AUGUST 2024

TEXAS
The University of Texas at Austin

MULTIMODAL CONTEXTUALIZED SEMANTIC PARSING FROM SPEECH

ACL 2024 Main, Bangkok, Thailand

ACL 2024
Bangkok, Thailand

Paper Link

JORDAN VOAS, RAYMOND MOONEY, DAVID HARWATH
The University of Texas at Austin

Introduction

SPICE - Semantic Parsing in Contextual Environments

Aims to formulate multi-turn, multimodal, dialogue through the iterative updates and utilization of knowledge graphs with Semantic Parsing.

Paper Link

Introduction

SPICE - Semantic Parsing in Contextual Environments

Aims to formulate multi-turn, multimodal, dialogue through the iterative updates and utilization of knowledge graphs with Semantic Parsing.

SPICE advances applications of Semantic Parsing to compliment dialogue focused tasks

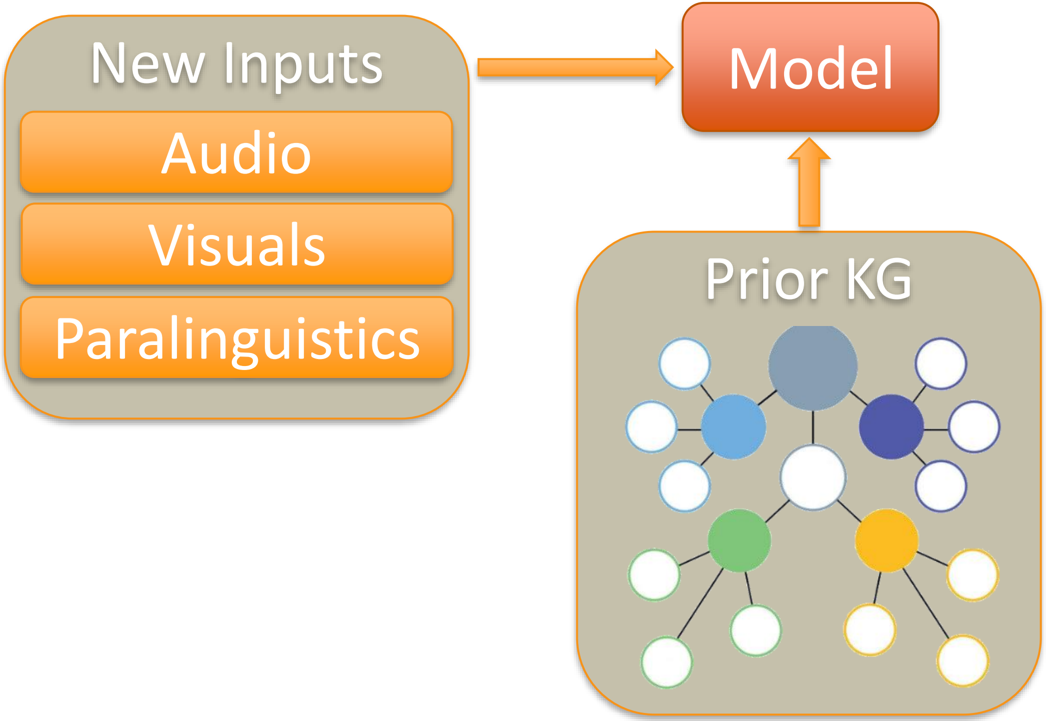
Paper Link

Motivation

Human Conversation is:

