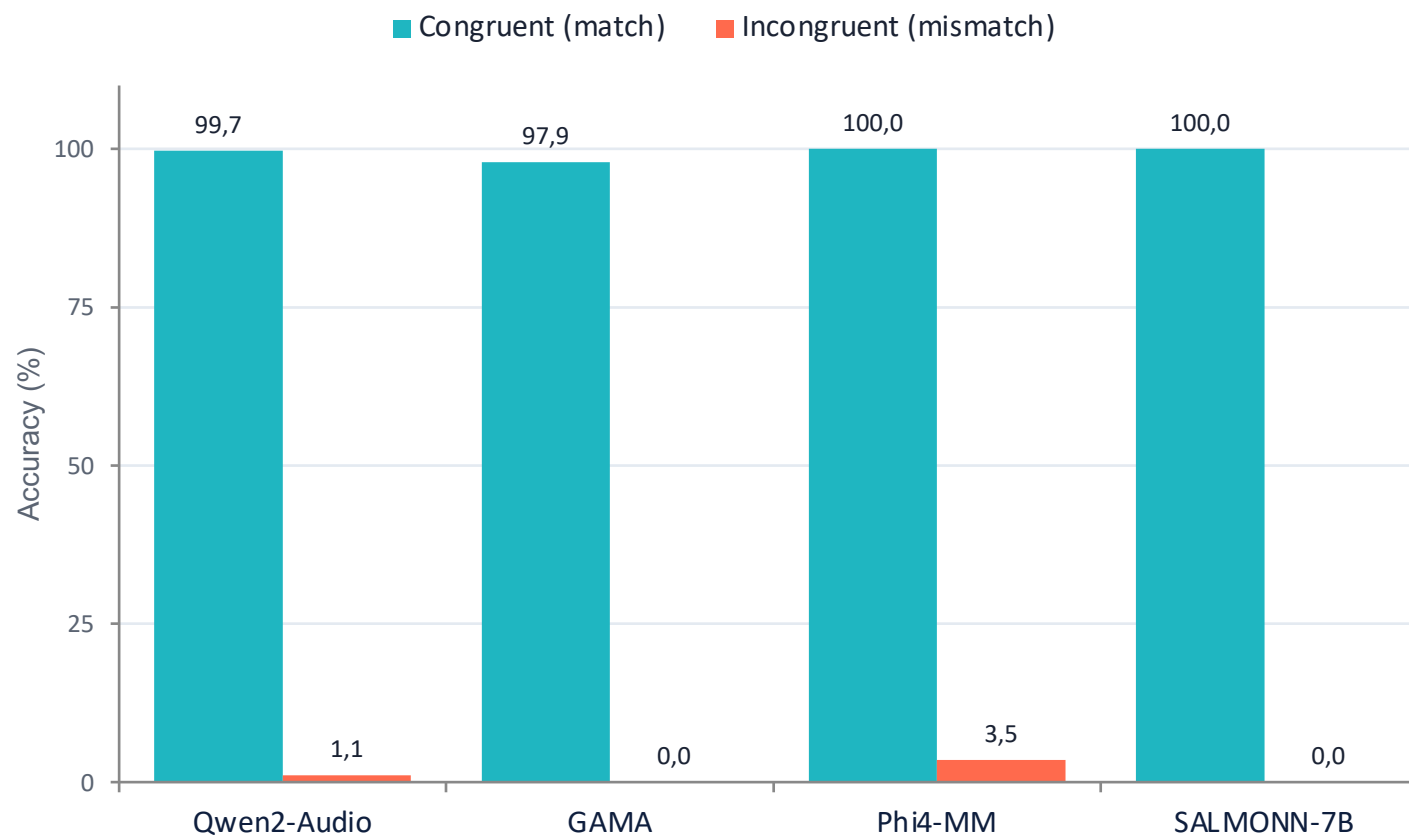
✗ Fails — attributes bound to wrong source

