Multimodal Contextualized Semantic Parsing from Speech

Jordan Voas and **Raymond Mooney** and **David Harwath**
jvoas@utexas.edu and mooney@utexas.edu and harwath@utexas.edu
The University of Texas at Austin

Abstract

We introduce Semantic Parsing in Contextual Environments (SPICE), a task designed to enhance artificial agents' contextual awareness by integrating multimodal inputs with prior contexts. SPICE goes beyond traditional semantic parsing by offering a structured, interpretable framework for dynamically updating an agent's knowledge with new information, mirroring the complexity of human communication. We develop the VG-SPICE dataset, crafted to challenge agents with visual scene graph construction from spoken conversational exchanges, highlighting speech and visual data integration. We also present the Audio-Vision Dialogue Scene Parser (AViD-SP) developed for use on VG-SPICE. These innovations aim to improve multimodal information processing and integration. Both the VG-SPICE dataset and the AViD-SP model are publicly available. [1] [2]

1 Introduction

Imagine you are taking a guided tour of an art museum. During the tour as you visit each piece of art, your guide describes not only the artworks themselves but also the history and unique features of the galleries and building itself. Through this dialog, you are able to construct a mental map of the museum, whose entities and their relationships with one another are grounded to their real-world counterparts in the museum. We engage in this type of iterative construction of grounded knowledge through dialog every day, such as when teaching a friend how to change the oil in their car or going over a set of X-rays with our dentist. As intelligent agents continue to become more ubiquitous and integrated into our lives, it is increasingly important to develop these same sorts of capabilities in them.

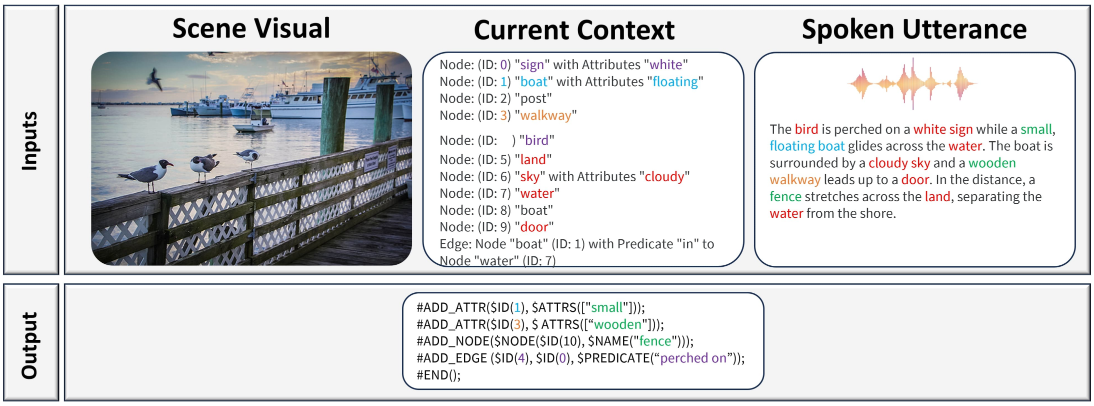
Toward this goal, this work introduces Semantic Parsing in Contextual Environments (SPICE), a task designed to capture the process of iterative knowledge construction through grounded language. It emphasizes the continuous need to update contextual states based on prior knowledge and new information. SPICE requires agents to maintain their contextual state within a structured, dense information framework that is scalable and interpretable, facilitating inspection by users or integration with downstream system components. SPICE accomplishes this by formulating updates as Formal Semantic Parsing, with the formal language defining the allowable solution space of the constructed context.

Because the SPICE task is designed to model real-world and embodied applications, such as teaching a mobile robot about an environment or assisting a doctor with medical image annotations, there are crucial differences between SPICE and traditional text-based semantic parsing. First, SPICE considers parsing language within a grounded, multimodal context. The language in cases like these may have ambiguities that can only be resolved by taking into account multimodal contextual information, such as from vision.

Furthermore, SPICE supports linguistic input that comes in the form of both speech and text. In real-world embodied interactions, language is predominantly spoken, not written. While modern automatic speech recognition (ASR) technology is highly accurate, it is still sensitive to environmental noise and reverberation, and representing the input language as both a waveform as well as a noisy ASR transcript can improve robustness. While we do not consider it here, the SPICE framework also supports paralinguistic input such as facial expressions, eye gaze, and hand gestures.

We present a novel dataset, VG-SPICE, derived from the Visual Genome (Krishna et al., 2016), an existing dataset comprised of annotated visual

[1] https://github.com/jvoas655/VG-SPICE
[2] https://github.com/jvoas655/AViD-SP



```
Node: (ID: 0) "sign" with Attributes "white"
Node: (ID: 1) "boat" with Attributes "floating"
Node: (ID: 2) "post"
Node: (ID: 3) "walkway"

Node: (ID:    ) "bird"
Node: (ID: 5) "land"
Node: (ID: 6) "sky" with Attributes "cloudy"
Node: (ID: 7) "water"
Node: (ID: 8) "boat"
Node: (ID: 9) "door"
Edge: Node "boat" (ID: 1) with Predicate "in" to
Node "water" (ID: 7)
```

The bird is perched on a white sign while a small, floating boat glides across the water. The boat is surrounded by a cloudy sky and a wooden walkway leads up to a door. In the distance, a fence stretches across the land, separating the water from the shore.

```
#ADD_ATTR($ID(1), $ATTRS(["small"]));
#ADD_ATTR($ID(3), $ ATTRS([“wooden"]));
#ADD_NODE($NODE($ID(10), $NAME("fence")));
#ADD_EDGE ($ID(4), $ID(0), $PREDICATE(“perched on”));
#END();
```

Figure 1: Example of VG-SPICE inputs as well as a plausible output to produce the correct next state context. New information that the agent is expected to add to the context is shown in green while already known information is noted in red. Grounding entities that have new information being added to them are noted in blue and orange. The current context is shown as a textually prompted representation of the actual knowledge graph (discussed in Section F).

scene graphs representing constituent entities and relational prepositions, enhanced with additional processing and synthetic augmentation to form a foundational representation for SPICE tasks. VG-SPICE simulates the conversational construction of visual scene graphs, wherein a knowledge graph representation of the entities and relationships contained within an image must be collected from the visual inputs and audio dialogue. This dataset, along with an initial model trained for VG-SPICE, sets the baseline for future efforts. Figure 1 shows an example of a typical VG-SPICE sample. The figure shows how potential semantic parses can be extracted from the visual scene and spoken utterance conditioned on what information is already known about the scene.

