Multimodal Contextualized Semantic Parsing from Speech

Anonymous ACL submission

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 and a novel multimodal fusion method, the Grouped Multimodal Attention Down Sampler (GMADS), within the AViD-SP model. These innovations aim to improve multimodal information processing and integration. Both the VG-SPICE dataset and the AViD-SP model are publicly available[1].

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 speech recognition 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/

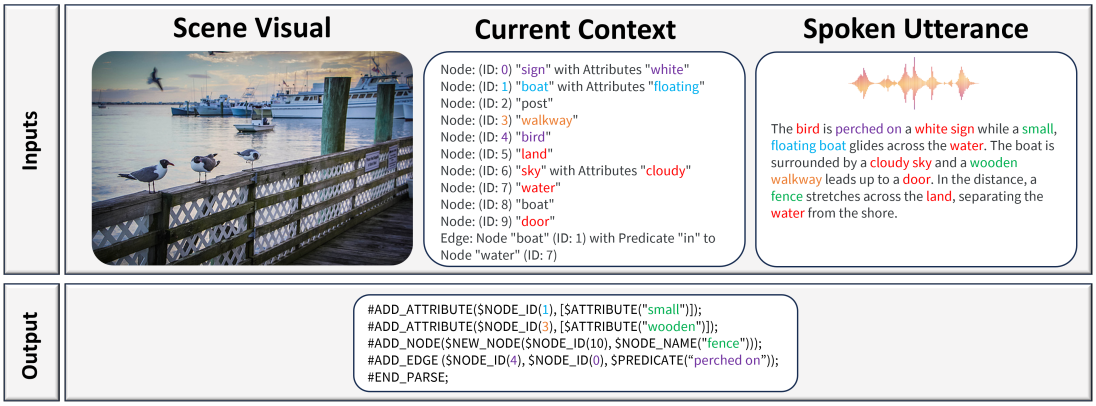

Figure 1: Example of VG-SPICE inputs as well as a plausible output to produce the correct next state context. New information that the agent is expected to add to the context is shown in green while already known information is noted in red. Grounding entities that have new information being added to 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 D).

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).

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 win-

dows (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.

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 historical 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, $P_i = a(F_i^m, C_i)$. 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 pre-

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 B.

vious 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 B 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 E.

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.

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.

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.

5 AViD-SP Model

To provide a foundational solution for VG-SPICE, our approach utilizes a range of pretrained models, specifically fine-tuned to boost SPICE-focused semantic parsing capabilities. Figure 2 provides an overview of our model architecture, named Audio-Vision Dialogue Scene Parser (AViD-SP). Central to our framework is the pretrained Llama 2 7B model (Touvron et al., 2023b). Despite using the smallest variant, the comprehensive pretraining of Llama 2 equips our model with solid functional capabilities, which are particularly advantageous for processing VG-SPICE's diverse semantic parses. However, Llama 2, being a model trained exclusively on textual data, does not inherently support the multimodal inputs typical of VG-SPICE.

To expand the allowed inputs we emulate previous research (Rubenstein et al., 2023; Gong et al., 2023; Lin et al., 2023) by projecting embeddings from pretrained modality specific feature extractors. This method has been shown to enable text-based LLMs to process information from various modalities. Nonetheless, directly incorporating these projected embeddings into the LLM's context window leads to significant computational overhead due to their typically long context lengths. While prior studies have often used pooling methods (Gong et al., 2023) to downsample embeddings by modality, this approach does not fully overcome the challenge of integrating diverse modalities embeddings for use by the LLM. For example, audio embeddings convey information at a finer temporal granularity compared to textual embeddings while the opposite may be true for vision embeddings, making the tuning of downsampling factors a difficult task. Moreover, even with optimized downsampling, pooled embeddings must maintain their original relative order and are limited to information from only the pooled region. Many applications might benefit from being able to establish downsampled features from both local and global features and to reorder these features to some extent.

