Multimodal Contextualized Semantic Parsing from Speech

Jordan Voas; Raymond Mooney; David Harwath
Computer Science Department, University of Texas at Austin, U.S.A.

TEXAS The University of Texas at Austin

ACL 2024 Bangkok, Thailand

Introduction

We present a novel application of multimodal semantic parsing, in the form of Semantic Parsing in Contextual Environments (**SPICE**), a SPICE-focused dataset, **VG-SPICE**, and a strong initial baseline model, **AViD-SP**.

SPICE

Overview

- **Purpose**: Boosts agents' contextual awareness through multimodal inputs and dynamic knowledge updates, catering to real-world communication.
- **Dialogue Enhancement**: Advances dialogue systems with structured, interpretable updates.

Core Components

- **Context Representation**: Employs a knowledge graph to capture and maintain context from ongoing interactions.
- **Multimodal Inputs**: Combines speech, text, and images to clarify ambiguities and enhance comprehension.
- **Real-Time Updates**: Enables continuous knowledge graph adjustments, reflecting natural conversational flow and iterative context enrichment.

VG-SPICE Dataset

Utilizes Visual Genome for simulating real-world visual scene graph construction in conversational settings. Challenges agents to build knowledge graphs from visual and auditory inputs (Fig. 1).

Motivation

- **Realistic Interaction**: Mimics natural human dialogue and visual perception.
- **Multimodal Integration**: Enhances processing and integration of diverse data types, crucial for real-world applications.
- **Data Selection**: Chose Visual Genome for its detailed scene graphs and diverse imagery.

Dataset Generation

- **Preprocessing**: Standardized terms in Visual Genome, removed duplicates, corrected inconsistencies.
- **Utterance Creation**: Generated realistic, multi-turn dialogues using LLMs and TTS models.
- **Clean Subset**: Includes human-annotated samples for realistic and out-of-domain evaluations.
- **Noise Robustness**: Applied noise augmentation with CHiME5 to replicate noisy environments.

Dataset Statistics

- Samples: 131,362
- Unique Scenes: 22,346
- Audio Hours: 10.56
- Avg. Words/Utterance: 71.83

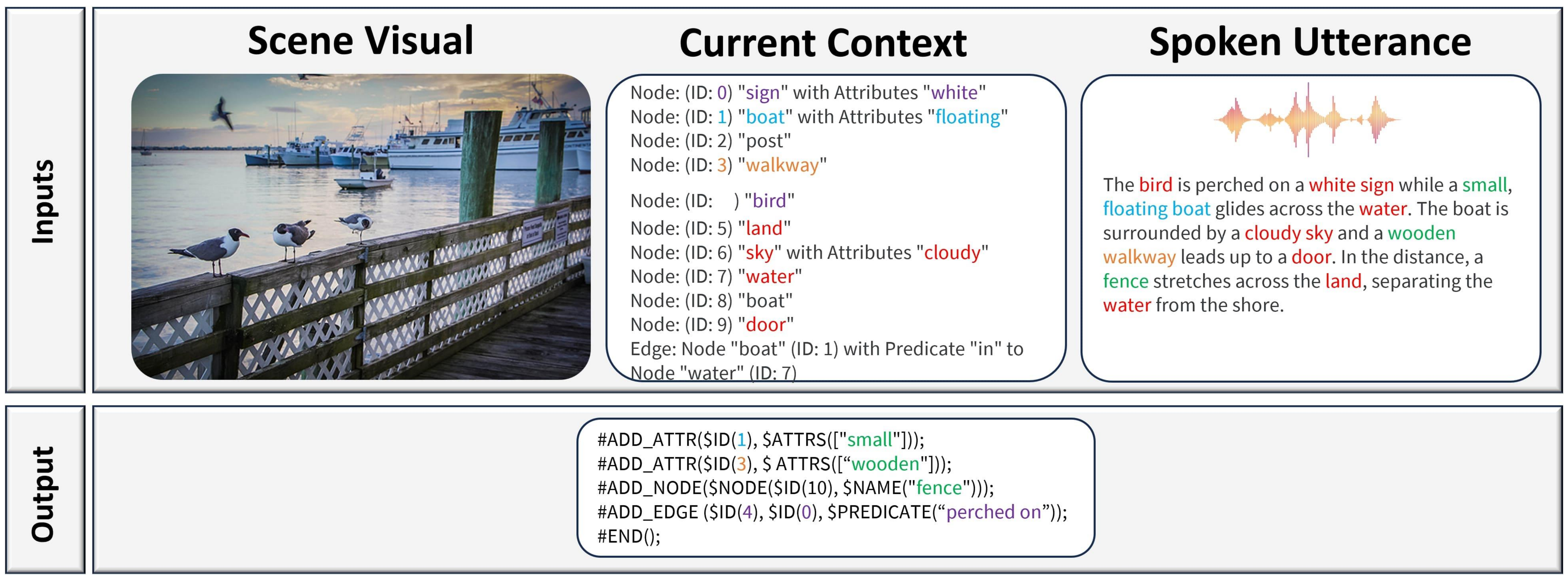

Figure 1. Example of VG-SPICE inputs and output showing the correct next state context. New information is in green, known information in red, and grounding entities in blue and orange. The current context is a textually prompted knowledge graph. .