The remainder of this paper is structured as follows: It begins with a detailed analysis of the SPICE task, introduces the VG-SPICE dataset, and presents our AViD-SP model. It then delves into experimental results, showcasing the model's ability to process and interpret context consistent with the SPICE framework. Finally we outline the implications and directions for future research. The main contributions include:

- A definition of the Semantic Parsing in Contextual Environments (SPICE) task, highlighting its challenges, scope, and significance in enhancing human-AI communication.
- The creation of a large, machine-generated SPICE dataset, VG-SPICE, leveraging existing machine learning models and the Visual Genome dataset, to motivate SPICE research.
- An initial baseline model, Audio-Vision Dialogue Scene Parser (AViD-SP), for VG-SPICE that integrates Language Models with Audio/Visual feature extractors, establishing a research benchmark for SPICE. As a component of AViD-SP, we also introduce a novel pretrained encoder adaption and multimodal fusion method, the Grouped Multimodal Attention Down Sampler (GMADS) to motivate the exploration of additional multimodal adaptation methods.

2 Related Work

The SPICE task intersects with research in dialogue systems and semantic parsing. While previous efforts in these areas have addressed some elements of SPICE, none have fully encapsulated the comprehensive requirements of the SPICE task.

2.1 Dialogue Systems and Multimodality

Dialogue systems share similarities with SPICE tasks, particularly in their aim to emulate human conversational skills, including referencing prior conversational context. However, SPICE differentiates itself by necessitating multimodal interactions, the utilization of structured and interpretable knowledge representations, and the capability for dynamic knowledge updates during conversations, setting it apart from conventional dialogue models.

Recent advancements in dialogue systems, particularly through large language models (LLMs) (Wei et al., 2022; Chowdhery et al., 2022; Ouyang et al., 2022; Jiang et al., 2023; Touvron et al., 2023a,b), have enhanced the ability to manage complex, multi-turn conversations. This is largely thanks to the employment of extensive context windows (Dao, 2023), improving language comprehension and generation for more coherent and contextually appropriate exchanges. Nevertheless, LLMs' reliance on broad textual contexts can compromise efficiency and interpretability in many applications. Not only must all prior inputs be reprocessed for future updates but the uncompressed format prevents easy end-user inspection of the information the model is tracking for future interactions.

Advances in multimodal dialogue systems, incorporating text, image, and audio inputs (Liu et al., 2023; Zhu et al., 2023; Dai et al., 2023; Zhang et al., 2023a; Maaz et al., 2023), edge closer to SPICE's vision of multimodal communication. Yet, these systems cannot often distill accumulated knowledge into concise, understandable formats, instead still relying on raw dialogue histories or opaque embeddings for prior context.

While some systems are beginning to interact with and update external knowledge bases, these interactions tend to be unidirectional (Cheng et al., 2022; Wu et al., 2021) or involve knowledge storage as extensive, barely processed texts (Zhong et al., 2023; Wang et al., 2023). Dialogue State Tracking (DST) (Balaraman et al., 2021) shares similarities with SPICE in that agents use and update their knowledge bases during dialogues. However, most DST efforts are unimodal, with limited exploration of multimodal inputs (Kottur et al., 2021). Moreover, existing datasets and models for DST do not align with the SPICE framework, as they often rely on regenerating the knowledge base with each dialogue step from all historical dialogue inputs without offering a structured representation of the prior context. SPICE, conversely, envisions sequential updates based on and directly applied to prior context, a feature not yet explored in DST. Further, we are unaware of any DST work that has attempted to utilize spoken audio.

2.2 Semantic Parsing

Semantic Parsing involves translating natural language into a structured, symbolic-meaning representation. Traditional semantic parsing research focuses on processing individual, short-span inputs to produce their semantic representations (Kamath and Das, 2019). Some studies have explored semantic parsing in dialogues or with contextual inputs, known as Semantic Parsing in Context (SPiC) or Context Dependent Semantic Parsing (CDSP) (Li et al., 2020). However, most CDSP research has been aimed at database applications, where the context is a static schema (Yu et al., 2019). While these tasks leverage context for query execution, they do not involve dynamic schema updates, instead maintaining a static context between interactions. Outside these applications, CDSP is mainly applied in DST (Ye et al., 2021; Cheng et al., 2020; Moradshahi et al., 2023; Heck et al., 2020), which we have previously differentiated from SPICE.

Furthermore, semantic parsing has traditionally been limited to textual inputs and unimodal applications. It has been extended to visual modalities, notably in automated Scene Graph Generation (SGG) tasks (Zhang et al., 2023b; Abdelsalam et al., 2022; Zareian et al., 2020). Although there has been exploration into using spoken audio for semantic parsing (Tomasello et al., 2022; Coucke et al., 2018; Lugosch et al., 2019; Sen and Groves, 2021), these efforts have been constrained by focusing on simple intent and slot prediction tasks, and have not incorporated contextual updates or complex semantic outputs.

As such, we believe SPICE to be considerably distinct from any works that have come previously. While individual components of SPICE's framework have been studied, such as semantic parsing from audio, context, or multimodal inputs, no work has utilized all of these at once. Additionally, SPICE goes beyond most semantic parsing and dialogue works, even those operating on some form of knowledge representation, by tasking the agent to produce continual updates to said knowledge graph and to maintain them in an interpretable format.

3 Task Definition

Semantic Parsing in Contextual Environments (SPICE) is defined as follows. Consider a model agent, denoted as a, designed to maintain and update a world state across interaction timesteps. Let C_i represent this world state during the i^{th} turn. For interpretability and downstream use C_i is represented as a formal knowledge graph (Chen et al., 2020). This state represents the accumulated context from prior interactions. Initially, C_i can be set to a default or empty state.

During each interaction turn, the agent encounters a set of new inputs, referred to as information inputs F_i^m, with m indicating the diversity of modalities the agent is processing. The agent's goal is to construct a formal semantic parse, $P_i =$

Dataset	#Scenes	#Nodes	#Predicates	Avg. Size
Visual Genome (Krishna et al., 2016)	108077	76,340	-	-
VG80K (Zhang et al., 2019)	104832	53304	29086	19.02
VG150 (Xu et al., 2017)	105414	150	50	6.98
Ours	22346	2032	282	19.64

Table 1: Comparison of our Visual Genome curation statistics to other works. Further details are in Section D.

$a(F_i^m, C_i)$. This parse is formulated by integrating the prior context C_i with the new information inputs F_i^m. With the aid of an execution function e, this results in an updated context $C_{i+1} = e(P_i, C_i)$.