Attribute Binding Hallucination

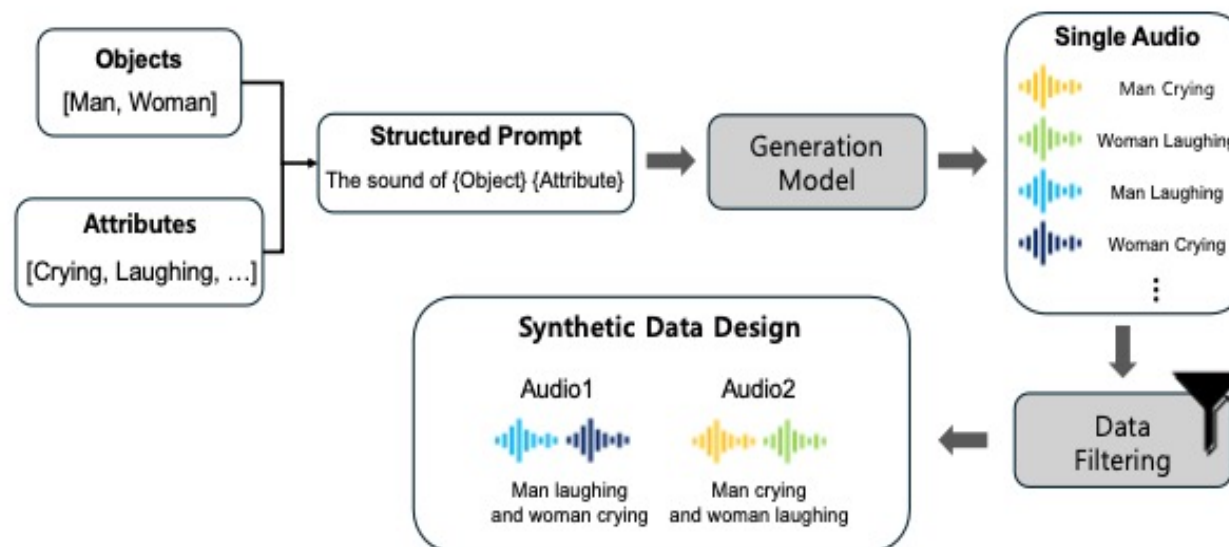
correct components detected, but incorrectly paired.

Motivation: The Evaluation Gap

Quantifying the Gap (With synthetic Dataset)



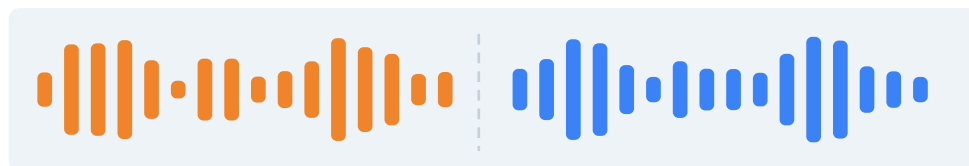
Method: Synthetic Dataset



- Source Data: FSD50K
- Generation Model: Stable Audio Open
- Filtering: PAM scoring (≈ 0.85 vs CompA)

Method: Synthetic Dataset

Synthetic (Ours)



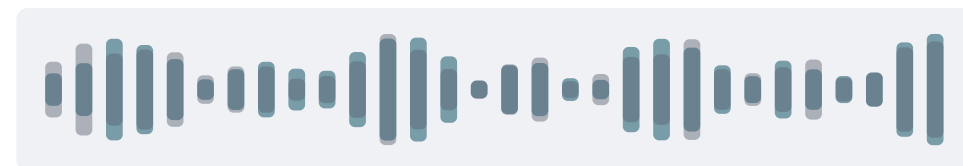
Man crying

Woman laughing

GT = {(man, crying), (woman, laughing)}

- Two single-source events, concatenated (non-overlapping)
- Exact ground truth
— any pairing error is a real binding failure
- Stable Audio Open, PAM-filtered for fidelity