AViD-SP Model

Incorporates Llama 2 7B, DINOv2, and Whisper-Large V3 into VG-SPICE for advanced semantic parsing. Uses a novel Grouped Modality Attention Down Sampler (**GMADS**) to efficiently fuse multimodal inputs. (Fig. 2)

Core Components & Integration

- Llama 2 7B: Forms the foundation for semantic parsing from multimodal data.
- DINOv2: Encodes visual inputs, boosting the ability to interpret complex or ambiguous scenes.
- Whisper-Large V3: Transforms speech into both latent representations and text.
- **GMADS:** Maps embeddings from audio and visual inputs into a unified space.
- Utilizes self-attention layers and mean pooling to dynamically downsample and integrate features, enhancing memory and processing efficiency.

Adaptations for Enhanced Performance

- ASR Transcription: Boosts parsing accuracy by integrating textual embeddings from audio transcriptions.
- Noise Augmentation: Trains with environmental noise from the CHiME5 dataset, improving resilience to real-world audio challenges.

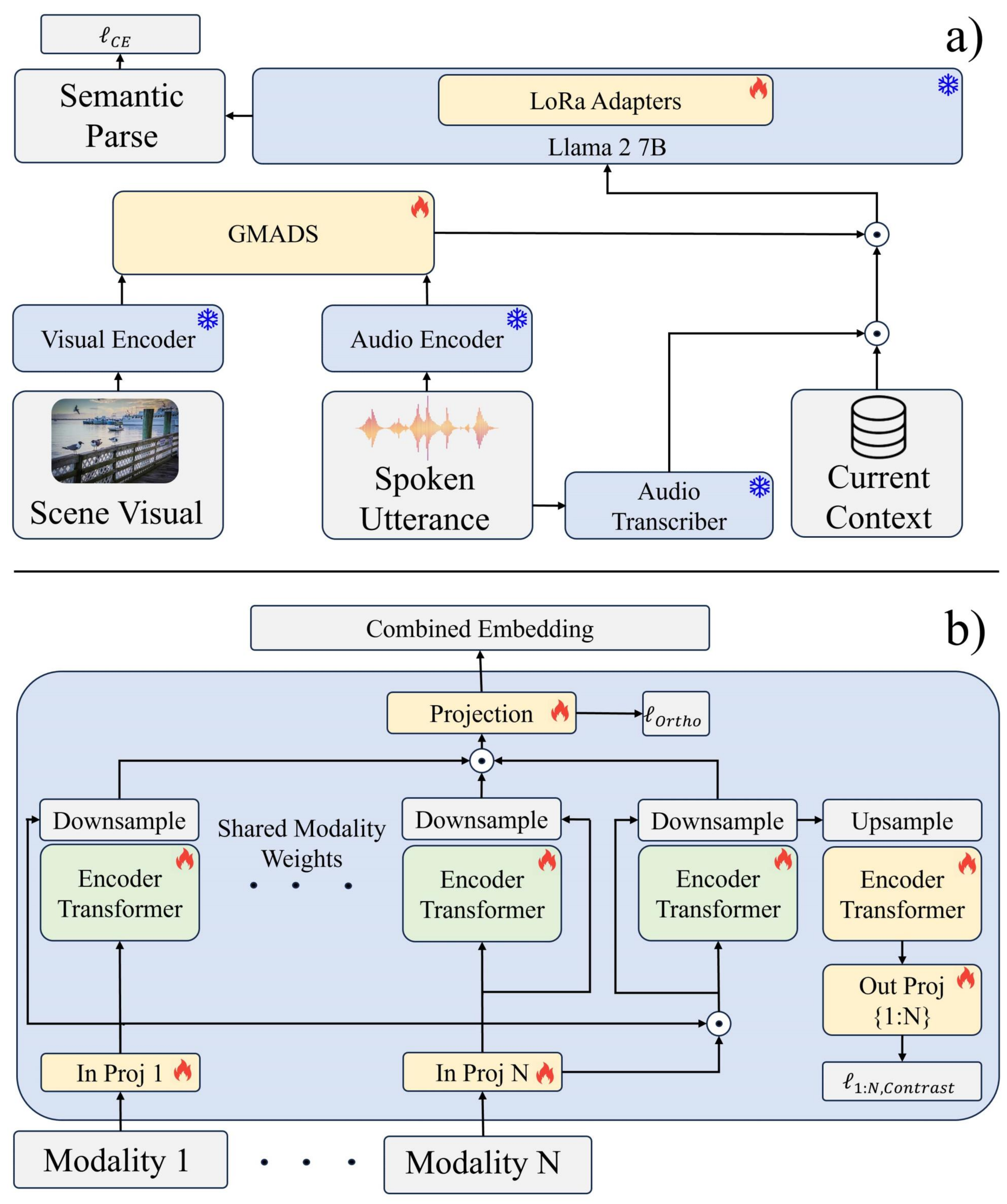

Figure 2. Architecture of the AViD-SP model for VG-SPICE.