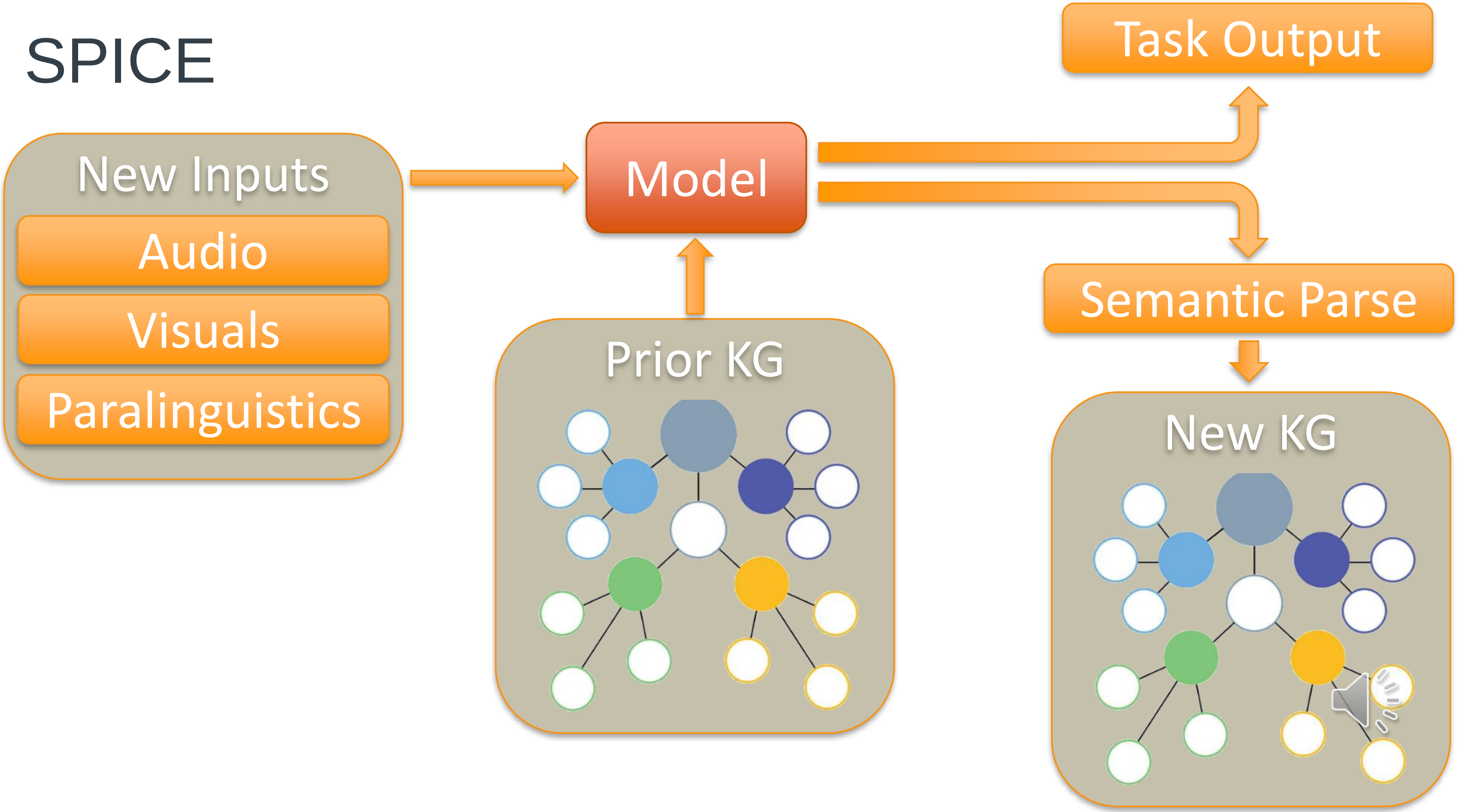
- Iterative
- Multimodal
  - e.q., Audio, Vision, Paralinguistics
- Exists within a structured knowledge base

SPICE



Paper Link

SPICE


New Inputs
Audio
Visuals
Paralinguistics
Model
Prior KG
Task Output
Semantic Parse
New KG

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

SPICE

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

- **Unimodal:** Processes primarily textual data

SPICE

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

- **Unimodal:** Processes primarily textual data
- **Non-Structure:** Rarely conditioned on dynamic structured contexts

SPICE

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

- **Unimodal:** Processes primarily textual data
- **Non-Structure:** Rarely conditioned on dynamic structured contexts
- **Single-Round:** Lacks integration of iterative applications

SPICE

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

- **Unimodal:** Processes primarily textual data
- **Non-Structure:** Rarely conditioned on dynamic structured contexts
- **Single-Round:** Lacks integration of iterative applications

SPICE

- **Multimodal:** Requires multimodal input utilization

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

- **Unimodal:** Processes primarily textual data
- **Non-Structure:** Rarely conditioned on dynamic structured contexts
- **Single-Round:** Lacks integration of iterative applications

SPICE

- **Multimodal:** Requires multimodal input utilization
- **Iterative:** Requires iteratively updating the context over multiple interactions

SPICE vs Prior Semantic Parsing Applications

Semantic Parsing

- **Unimodal:** Processes primarily textual data
- **Non-Structure:** Rarely conditioned on dynamic structured contexts
- **Single-Round:** Lacks integration of iterative applications

SPICE

- **Multimodal:** Requires multimodal input utilization
- **Iterative:** Requires iteratively updating the context over multiple interactions
- **Structure Conditioning**: Requires conditioning on both novel inputs and prior contexts at each update

SPICE Benefits

Computationally Efficient

Paper Link

SPICE Benefits

Computationally Efficient

Human Comprehensible

Paper Link

SPICE Benefits

Computationally Efficient

Human Comprehensible

Modular and Adaptable

Paper Link

How can we develop and measure current SPICE capabilities?