CompA-Attribute

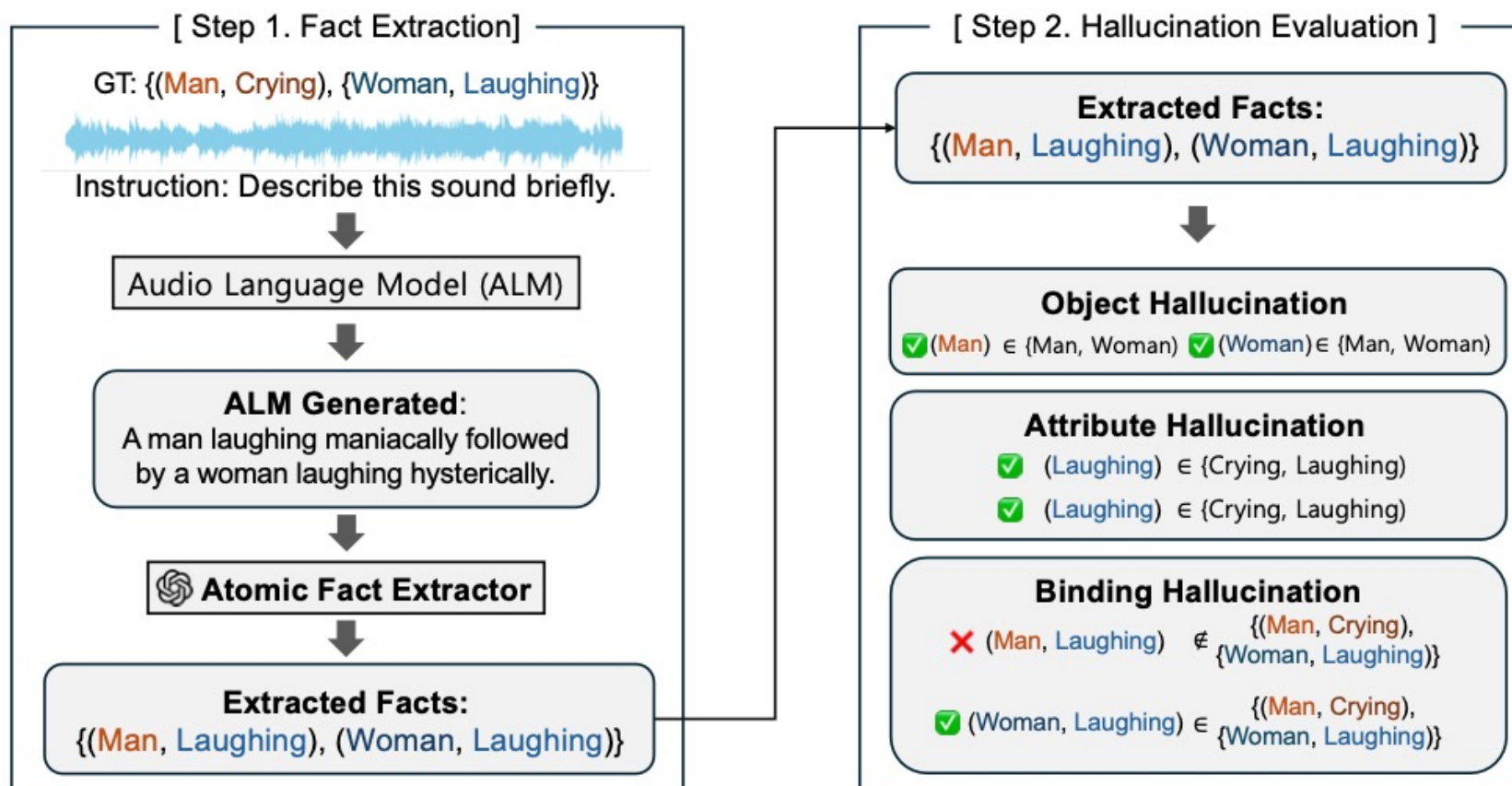


Overlapping real recordings