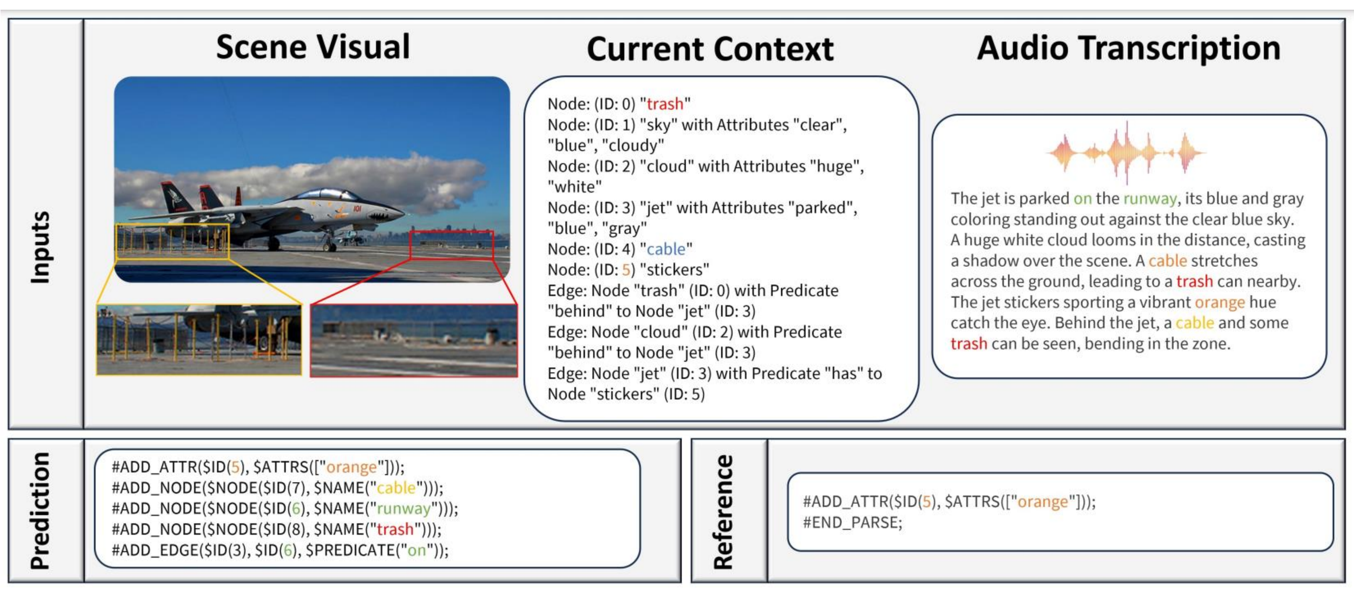

Figure 3. Example output from AViD-SP on the VG-SPICE dataset. Extraneous information are often valid, justifying use of Soft metrics.

Evaluation

VG-SPICE applies Representation Edit Distance (RED) to measure the accuracy of semantic parses from the AViD-SP model, alongside metrics like Graph Edit Distance and H-RED.

Representation Edit Distance (RED)
Evaluates accuracy by incorporating semantic similarity; groups nodes and attributes into phrases, and aggregates semantic differences, normalized to indicate the percentage of precisely captured information.

Metric Variants

- Hard (H): Penalizes missing and extraneous details.
- Soft (S): Primarily assesses omissions, allowing for a lenient performance evaluation.

We evaluate the GMADS method's performance with traditional pooling along, highlighting the advantages of advanced multimodal integration.

Results

Core Performance Metrics (Table 1, Figure 3)

- **High Accuracy in Information Assimilation**: Achieves S-RED scores below 0.4, demonstrating substantial effectiveness.
- **Resilience to Background Noise**: Maintains strong performance across various SNR levels, showcasing robustness to environmental noise.
- **Enhanced with Gold Standard Transcriptions**: Significant improvement in parsing accuracy when utilizing perfect transcriptions.
- **Handling of Irrelevant Information**: Some irrelevant information erroneously introduced.

Multimodal Feature Utilization

- **Effective Multimodal Integration**: Minor performance declines when omitting visual inputs or using incorrect images, highlighting effective but partial utilization of multimodal features.
- **Superior Handling of Out-of-Domain Audio**: In tests on a clean-challenge subset, GMADS outperforms traditional mean pooling, particularly with human-annotated audio (Table 2)

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

Table 1. Results on the VG-SPICE test set for our AViD-SP model..

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

Table 2. Results on the VG-SPICE-C test set..

Paper Link

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