Paper Link

VG-SPICE

- A novel training and evaluation dataset matching SPICEs formulation
- Derived from Visual Genome annotations refined with synthetic augmentations
- Simulates iterative construction of scene graphs from single-perspective dialogue

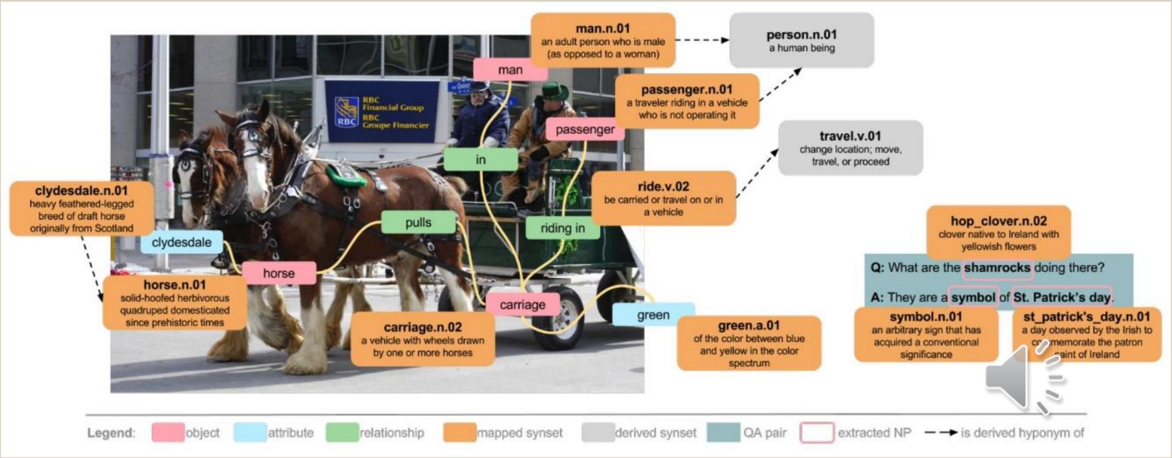
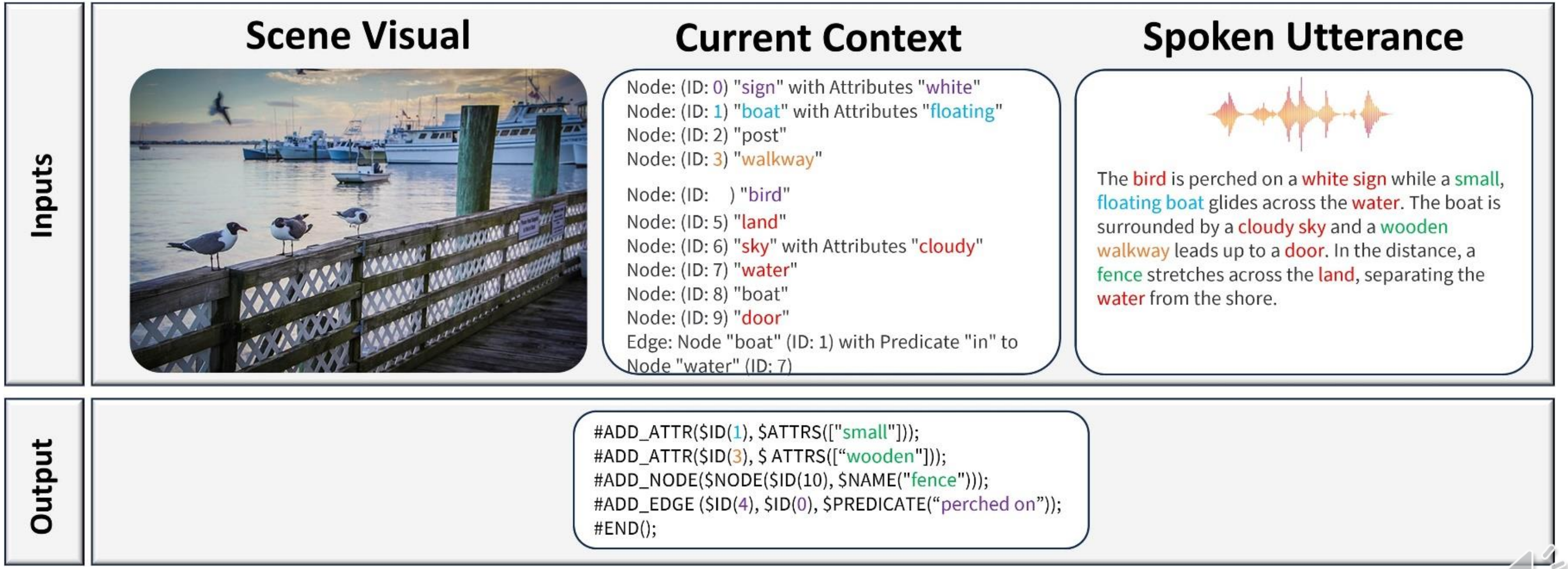