This newly formed context C_{i+1} should represent all task essential information, both from previous context C_i and the most recent interaction round, for future rounds. C_{i+1} is expected to align with a reference context, denoted as $\hat{C}_{i+1}$, which represents the ideal post-interaction state.

4 Dataset

This section introduces VG-SPICE, a novel dataset for SPICE tasks, providing a structured benchmark for model training and evaluation. To our knowledge, VG-SPICE is the first of its kind and is derived from the Visual Genome dataset (Krishna et al., 2016) to simulate a "tour guide" providing sequential descriptions of aspects of the environment. In these scenarios, the tour guide describes a visual scene with sequential utterances, each introducing new elements to the scene. These descriptions, combined with a pre-established world state of the scene, mimic the accumulation of world state information through successive interactions.

VG-SPICE utilizes the Visual Genome's 108k images with human-annotated scene graphs for entity identification via bounding boxes, originally detected using an object identification model. The graphs include named nodes, optional attributes, and directed edges for relational predicates.

The dataset is constructed by extracting subgraphs from scene graphs as the initial context, C_i, sampled from empty to nearly complete. These are then augmented by reintegrating a portion of the omitted graph to form the updated context, C_{i+1}. Before extracting our samples, the Visual Genome data underwent preprocessing to enhance dataset quality (Section D and summary results shown in Table 1). The dataset allows flexible model implementation with semantic parses (P_i) and parsing functions (e) not predefined, allowing flexibility in modeling implementation. Our model's semantic parse format is discussed in Section G.

For each context pair (C_i, C_{i+1}), features from C_i and modified features for C_{i+1} are structured into natural language prompts. These prompts are processed by the Llama 2 70B LLM (Touvron et al., 2023a) to generate plausible sentences that describe the difference between C_i and C_{i+1}. We then synthesize spoken versions of these sentences via the Tortoise-TTS-V2 (Betker, 2022) text-to-speech (TTS) synthesis system. We configure the TTS model to randomly sample speaker characteristics from its pretrained latent space, and use the built-in "high_quality" setup for other generation settings. Before TTS conversion filtering is performed on the textual utterances to remove common recurrent terms indicative of new information (eg., "there now is a" versus "there is a"). The audio recordings and visual images are the multimodal inputs F_i^m of VG-SPICE, emphasizing spoken audio for practicality in real-world applications and necessitating addressing the challenges of semantic parsing from audio such as speaker diversity and noise robustness. The presence of both textual and spoken audio representations for the update utterances allows VG-SPICE to be utilized for semantic parsing evaluations in either modality.

VG-SPICE includes over 131k SPICE update samples from 20k unique scenes, with 2.5% allocated to each of the validation and test sets, ensuring distinct scenes across splits. We perform noise augmentation on the input speech using the CHiME5 dataset (Barker et al., 2018) to simulate realistic noise conditions, with performance evaluated at various Signal to Noise Ratios (SNR). VG-SPICE samples and summary statistics are presented in Figure 1 and Table 2, respectively.

4.1 Challenge Subset

In addition to the standard test set, we augment VG-SPICE with an additional Challenge Subset, VG-SPICE-C. Although this subset is small, spanning only 50 individual visual scenes, it provides distinct capabilities not present in the primary 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

Table 2: Summary statistics for our VG-SPICE dataset.

test dataset, as detailed below.

Broad Visual Representation: To sample the Challenge Subset, we used a representation-based process to promote diverse image types. We obtained the CLIP[3] representations for each image in the original VG-SPICE test split. Using KMeans clustering, the dataset was partitioned into 50 distinct groupings of visual representations, with a single sample taken from each cluster.

Manual Scene Graph Quality Enhancements: Despite automated generation processes in VG-SPICE aiming to improve scene graph quality, persistent issues remain. To ensure a clean and reliable testing subset, manual scene graph improvements were made to ensure the final scene graph for each image was accurate. This involved removing incorrect, low-quality, or duplicate scene features and enhancing the scene graphs to achieve far greater density than originally present in VG-SPICE or Visual Genome, particularly for Edges and Attributes.

Coherent Iterative Updates: To improve sample diversity, VG-SPICE was generated in an iteratively incoherent fashion, meaning samples for a single update cannot be used to coherently evaluate end-to-end SPICE evaluations. For the Challenge Subset, we manually annotated each of the 50 sampled scenes with five individual utterances, each adding novel information while referring to previously mentioned details. These utterances are of greater diversity and quality (due to manual annotation rather than LLM production) and can be used sequentially to evaluate scene graph generation errors over multiple interaction rounds.

OOD and Real Speech: To enhance the evaluative capabilities of the Challenge Set, we provide speech samples for the utterances from two sources: Tortoise-TTS as used for the remainder of VG-SPICE (with three random voice samples per utterance) as well as manual recordings of the spoken utterances by a individual human annotator.

This Challenge Subset offers a rigorous evaluation framework for models, promoting advancements in handling diverse visual representations, maintaining high-quality scene graphs, performing coherent iterative updates, and managing out-of-domain and real-world speech scenarios.

5 AViD-SP Model

To address the challenges of VG-SPICE, our approach utilizes a range of pretrained models, specifically fine-tuned to enhance SPICE-focused semantic parsing capabilities. Figure 2 illustrates our model architecture, termed Audio-Vision Dialogue Scene Parser (AViD-SP). At the core of our framework lies the pretrained Llama 2 7B model (Touvron et al., 2023b). Despite deploying its smallest variant, the extensive pretraining endows our model with robust functional abilities, particularly beneficial for processing the diverse semantic parses inherent to VG-SPICE. However, Llama 2, trained on textual data, lacks inherent support for the multimodal inputs typical in VG-SPICE.

To accommodate diverse inputs, we extend techniques from prior studies (Rubenstein et al., 2023; Gong et al., 2023; Lin et al., 2023) by projecting embeddings from pretrained modality-specific feature extractors. This approach has been proven to enable text-based LLMs to process information across various modalities. Directly integrating these projected embeddings into the LLM's context window, however, introduces significant computational overhead due to their typically extensive context lengths. While previous research often employed pooling methods (Gong et al., 2023) to condense embeddings by modality, this strategy incompletely addresses the challenges of merging varied modality embeddings for LLM use. For instance, audio embeddings offer finer temporal granularity than textual embeddings, and the reverse is often true for vision embeddings, complicating the adjustment of downsampling factors. Moreover, even with optimized downsampling, pooled embeddings must preserve their original sequential order and are restricted to information from solely the pooled segments. Many applications could benefit from capabilities to establish downsampled features encompassing both local and global contexts and to rearrange these features to an extent.