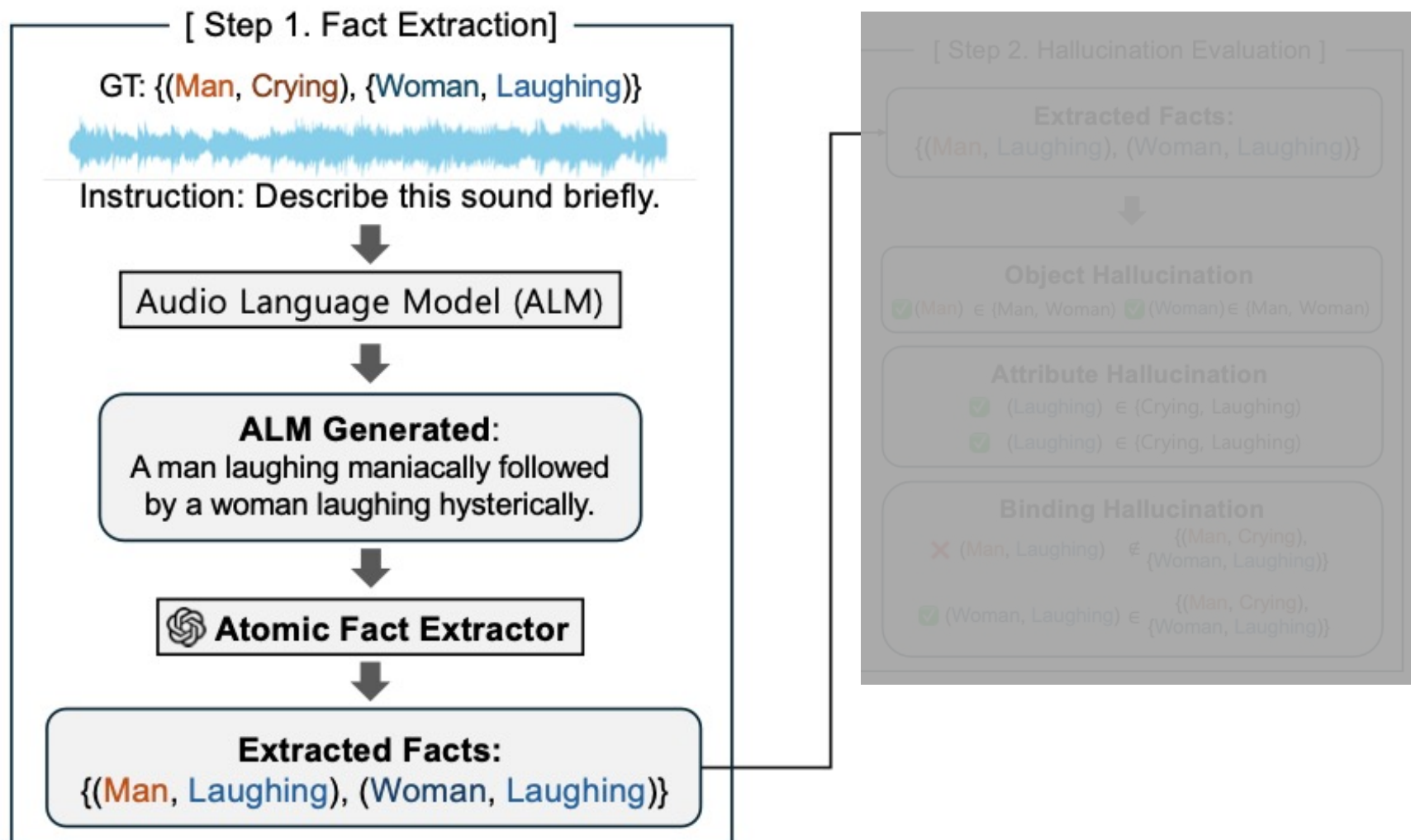
Same events, different compositions — matched to captions.