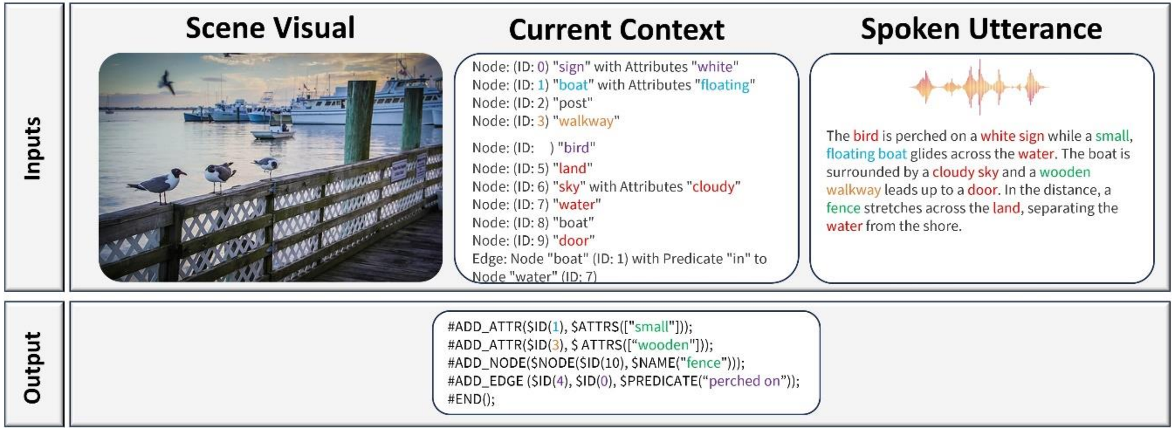
Image: Visual Genome (Krishna et al.)

VG-SPICE



Paper Link

VG-SPICE



Statistic	Value
# Samples	131362
# Unique Scenes	22346
Hours of Audio	10.56
Avg. Words per Utterance	71.83
Avg. Nodes Added	1.27
Avg. Attributes Added	0.93
Avg. Edges Added	0.60

Paper Link

VG-SPICE Clean-Challenge Set

- Sample evaluation subset of 50 visual scenes over 250 samples

Paper Link

VG-SPICE Clean-Challenge Set

- Sample evaluation subset of 50 visual scenes
- Human annotated for high quality and dense scene graphs and dialogue utterances

Paper Link

VG-SPICE Clean-Challenge Set

- Sample evaluation subset of 50 visual scenes
- Human annotated for high quality and dense scene graphs and dialogue utterances
- Includes both TTS audio samples and single voice real human speech for out of domain evaluation

Paper Link

Audio-Visual Dialogue Scene Parser (AViD-SP)

- To set a baseline for VG-SPICE we produce a initial model, AViD-SP, built on LLaMa 2 7B, using pretrained per-modality encoders (DINOv2 and Whisper-Large)

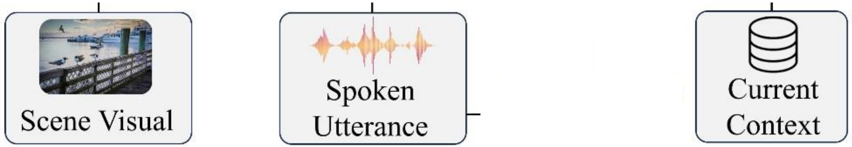
Paper Link

Audio-Visual Dialogue Scene Parser (AViD-SP)