To surmount these challenges, we introduce a novel Grouped Modality Attention Down Sam-

[3] openai/clip-vit-base-patch32 from Huggingface

pler (GMADS) module. This module initially projects embeddings from non-textual modalities into a unified, fixed-dimensional space. We form a set of modality groupings, one for each input modality (audio and visual with VG-SPICE), and a cross-modality grouping derived from concatenating all modality embeddings, each prefixed with a modality-specific token. A series of self-attention layers processes each embedding sequence and downsamples the outputs by a factor of S through mean pooling. These values are then concatenated with the mean-pooled pre-self-attention embeddings along the embedding dimension, akin to a skip connection. A final projection adjusts the outputs to match the dimensionality of the Llama 2 7B decoder, and all embedding sequences are concatenated. This process yields an embedding output that is effectively downsampled by a factor of $S/2$. All weights in the GMADS module are shared across the groups, substantially reducing the parameter count. Additionally, we employ a self-supervised representation learning objective on the embeddings from the downsampled cross-modality group outputs by upsampling them to their original size and then processing them through a secondary set of self-attention layers. The reconstructed cross-modality embeddings are then segmented by modality, with per-modality projections striving to restore them to their original input size. We apply a contrastive reconstruction loss objective as outlined in Eq. 1, using the corresponding ground truth embedding as an anchor and all other embeddings in the batch as contrastive samples.

$$\ell_{n,Contrast} = \sum_{j=1}^{B*K} \log \frac{\exp(sim(z_i,z_j)/\tau)}{\sum_{k=1}^{B*K} [k \neq i] \exp(sim(z_i,z_k)/\tau)} \tag{1}$$

In this equation z_i denotes the reconstructed input embedding, K represents the length of each sequence, B denotes the batch size, and τ is a tunable temperature hyperparameter.

We also observed that non-textual modality inputs tended to collapse when combined with simpler textual inputs, such as prior context or ASR transcripts. To counter this, we include an additional orthogonality loss, designed to encourage maximal dissimilarity among aligned embeddings in each batch sequence. This methodology is similar to previous efforts to promote distinct class embeddings (Ranasinghe et al., 2021), but in our case, we treat each embedding as a distinct class sample. However, given the nature of these embedding sequences, some level of similarity is expected, and entirely dissimilar values (cosine similarity less than zero) are not feasible. Thus, we modify Eq. 2 to include a slight margin allowing for minimal similarity. Below, e_i represents a single GMADS output embedding (pre-output projection) within a batch of B sequences, each of length K.

$$\ell_{Ortho} = \frac{2\sum_{i=1}^{B*K-1} \sum_{j=i+1}^{B*K} max(\frac{e_i * e_j}{\|e_i\| * \|e_j\|} - h, 0))}{B*K*(B*K-1)} \tag{2}$$

The GMADS module attempts to provide several advantages over the direct use of raw modality embeddings with the LLM decoder or mean pooling. Firstly, GMADS operates at reduced dimensional scales compared to the pretrained LLM, which significantly lowers memory requirements, requiring the much larger decoder to process shorter (reduced to only $2/S$ the size) input sequences. Moreover, the modality inputs do not necessitate autoregressive generation alongside these inputs, further conserving cost. Secondly, GMADS empowers the model to selectively learn its downsampling process, including choices on whether to focus locally or integrate global features, allowing some degree of information restructuring. The incorporation of cross-modality encoding enables parts of the downsampled embeddings to capture essential information across modalities while maintaining individual modality components in the outputs ensuring that some portion of the output embeddings is conditioned on each modality, requiring the attention mechanisms to remain sensitive to all modalities.

For feature extraction, we utilize the visual encoder from DINOv2 (Oquab et al., 2024) for visual inputs and the encoder from Whisper-Large V3 (Radford et al., 2022) for audio. We retain only the necessary encoder portions of these pretrained models. In alignment with successful semantic parsing efforts from speech (Arora et al., 2023), we perform ASR transcription on the audio, appending these textual embeddings to the prior context embeddings. ASR transcriptions are generated using the Whisper-medium.en model. To enable scalable fine-tuning, we integrate LoRa adaptation layers into Llama 2 7B and freeze all feature extractors.

5.1 Training Routine

We train AViD-SP using cross-entropy loss (Eq. 3) between the predicted and reference Formal Semantic Parses, alongside the objectives in Eq. 1 and 2. Our comprehensive loss function is outlined below

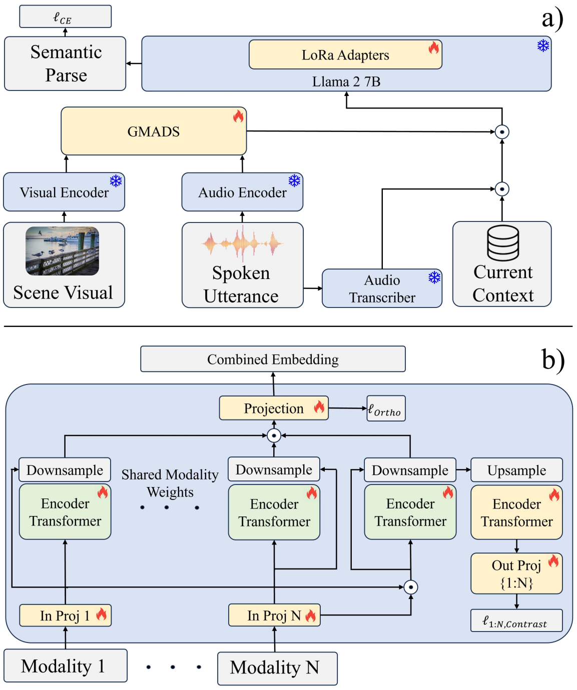


Figure 2: a) The architecture of the AViD-SP model for VG-SPICE, integrating pretrained encoders and large language models (LLMs) with LoRa adapters and feature fusion modules. Trained and frozen segments of the model are denoted by fire and snowflake icons, respectively. b) Our novel Grouped Modality Attention Down Sampler module, enabling integrated cross-modality fusion and downsampling. Green modules share weights. For downsampling, we utilize meanpooling, and for upsampling we linearly interpolate the embeddings.

in Eq. 4, where $p_{i,k}$ denotes the softmax prediction for each of the k tokens in P_i, and $t_{i,k}$ represents the corresponding ground-truth token label.

$$\ell_{CE} = -\sum_{k=1}^{n} t_{i,k} \log(p_{i,k}) \tag{3}$$

$$L = \alpha\ell_{CE} + \beta\ell_{Ortho} + \frac{\gamma}{N}\sum_{n=1}^{N} \ell_{n,Contrast} \tag{4}$$

AViD-SP employs a three-layer self-attention transformer as the primary encoder transformer, each layer having an embedding dimensionality of 1024 and 8 attention heads. The secondary encoder transformer, used for the upsampled reconstruction training objective, is of the same configuration. The GMADS module employs a downsampling factor, S, of 16. Additionally, we enhance the key, query, and value layers of the Llama 2 7B model with Low-Rank Adaptation (LoRa) layers. No hyperparameter optimization was conducted.