To tackle these challenges, we introduce a novel Grouped Modality Attention Down Sampler (GMADS) module. This module projects embeddings from non-textual modalities into a unified, fixed-dimensional space and appends them with modality-specific positional embeddings. For our two-modality inputs, audio and visual, we collect all individual modality embeddings and a single cross-modality embedding formed from the concatenation of all modalities. This creates a group set of embeddings, with most corresponding to individual inputs and one representing all inputs combined. We apply learned sampling tokens to each of these embedding sequences, adding a sampling signifier token to every S embedding and a non-sampling signifier token to the subsequent $S-1$ indices. A series of self-attention layers then processes each embedding sequence, retaining only the output indices marked with a sampling signifier. A linear projection adjusts the outputs to the dimensionality of the Llama 2 7B decoder, and all embedding sequences are concatenated. This process results in an embedding output that is downsampled by a factor of $S/2$. All weights in the GMADS module are shared across the groups, significantly reducing the parameter size.

The GMADS module offers several advantages over providing all raw modality embeddings directly to the LLM decoder or using traditional pooling. First, GMADS operates at lower dimensional scales than the pretrained LLM, which, despite needing to attend to the concatenation of all modality inputs, greatly reduces memory costs. Furthermore, the modality inputs do not require autoregressive generation to be done alongside those inputs, thereby saving additional memory. The output contexts will be only $2/S$ the size when reaching the more resource-intensive LLM decoder. Second, GMADS enables the model to selectively learn its downsampling process, including deciding whether to focus near each sampling index or incorporate longer-range features, allowing some level of information restructuring. The inclusion of cross-modality encoding allows parts of the downsampled embeddings to capture valuable information across modalities. However, having individual modality components in the outputs ensures that a portion of the output embeddings will be conditioned on each modality, requiring the attention mechanisms to be sensitive to all modalities.

For feature extraction, we employ the following: For visual inputs, we adopt the strategy from (Lin et al., 2023) and utilize a mix of visual encoders for distinct task-specific embeddings, but utilizing more recent variants then (Lin et al., 2023). These include the encoder portion of DINOv2 (Oquab et al., 2024), the visual components of CLIP-ViT (Radford et al., 2021) and ConvNeXt V2 (Woo et al., 2023), and the visual encoder of BLIP-2 (Li et al., 2023). For audio, we use the encoder part of Whisper-Large V3 (Radford et al., 2022).

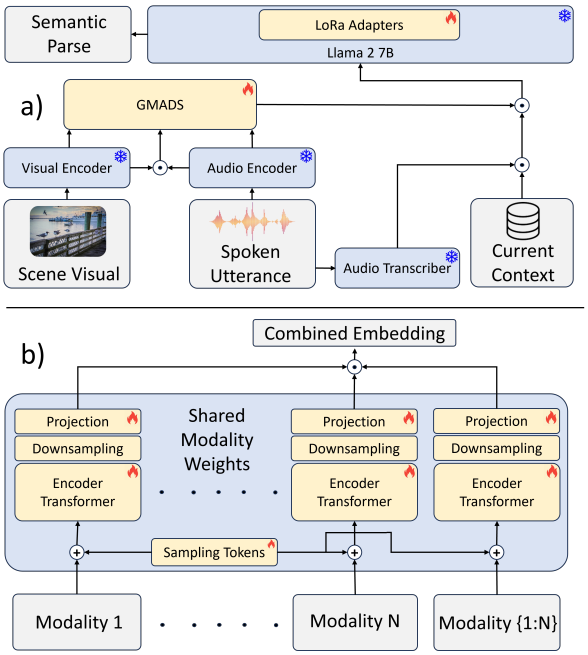

Figure 2: a) The architecture of the AViD-SP model for VG-SPICE, integrating pretrained encoders and large language models (LLMs) with LoRa adapters and feature fusion modules. Trained and frozen segments of the model are denoted by fire and snowflake icons, respectively. b) Our novel Grouped Modality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

All irrelevant portions of the pretrained models are discarded, retaining only the utilized encoder portions. In line 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 Whiser-Medium.en model.

To enable scalable fine-tuning, we incorporate LoRa adaptation layers to Llama 2 7B and freeze all pretrained feature extractors. As a result, our model encompasses over 11.5 billion parameters but maintains a manageable trainable parameter count of 146 million.