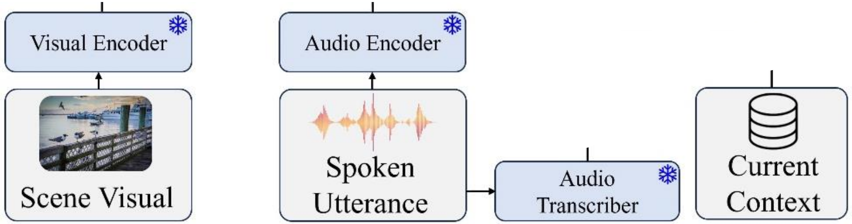
- To set a baseline for VG-SPICE we train a LLM based model, AViD-SP, built on LLaMa 2 7B, using pretrained per-modality encoders (DINOv2 and Whisper-Large)
- We evaluate AViD-SP with two forms of multimodal features adaptation modules
  - Linear Projection + Meanpooling
  - A novel Grouped Modality Adaptation Down Sampler (GMADS)

Paper Link

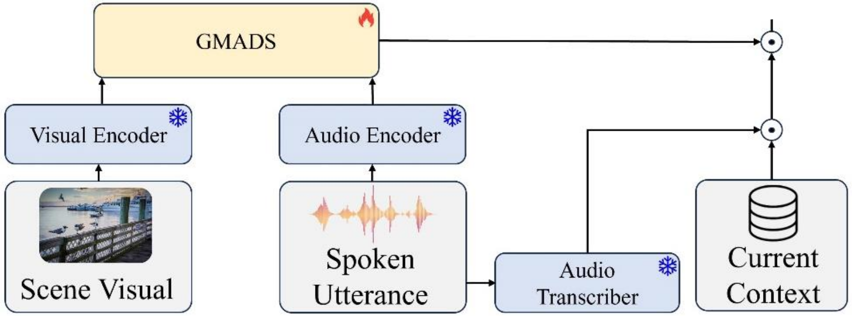
Audio-Visual Dialogue Scene Parser (AViD-SP)



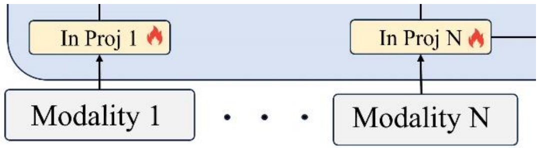
Audio-Visual Dialogue Scene Parser (AViD-SP)

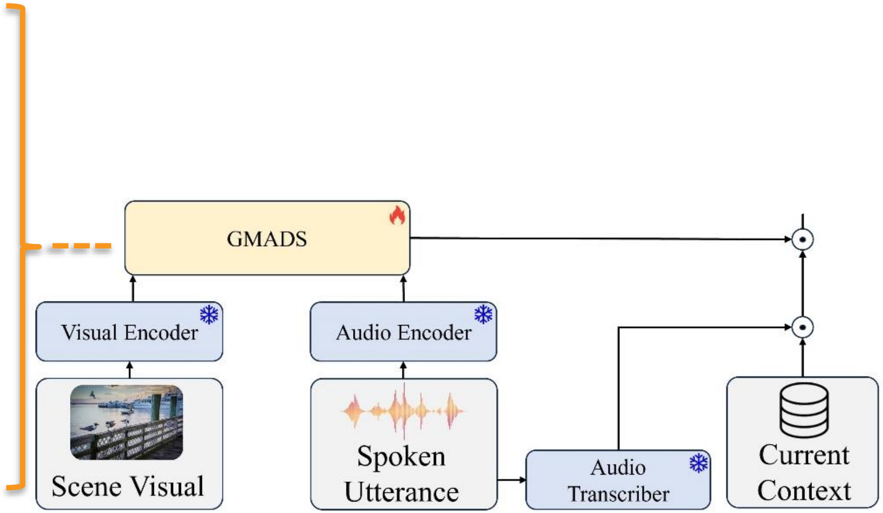


Audio-Visual Dialogue Scene Parser (AViD-SP)