We train AViD-SP by incorporating randomly sampled CHiME5 noise to simulate audio corruption, adding this noise at various Signal-to-Noise Ratios (SNR) of 0, 2, 5, 10, or 20dB. Further details on training and inference hyperparameters are discussed in Section E. To ensure robustness to various input feature combinations, we implement random input dropout with a probability of 30%. In these instances, we randomly omit one of the input modalities, either audio embeddings, visual embeddings, or audio transcriptions. We do not omit the prior context, as we found the task too difficult to learn under such conditions since it requires both the already known information as well as their current assigned labels under our semantic parsing framework. AViD-SP is trained in a two-stage pipeline, with the initial stage acting as pretraining without the ASR transcriptions to allow the GMADS module to reach a semi-trained state for enhanced efficiency. Subsequently, we continue fine-tuning the model with ASR transcriptions until convergence. Our initial pretraining lasts one full epochs, followed by the fine-tuning stage.

5.2 Evaluation Metrics

We use several metrics to measure how closely the generated semantic parse aligns with the ground truth and how accurately the scene graph context updates match the reference. Unlike conventional semantic parsing assessments (Tomasello et al., 2022), we omit exact-match metrics due to their unsuitability for our problem, which allows for permutation invariance in the formal-language output (see Section G). This permits the parser to generate scene-graph updates in any order and assign node IDs freely, as long as the resulting scene graph is isomorphic to the reference.

For each below metric, we examine hard ("H") and soft ("S") variants. The hard variant penalizes missing and unnecessary information, while the soft variant only penalizes omissions. This approach accounts for the Visual Genome dataset's sparsity and the possibility of LLMs generating extraneous yet potentially valid content. For example, an LLM might enhance a "blue table" to a "vibrant blue table," making "vibrant" an acceptable attribute. Our analysis shows such inclusions are common in the VG-SPICE dataset, leading us to focus on the soft metric and qualitatively show in Section 6 how updated utterances accommodate these extraneous additions. We include results for GED in the supplement Table 5.

Graph Edit Distance (GED): GED calculates the normalized cost to transform the predicted context to the reference one, considering only perfectly semantically equivalent Nodes, Attributes, and Edges. Missing or extra Nodes or Edges increase the error by one, while incorrect Attributes have a smaller penalty of 0.25. GED is not normalized and should be interpreted as the magnitude of incorrect features compared with the reference solution and not as a recall or precision metric. GED is particularly reliant on exact matches, so minor discrepancies (like "snow board" vs. "snowboard") can incur significant penalties, with misalignments doubly penalized in the hard variant.

Representation Edit Distance (RED): RED addresses the limitations of GED by employing a “softer” semantic similarity to evaluate entity pairings. Using a transformer model for sentence semantic similarity[4], RED groups Nodes and their Attributes into descriptive phrases (for example, a "table" Node with "vibrant" and "blue" Attributes becomes "vibrant blue table") and assesses the dissimilarity between potential pairings, using an exhaustic search for optimal pairings of Nodes and Edges. Unmatched Nodes and Edges are considered entirely dissimilar. Since unmodified graph portions from the prior context are pre-matched and excluded from the exhaustive search, the computation of the pairings remains manageable. RED is normalized by the representation edit distance needed to transform the prior context into the reference context, and so numerically can be interpreted as the percentage of missing and/or extra information relative to the reference context.

5.3 Baselines and Evaluation

To thoroughly evaluate our AViD-SP model, we conducted a series of ablation studies to explore the impact of various input modality combinations. Given that AViD-SP was trained under diverse noise conditions, its performance was tested across noise levels of 0, 2, and 20 dB using the CHiME5 dataset. We assessed the model’s capability to resolve ambiguities in audio input by introducing tests with and without visual modality, and by evaluating the model with incorrectly matched images in the GMADS module. Additionally, we explored potential enhancements in ASR performance by incorporating ground truth ASR transcriptions in our evaluations. To ablate the effects our GMADS module has on performance we compare against a version of AViD-SP trained using traditional meanpooling after a per modality projection layer to downsample the audio and visual input embeddings, with all hyperparameter and training methods matched between the two except the meanpooling baseline only utilizing the cross entropy component of the full training objective.

We also extended our evaluations to the VG-SPICE-C Subset. Here, we analyze the subset through a single-step evaluation approach, with ground truth prior context provided and metrics measured after each individual SPICE update.

6 Results

The performance of the AViD-SP model on the VG-SPICE test set, as shown in Table 3, demonstrates that the baseline AViD-SP achieves S-RED scores just below 0.4, with the meanpooling variant slightly lower, approaching 0.38. This performance suggests a substantial effectiveness (over 60%) in assimilating desired information into the scene graph. However, the H-RED metrics indicate the introduction of moderate quantities of irrelevant information, particularly in the GMADS version. Given that VG-SPICE scene graphs are often overly sparse, the elevated H-RED values for GMADS may reflect an increased utilization of visual inputs, possibly learning to incorporate non-essential features detected through visual cues. While this interpretation is speculative, some level of elevated H-RED could be reasonable for VG-SPICE in its current state (Section C).

Under varying SNR conditions, both GMADS and meanpooling configurations of AViD-SP show minimal performance degradation at lower SNRs, indicating resilience to reasonable background noise levels. The use of accurate ASR transcriptions substantially boosts parsing accuracy, emphasizing the benefits of reliable ASR.