- Expert-annotated real-world audio
- Complex, overlapping acoustic scenes
- Noisier audio; ground truth can be ambiguous

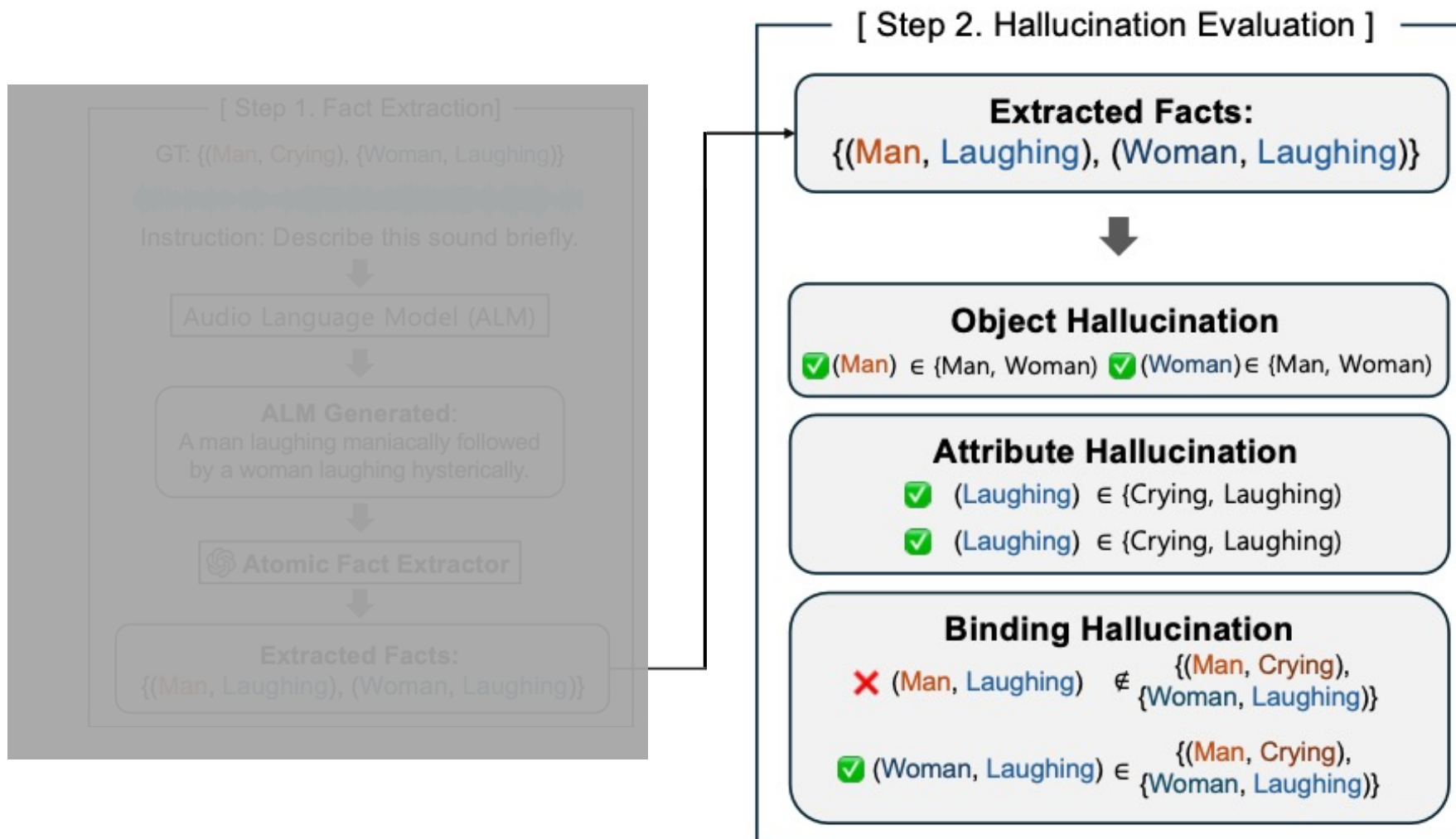
Method: Evaluation Framework



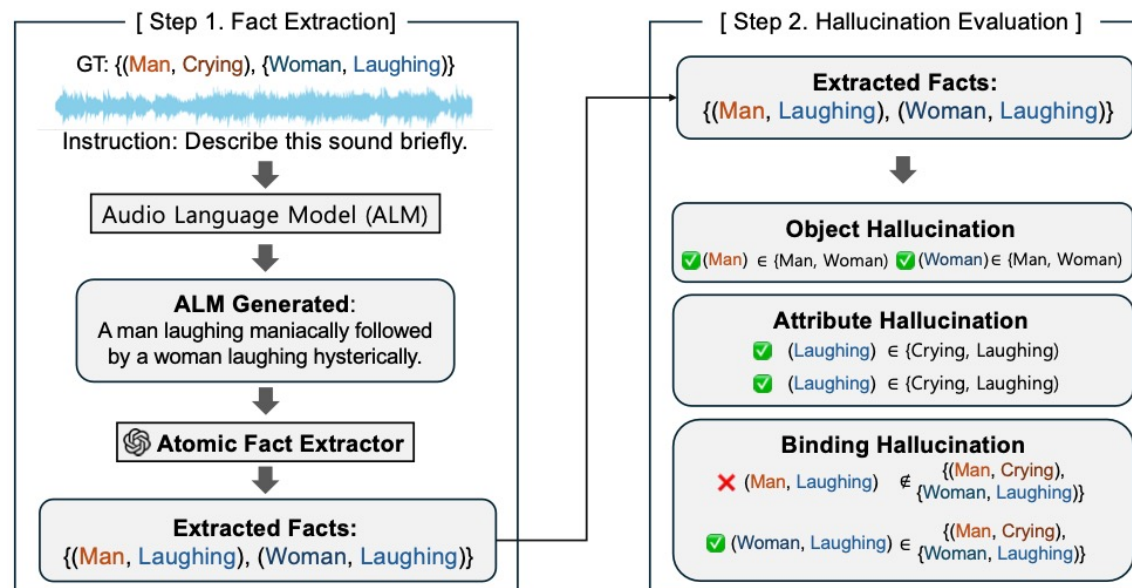
Method: Evaluation Framework



Method: Evaluation Framework



Method: Evaluation Framework



H_{obj}

Object hallucination

$obj \notin GT \text{ objects}$

H_{attr}

Attribute hallucination

$attr \notin GT \text{ attributes}$

H_{bind}

Binding hallucination

$(obj, attr) \notin GT \text{ pairs}$

Main Results: Open-ended Generation

74–97%

Binding hallucination is the dominant failure

Consistent across every model
persisting in both Synthetic and CompA Dataset

Model	Acc ↑	H_obj ↓	H_attr ↓	H_bind ↓
Qwen2-Audio	32.9	93.9	40.3	95.9
GAMA	36.0	88.9	39.1	92.0
Phi4-MM	16.8	92.5	74.0	97.6
SALMONN-7B	64.4	46.2	25.0	74.2

Synthetic dataset (N=376), scores in %. Best-performing SALMONN-7B still hallucinates binding 74% of the time.

Validation

$$\rho > 0.88$$

Spearman correlation with human judgment

Robustness

- Bootstrap: binding rate stable at 74.2–75.3%
- Cross-LLM (GPT-4 vs GPT-4.1): 92.75% consistency
- High inter-rater reliability (Krippendorff's α)

Conventional metrics miss binding errors

Metric	BLEU-4	METEOR	CIDEr	CLAP	ABEF (ours)
Human corr (ρ)	0.19	0.31	0.28	0.12	0.88

n-gram & embedding metrics reward lexical overlap even when attributes are wrong.

Conclusion

1 A hidden failure mode

Attribute Binding Hallucination is pervasive (74–97%) and invisible to discriminative evaluation.

2 A new standard

ABEF + a controlled synthetic dataset measurement to verify logical relationships.

3 A path forward

Results call for binding-aware architectures for true audio-grounded reasoning.

Takeaways

1

Existence $\neq$ understanding

Passing 'is X present?' says nothing about correct attribute binding.

2

Binding dominates

It is the #1 error type:
far above object or attribute hallucination.

3

Fix the attention

Text-prior bias in audio grounding

Thank you!

Question & Contact: mingi8233@cau.ac.kr