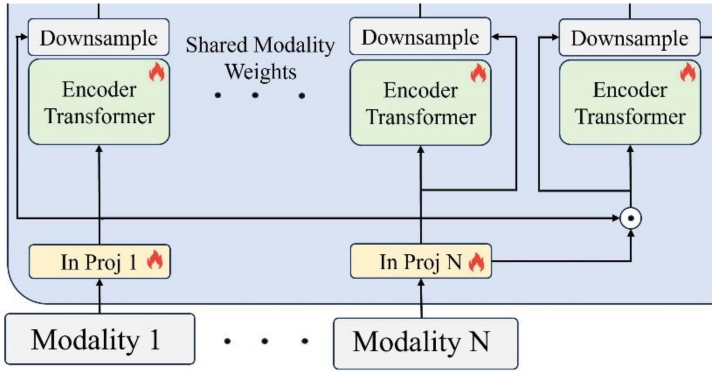


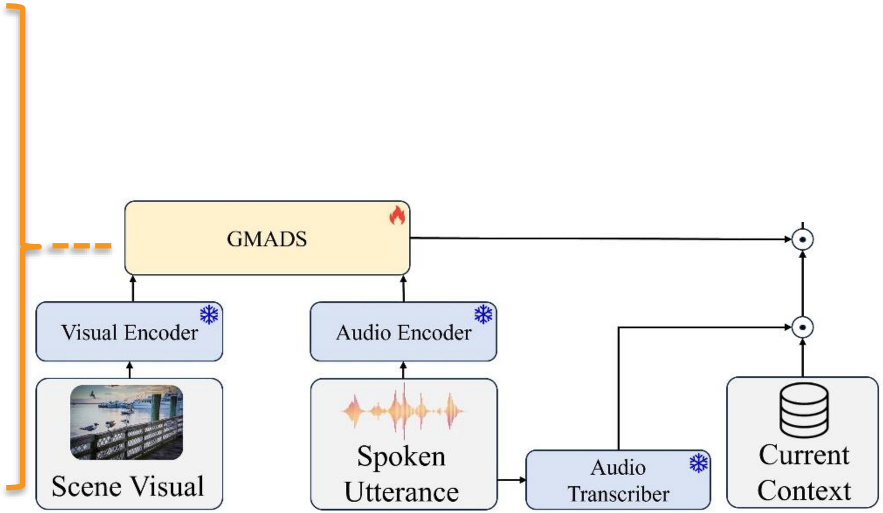
Audio-Visual Dialogue Scene Parser (AViD-SP)



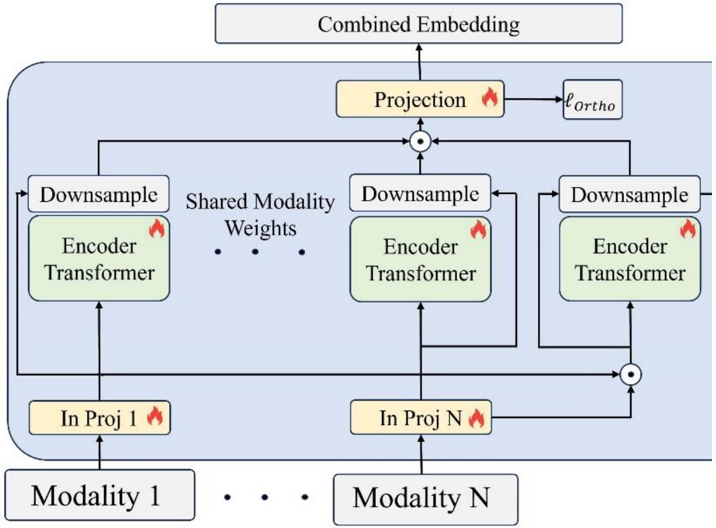


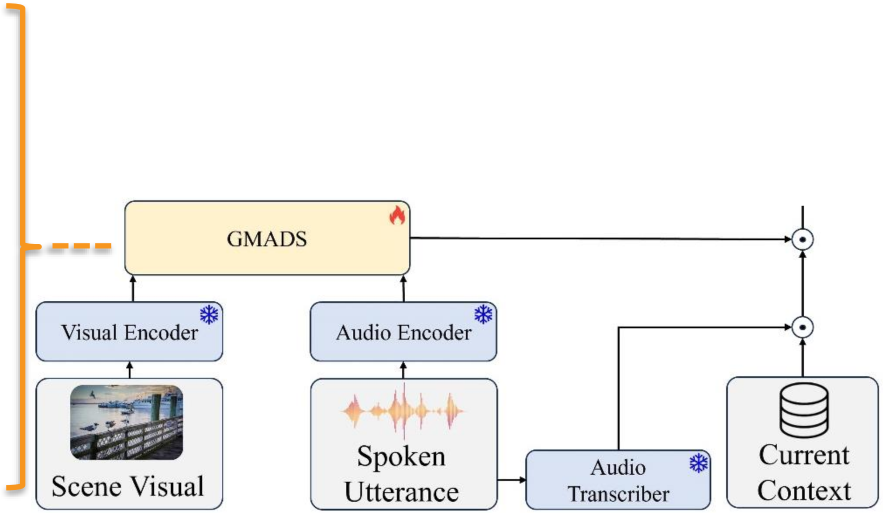
Audio-Visual Dialogue Scene Parser (AViD-SP)