Experiments omitting visual inputs or incorporating incorrectly paired visual inputs exhibit minor performance declines. For the meanpooling based AViD-SP a slightly larger, but still quite minor, degradation in metric performance is observed when audio inputs are excluded, with only ASR transcriptions being provided. However, a more significant degradation is observed for the GMADS variant of AViD-SP under these same conditions. This implies that the GMADS multimodal adapta-

[4] The “en_stsb_roberta_base” model from https://github.com/MartinoMensio/spacy-sentence-bert

Model Type		H-RED↓				S-RED↓			
		0dB	2dB	20dB	Gold*	0dB	2dB	20dB	Gold*
AViD-SP + GMADS	Base	1.618	1.517	1.412	1.272	0.402	0.383	0.3765	0.348
	w/o Image	1.611	1.527	1.430	1.33	0.407	0.393	0.384	0.364
	w/o Audio	1.660	1.607	1.590	1.540	0.570	0.559	0.538	0.481
	w Incorrect Image**	-	-	1.423	-	-	-	0.381	-
	w/o Prior Context***	-	-	3.428	-	-	-	0.478	-
AViD-SP + Meanpool	Base	1.083	1.038	0.940	0.817	0.377	0.368	0.359	0.323
	w/o Image	1.051	0.980	0.911	0.826	0.386	0.377	0.362	0.330
	w/o Audio	0.946	0.897	0.804	0.759	0.414	0.397	0.385	0.363

Table 3: RED results on the VG-SPICE test set for our AViD-SP model. AViD-SP was trained with CHiME5 noise augmentation sampled between 0db and 20dB SNR (all CHiME5 noise followed the provided train/eval/test splits). *Given the ground truth utterance transcripts in place of the ASR transcriptions. **Evaluated by offsetting visual features within batch so incorrect image features are paired with the other input components. ***Evaluated with "Empty Context" prior state scene graphs summaries instead of the correct ones.

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

Table 4: RED results on the VG-SPICE-C challenge test set for AViD-SP with Single Step (ground truth prior context provided for each step) metrics reported.

tion process has resulted in a model which is more sensitive to the raw audio inputs than when meanpooling is used, which seems to dominantly rely on the natively textual ASR transcriptions. We theorize that the enhanced capability of GMADS to process multimodal inputs may lead to its overall worse results, as it produces a more complex optimization landscape compared with simply collapsing to utilize only the native textual ASR transcripts. Additionally, the absence of prior context markedly increases error rates, underscoring the importance of historical context for accurate SPICE updates.

Table 4 presents the performance of AViD-SP on the VG-SPICE-C test set. For TTS audio, the metrics diverge significantly from those of the standard VG-SPICE test set, featuring higher S-RED and lower H-RED scores. The higher density of VG-SPICE-C's scene graphs, which include fewer visually or auditorily supported features that are untracked in reference scene graphs, likely contributes to these lower Hard metric scores. However, this increased density also presents a greater challenge in achieving improved Soft metric scores, as the model must correctly incorporate a substantial amount of information at each update step.

For the GMADS-based AViD-SP, performance metrics on the read audio portion of VG-SPICE-C align closely with those observed in the TTS portion. Conversely, the meanpooling variant shows a substantial performance reduction. This discrepancy suggests that GMADS possesses more robust multimodal processing capabilities, especially in processing out-of-domain real audio distributions. Since both model variants use the same ASR model without parameter tuning, the observed differences indicate that GMADS compensates more effectively for poorer ASR performance.

7 Conclusion

In this paper, we introduced Semantic Parsing in Contextual Environments (SPICE), an innovative task designed to enhance artificial agents' contextual understanding by integrating multimodal inputs with prior contexts. Through the development of the VG-SPICE dataset and the Audio-Vision Dialogue Scene Parser (AViD-SP) model, we established a framework for agents to dynamically update their knowledge in response to new information, closely mirroring human communication processes. The VG-SPICE dataset, crafted to challenge agents with the task of visual scene graph construction from spoken conversational exchanges, represents a significant step forward in the field of semantic parsing by incorporating both speech and visual data integration. Meanwhile, the AViD-SP model, equipped with the novel Grouped Multimodal Attention Down Sampler (GMADS), provides a strong initial baseline for VG-SPICE as well as insights into potential methods to improve multimodal information processing and integration.

Our work highlights the importance of developing systems capable of understanding and interacting within complex, multimodal environments. By focusing on the continuous update of contextual states based on new, and multimodal, information, SPICE represents a shift towards more natural and effective human-AI communication.

8 Limitations

While VG-SPICE and AViD-SP are novel approaches, they have several limitations and should be treated as initial attempts toward further SPICE implementations and benchmarks. The main limitation stems from the extensive use of synthetic data augmentation in VG-SPICE's creation. The process involved several steps, including dataset preprocessing with BERT-like POS taggers, crafting update utterances using the Llama 2 70B LLM, and generating synthetic TTS audio. These stages may introduce errors, hallucinations, or overly simple data distributions, potentially misaligning with real-world applications. For example, our models' resilience to background noise may reflect the specific TTS audio distribution, possibly simplifying the ASR model's speech discernment. Additionally, the Visual Genome, our work's foundation, suffers from notable quality issues, such as poor annotations and unreliable synthetic object segmentation, which, despite efforts to mitigate, remain challenges in VG-SPICE. While the included VG-SPICE-C test subset attempts to improve these limitations, and indeed the hard versions of are metrics are significantly improved on the manually cleaned samples of this subset, they are still comprised of intentionally crafted utterances with read audio, which may not transfer to real-world applications and natural spoken audio. Further, this work only includes analysis of the VG-SPICE-C challenge subset in the simple Single Step task and does not evaluate in end-to-end sequence-based analysis.

The various version of AViD-SP we introduce also provides indications of further development for efficient multimodal adaptation methodologies. While the version utilizing GMADS generally failed to outperform the results of the traditional meanpooling version the GMADS method also provided a stronger indication of cross-modality feature utilization, whereas integration of simplistically downsampled multimodal features alongside native textual features appears to cause strong underutilization and feature collapse for the multimodal features. This is further supported by the poor performance achieved by the meanpooling version of AViD-SP, relative to the GMADS version, on real human recorded audio, indicating the meanpooling version adapts much worse to out-of-domain multimodal inputs. We suggest future work to continue investigating methods similar to GMADS to further realize their theoretical benefits.

Moreover, VG-SPICE, while pioneering in SPICE tasks, is only a start, limited to audio and images, with a basic language for knowledge graph updates. Future research should address these limitations by incorporating more realistic inputs, like video, 3D environments, and paralinguistic cues, and by exploring dynamic tasks beyond simple scene graph updates. Environments like Matterport3D (Chang et al., 2017) or Habitat 3.0 (Puig et al., 2023) offer promising avenues for embodied SPICE research. Expanding SPICE to include secondary tasks that rely on an agent's contextual understanding can also enhance its utility, such as aiding in medical image annotation with co-dialogue.

References

Mohamed Ashraf Abdelsalam, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat Bhatt, Vladimir Pavlovic, and Afsaneh Fazly. 2022. Visual semantic parsing: From images to Abstract Meaning Representation. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 282–300, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Siddhant Arora, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, Brian Yan, and Shinji Watanabe. 2023. Integrating pretrained asr and lm to perform sequence generation for spoken language understanding. *ArXiv*, abs/2307.11005.

Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, Singapore and Online. Association for Computational Linguistics.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines.

