1. The Tools Used (Models)

- **VideoMAE:** video encoder.
- **AST (Audio Spectrogram Transformer):**audio encoder
- **GIT:** fused model

2. Measuring Success (Metrics)

Metric	**What it is**	**How it was measured**	**Interpretation & Ranges**	**Meaning for your Video**
Total Effect (TE)	The overall impact the sound had on the final caption.	Comparing the model's confidence with sound versus without sound.	**Range:** -1 to 1. **Positive** means the sound helped; **Negative** means it caused a conflict.	**0.1052**: The sound of glass breaking correctly helped the model describe the action.
Natural Direct Effect (NDE)	The direct influence of sound on the words used.	Adding sound while keeping the video processing "fixed" to see the direct change.	**Range:** -1 to 1. If it's close to the TE, the sound had a very direct impact.	**0.1052**: All of the sound's influence went directly into the caption words.
Natural Indirect Effect (NIE)	How sound changed the way the model "saw" the objects.	Measuring if the sound caused the visual tool to pay more attention to certain areas.	**Range:** -1 to 1. A value of **0** means the sound did not change visual perception.	**0.0000**: The sound added new information but didn't change what the model saw.
Modality Sensitivity	How much the model relies on the audio.	A "mute test" that measures the total change in prediction when audio is removed.	**Higher is better.** A high value proves the model is actually listening, not just looking.	**0.8266**: The model showed a strong reliance on the audio for its final answer.

3. The Experiments (The Tests)

Test Name	**Video + Audio Used**	**Final Caption Produced**
Matched	Glass Video + Breaking Sound	"A city in the game with an underlying sound of breaking."
Swapped	Glass Video + Bird Squawk	"A city in the game with an underlying sound of squawk."

What this implies:

Because the model changed its description from "breaking" to "squawk" when the audio was swapped, it proves the audio is a direct cause of what the model says. If it were biased only toward vision, it would have kept saying "breaking" despite the bird sound.

4. Summary of Results

In the "Swapped" test, negative **Total Effect (-0.1725)**. This is a key finding: it indicates the model "realized" the bird sound didn't match the visual of the glass and city. This proves you have a high-quality model that successfully weighs both visual and audio cues to understand a scene.