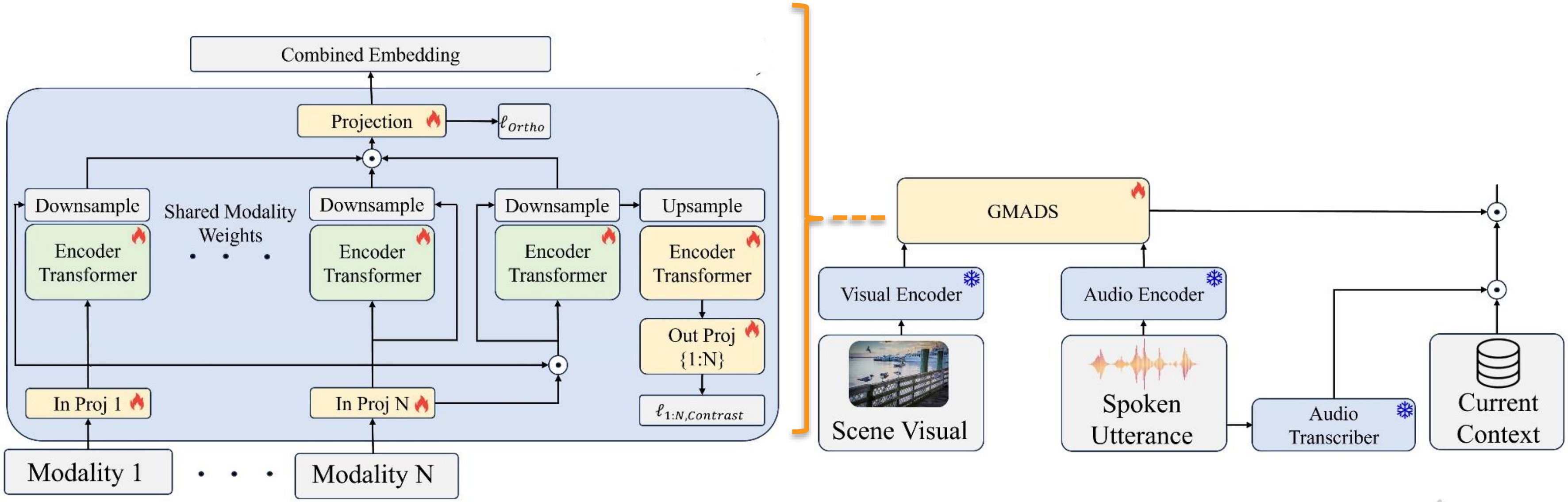


Audio-Visual Dialogue Scene Parser (AViD-SP)

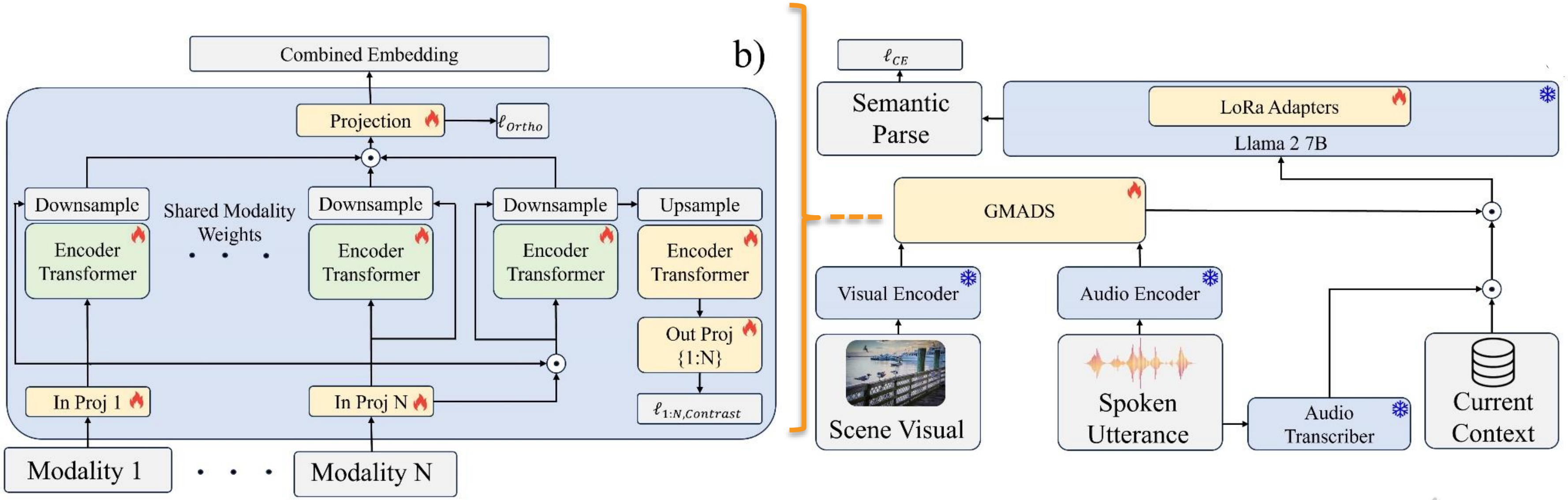




Audio-Visual Dialogue Scene Parser (AViD-SP)