James Betker. 2022. TorToiSe text-to-speech.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*.

Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Aishwarya Kamath and Rajarshi Das. 2019. A survey on semantic parsing.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2020. Context dependent semantic parsing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2509–2521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models.

Neau Maëlic, Paulo E. Santos, Anne-Gwenn Bosser, and Cédric Buche. 2023. Fine-grained is too coarse: A novel data-centric approach for efficient scene graph generation.

Mehrad Moradshahi, Victoria Tsai, Giovanni Campagna, and Monica Lam. 2023. Contextual semantic parsing for multilingual task-oriented dialogues. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 902–915, Dubrovnik, Croatia. Association for Computational Linguistics.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. 2021. Orthogonal projection loss.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. Audiopalm: A large language model that can speak and listen.

Priyanka Sen and Isabel Groves. 2021. Semantic parsing of disfluent speech. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1748–1753, Online. Association for Computational Linguistics.

Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, Robin Algayres, Tu Ahn Nguyen, Emmanuel Dupoux, Luke Zettlemoyer, and Abdelrahman Mohamed. 2022. Stop: A dataset for spoken task oriented semantic parsing.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2023. Recursively summarizing enables long-term dialogue memory in large language models.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Sixing Wu, Ying Li, Minghui Wang, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2286–2300, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing.

Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, Shenghui Li, and Emine Yilmaz. 2021. Slot self-attentive dialogue state tracking.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, and Dragomir Radev. 2019. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. 2020. Weakly supervised visual semantic parsing.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding.

Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. 2019. Large-scale visual relationship understanding.

Yong Hong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang Wen Chen. 2023b. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2915–2924.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

A Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2238605

B Additional AViD-SP Results

We report the Graph Edit Distance (GED) results for AViD-SP, and the tested baselines, here.

C Qualitative AViD-SP Examples

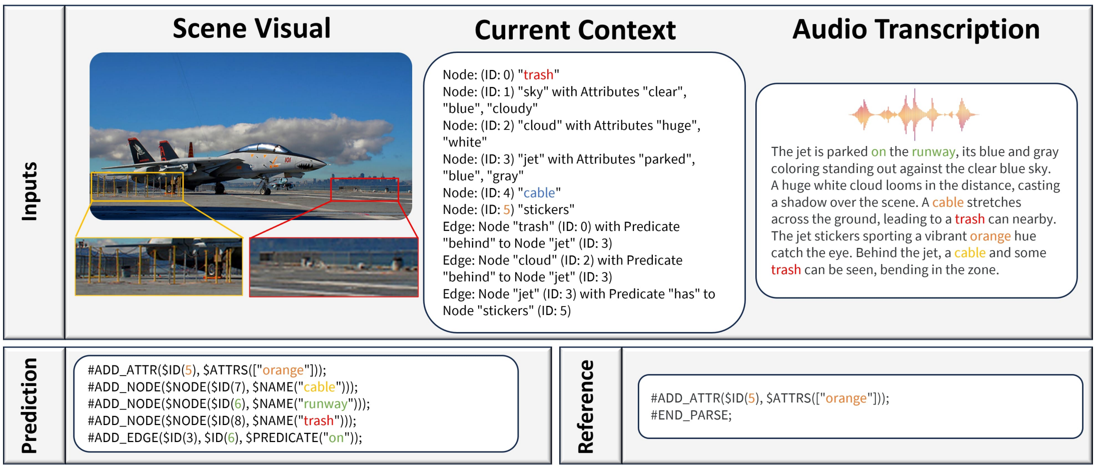
We include an example of a typical AViD-SP generation in Figure 3, with metric scores approximately at the average obtained across the full testing set. In this example it is evident that all of the ground truth reference information was successfully added to the updated scene graph, leading to the Soft-RED score of 0.0. However, considerable extraneous information is also observed to have been added. In Figure 3 three additional Nodes are added, with two of them being duplicates of ones that already exist in the scene graph, along with one Edge.

However, considering the Transcription and Visual Scene for the illustrated sample reveals that these features, while not included in the reference, likely are logically reasonable for the agent to include. For the additional Node of "runway" the motivation is obvious. Not only is the runway and its corresponding edge relationship mentioned by the LLM, but a runway is even present in the scene visual. Similar conditions apply to the two duplicate nodes added. While those nodes already exist, they are mentioned in the Audio Transcription at two distinct times. Inspection of the highlighted and blown-up parts of the image also reveals that there are in fact duplicates of these entities in the scene, making their addition to the updated context reasonable.

This is not to say all extraneous additions should be treated as correct since many should not. However, it does illustrate a key area to seek further improvement in the VG-SPICE dataset and why, for this work, we focus more on the "soft" capability to add all known good information tot he graph.

Model Type		H-GED↓				S-GED↓			
		0dB	2dB	20dB	Gold*	0dB	2dB	20dB	Gold*
AViD-SP + GMADS	Base	2.010	1.921	1.811	1.621	0.924	0.889	0.862	0.778
	w/o Image	2.044	1.973	1.816	1.642	0.944	0.923	0.878	0.791
	w/o Audio	2.168	2.101	2.071	1.863	1.209	1.186	1.158	1.004
	w Incorrect Image**	-	-	1.806	-	-	-	0.861	-
	w/o Prior Context***	-	-	4.656	-	-	-	0.909	-
AViD-SP + Meanpool	Base	1.739	1.617	1.514	1.295	0.935	0.889	0.859	0.759
	w/o Image	1.732	1.599	1.514	1.285	0.939	0.910	0.872	0.759
	w/o Audio	1.622	1.560	1.428	1.244	1.002	0.964	0.909	0.815
	w Incorrect Image**	-	-	1.517	-	-	-	0.857	-
	w/o Prior Context***	-	-	4.778	-	-	-	0.905	-

Table 5: GED results on the VG-SPICE test set for our AViD-SP model. AViD-SP was trained with CHiME5 noise augmentation sampled between 0db and 20dB SNR (all CHiME5 noise followed the provided train/eval/test splits). *Given the ground truth utterance transcripts in place of the ASR transcriptions. **Evaluated by offsetting visual features within batch so incorrect image features are paired with the other input components. ***Evaluated with "Empty Context" prior state scene graphs summaries instead of the correct ones.