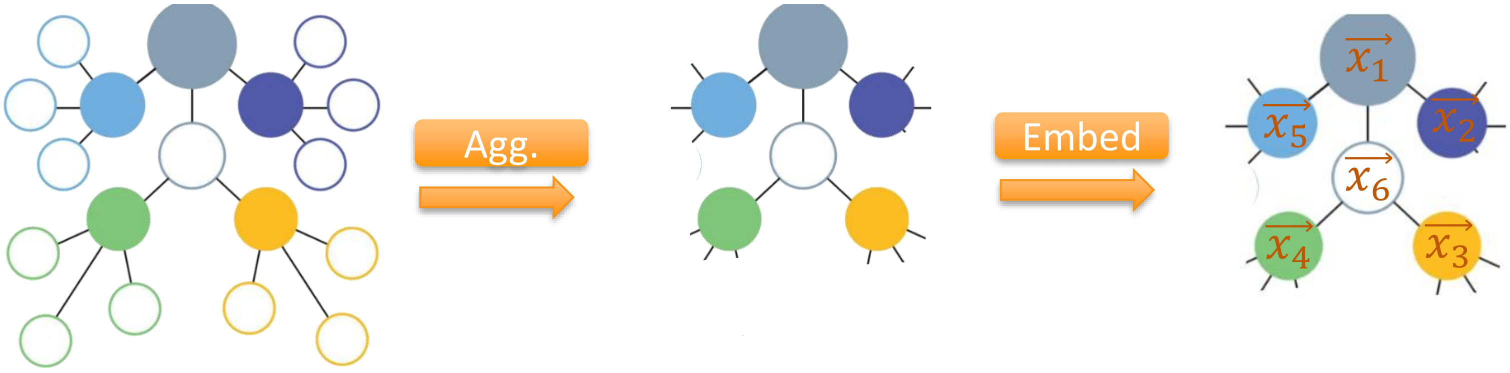


Audio-Visual Dialogue Scene Parser (AViD-SP)



Evaluation Metrics

- Evaluations are performed using Representation Edit Distance
  - Group Attributes and Nodes together
  - Uses sentence embedding representations to identify semantic edit distance between reference and prediction
  - We include both Soft (penalizes only omissions) and Hard (penalizes erroneous additions as well) metric variants



Paper Link

Evaluation Results

Model Type	S-RED↓		
	0dB	20dB	Gold*
AViD-SP + GMADS			
Base	0.402	0.3765	0.348
w/o Image	0.407	0.384	0.364
w/o Audio	0.570	0.538	0.481
w Incorrect Image**	-	0.381	-
w/o Prior Context***	-	0.478	-
AViD-SP + Meanpool			
Base	0.377	0.359	0.323
w/o Image	0.386	0.362	0.330
w/o Audio	0.414	0.385	0.363

Paper Link

Evaluation Results

Model Type	S-RED↓		
	0dB	20dB	Gold*
AViD-SP + GMADS			
Base	0.402	0.3765	0.348
w/o Image	0.407	0.384	0.364
w/o Audio	0.570	0.538	0.481
w Incorrect Image**	-	0.381	-
w/o Prior Context***	-	0.478	-
AViD-SP + Meanpool			
Base	0.377	0.359	0.323
w/o Image	0.386	0.362	0.330
w/o Audio	0.414	0.385	0.363

Paper Link

Clean Challenge Set Results

- Our novel multimodal fusion method, GMADS, manages to far exceed meanpooling on out of domain real-world performance.

Variant	TTS		Read	
	H-RED↓	S-RED↓	H-RED↓	S-RED↓
GMADS	0.739	0.497	0.731	0.497
Meanpool	0.640	0.460	1.415	0.628

Paper Link

AUGUST 2024

The University of Texas at Austin

Contact

Jordan Voas
University of Texas at Austin
Email: jvoas@utexas.edu
Website: jordanvoas.com
Phone: (320) 267-2665

ACL 2024 Main, Bangkok, Thailand

JORDAN VOAS
PhD, The University of Texas at Austin