Node: (ID: 0) "trash"
Node: (ID: 1) "sky" with Attributes "clear", "blue", "cloudy"
Node: (ID: 2) "cloud" with Attributes "huge", "white"
Node: (ID: 3) "jet" with Attributes "parked", "blue", "gray"
Node: (ID: 4) "cable"
Node: (ID: 5) "stickers"
Edge: Node "trash" (ID: 0) with Predicate "behind" to Node "jet" (ID: 3)
Edge: Node "cloud" (ID: 2) with Predicate "behind" to Node "jet" (ID: 3)
Edge: Node "jet" (ID: 3) with Predicate "has" to Node "stickers" (ID: 5)

The jet is parked on the runway, its blue and gray coloring standing out against the clear blue sky. A huge white cloud looms in the distance, casting a shadow over the scene. A cable stretches across the ground, leading to a trash can nearby. The jet stickers sporting a vibrant orange hue catch the eye. Behind the jet, a cable and some trash can be seen, bending in the zone.

```
#ADD_ATTR($ID(5), $ATTRS(["orange"]));
#ADD_NODE($NODE($ID(7), $NAME("cable")));
#ADD_NODE($NODE($ID(6), $NAME("runway")));
#ADD_NODE($NODE($ID(8), $NAME("trash")));
#ADD_EDGE($ID(3), $ID(6), $PREDICATE("on"));
```

```
#ADD_ATTR($ID(5), $ATTRS(["orange"]));
#END_PARSE;
```

Figure 3: Sample generation output with corresponding inputs from AViD-SP. Scored a Soft-RED of 0.0 and Hard-RED of 6.727. Significant features highlighted in colors. Qualitative evaluation reveals that the majority of extraneous additions were either supported by the Audio Transcription, the scene image, or both.

D Visual Genome Preprocessing

The Visual Genome serves as a strong basis for VG-SPICE but has quality issues such as inconsistent naming for Nodes, Attributes, and Predicates, duplicate Nodes, and unnecessary Nodes (e.g., **<man, has, head>**). Prior solutions for Scene Graph Generation (SGG) tasks (Liang et al., 2019; Zhang et al., 2019; Xu et al., 2017; Maëlic et al., 2023) curated versions by limiting predicates and node names, reducing predicates from 27k to 50 and node names from 53k to 150. While the Visual Genome contains a substantial portion of single-sample terms, typically of lower quality, such restrictions can oversimplify and yield smaller, less representative scene graphs.

Our approach refines the Visual Genome by:

Standardization and Correction: We applied rule-based systems with Sentence Transformer Part of Speech taggers [5] to fix inconsistencies and improve scene graph density by retaining rare Node names (e.g., "red table", identifying "red" as an attribute). We removed low-quality attributes and predicates by limiting them to specific parts of speech conditions, such as removing proper and common nouns from attributes/edges. Furthermore, we imposed several straightforward constraints to refine the scene graph structure. These included setting limits on the word counts for individual scene graph elements and consolidating attributes when redundancy was detected within a specific node, for instance, merging "reddish" and "red" when both attributes described the same entity.

Duplicate Node Elimination: We added a post-standardization phase to remove duplicate nodes. Unlike earlier methods (Maëlic et al., 2023) relying solely on a high Intersection over Union (IoU) threshold for exact node matches, we included a semantic similarity check from the contextualized embeddings from the same Sentence Transformer utilized in the Standardization and Correction phase. This allows for the detection of duplicate Nodes with significant name similarities and IoUs. With a preference for visually supported scene graphs over the potential exclusion of some valid Nodes, we set a lower IoU threshold (0.5, compared to prior works' 0.9) and a semantic similarity threshold of 0.7.

Term Frequency Analysis: Next, we manually curated terms in the filtered dataset to establish a relevant set for the SPICE task, excluding single-occurrence terms for their low quality, and filtered scene graphs based on this list.

Scene Graph Size Restriction: Finally, we filtered out small graphs to ensure a diverse set for VG-SPICE, excluding graphs with fewer than four Nodes or Edges and applying dynamically increased threshold for graphs with duplicate nodes.

These methods enhanced the Visual Genome's graphs, yielding a dataset with improved quality and annotation density, as illustrated in Table 1.

E Training and Inference Hyperparameters

The training regimen for AViD-SP spans two epochs across the dataset, using a combined batch size of 72 on six Nvidia L40 GPUs. An initial learning rate of 5×10^{-5} is applied, followed by exponential decay. We employ cross-entropy loss for the prediction of target semantic parses, introducing loss masking for padding and for the prompt that combines prior context with multimodal inputs. We utilize loss factors of $\alpha = 1.0$, $\beta = 0.1$, and $\gamma = 0.1$.

Inference leverages a greedy decoding strategy with a max generation length of 160 tokens and otherwise default generation parameters for LLAMA 2 7B.

F Contextual State Representation

SPICE formulates the prior context to be utilized by the agent as a structured knowledge graph. However, top-performing semantic parsing generation models, such as those best on the Llama architecture as used in this work, are decoder-only models that can accept inputs from linear text sequences only. This requires utilizing either a compatible knowledge graph encoder which can embed and project the knowledge graph representation for use by the semantic parse generation model, or representing the knowledge graph in the form of a textually formatted prompt. For AViD-SP developed in this work, we utilized the second, with the format of the textually prompted representation of the prior context shown in Figure 1.

When generating the context representations all existing Nodes are assigned Node IDs, and semantic parses are expected to operate in reference to these Node IDs (Section G). We provide Nodes

[5] Using "all-mpnet-base-v2" from Python Sentence Transformers

and Attributes first, followed by any Edges. The ordering of all information is sorted by Node ID in ascending order. In practice, all Node IDs are randomly assigned for each training iteration to diversity training inputs.

G Formal Language Definition

The formal language we used in the semantic parses P_i and the corresponding execution function e contained the following executable function, which together could deterministically update the scene graph prior context C_i to the next context state C_{i+1}. Since VG-SPICE only represents the conversational construction of scene graphs, and not deletion or alterations, our formal language is comprised of three distinct operations: 1) *#ADD_NODE* accepting a new Node ID, name, and optionally a set of attributes to add along with it, 2) *#ADD_ATTR* accepting an existing Node ID as well as a set of attributes to be added to the specified node, and 3) *#ADD_EDGE* accepting a source and target pair of existing node IDs along with the predicate to be assigned between them. Our formal language always generates reference semantic parses with new attributes added first, followed by new Nodes (and assigned attributes), and lastly new edges. However, when evaluating our model outputs the execution function e can accept these commands in any order, so long as the referenced node IDs already have been added.

H Licensing

Our paper utilized the Visual Genome dataset which is listed under a Creative Commons license. All other tools utilized are available from either Pythons Spacy or Huggingface and are available for academic use. To the best of our knowledge, all artifacts utilized are aligned with their intended use cases.

I AI Assistance

A minor portion of code development was done with the assistance of ChatGPT. All research ideas and writing are of the author's original creation. Grammarly was utilized for writing assistance.