5.1 AViD-SP Training and Evaluations

We train AViD-SP using the cross entropy loss between the predicted and reference Formal Semantic Parses. However, we observed that in cases where the ASR transcriptions were moderately reliable, the model learned to rely on them too much, leading to the GMADS outputs severely collapsing. To avoid this we apply an additional orthogonality loss to the outputs of the GMADS module. This loss attempts to minimize the inverse cosine similarity between distinct samples in an input batch, therefore preventing embedding collapse and promoting discriminative ability between the various samples. We began with the orthogonality loss weighted high until the model was capable of reaching a per batch average cosine similarity of 0.6, after which we reduced the weighting significantly to allow the training to focus on semantic parsing performance. Our full loss function is shown below, where p is the softmax predictions for P_i, t is the ground truth labels, E_o is the GMADS output embedding for index o of a tensor with a batch size of B.

$$\ell_{CE} = -\sum_{k=1}^{n} t_k log(p_k)$$

$$\ell_{Ortho} = 1 - \frac{2}{B(B-1)} \sum_{i=1}^{B-1} \sum_{j=i+1}^{B} \frac{E_i * E_j}{\|E_i\| * \|E_j\|}$$

$$L = \alpha * \ell_{CE} + \beta * \ell_{Ortho}$$

AViD-SP utilizes a six-layer self-attention transformer as part of its GMADS module, each with an embedding dimensionality of 1024 and eight attention heads. The GMADS module is constructed to utilize a downsampling factor, S, of 32. Additionally, we enhance the Llama 2 7B model's key, query, and value layers with a rank 64 Low-Rank Adaptation (LoRa). No hyperparameter search was done to optimize these settings.

We train AViD-SP by integrating randomly sampled CHiME5 noise to induce audio corruption, adding this noise at various Signal-to-Noise Ratios (SNR) of 0, 5, 10, or 20dB. Additional details on training and inference hyperparameters are discussed in Section C. Due to computational constraints, AViD-SP was trained for only a single epoch on VG-SPICE and did not exhibit signs of having fully converged. Further details on AViD-SP checkpoints and evaluations will be made available in the code repository.

We evaluate the AViD-SP model across multiple ablation evaluations to assess the impact of different features. These studies include scenarios with and without visual modality inputs, with CHiME5 noise additions at -2, 0, and 20dB, when gold transcription labels are provided to the model, with mismatched images, without access to previous scene graph context, and finally, when the model operates without explicit ASR integration.

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 E). 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 weak metric and qualitatively show in Section 6 how updated utterances accommodate these extraneous additions.

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 normalized by the edit distance needed to transform the prior context into the reference, so the result can be viewed as the percentage information error. GED is particularly sensitive to 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 sentence transformer model[2] for semantic similarity, 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 in the same manner as GED and so numerically can be interpreted as the percentage of missing and/or extra information.

[2]The "en_stsb_roberta_base" model from https://github.com/MartinoMensio/spacy-sentence-bert

6 Results

The results for AViD-SP on the VG-SPICE test set, as depicted in Table 3, illustrate that the baseline AViD-SP model achieves Soft metrics around 0.4. This indicates a 60% effectiveness in identifying and incorporating desired information into the scene graph. However, the Hard metrics highlight the introduction of significant irrelevant information. A deeper analysis and qualitative example examines the reasons behind this, including why high Hard edit distances may be desirable for VG-SPICE to a degree, are discussed in Section A.

The performance of the AViD-SP model under various SNR conditions exhibits slight performance degradation at low SNR levels. This suggests that, although background noise adversely affects the model's performance, it remains generally resilient to moderate levels of noise. Moreover, the provision of gold transcripts significantly enhances parsing accuracy, underlining the advantages of accurate ASR (Automatic Speech Recognition) transcriptions for our model.

Experiments conducted without visual inputs underscore their indispensable role. The omission of images leads to increased GED and RED, affirming that visual context is critical for precise semantic parsing. Additionally, the utilization of incorrect images still showcases similar performance decreases, attributing gains not only to maintaining AViD-SP's visual pathways but also to accessing relevant visual features.

The lack of prior context notably raises error rates, highlighting the critical role of historical context in enabling the model to update the scene graph accurately. Similarly, the complete removal of ASR transcriptions emphasizes the importance of textual information from audio inputs for semantic parsing. Even if this information is processed through a secondary pathway in the model, its absence markedly

Model Type	Hard-GED	Soft-GED	Hard-RED	Soft-RED
AViD-SP				
-2dB SNR Noise	3.626	0.458	3.537	0.441
0dB SNR Noise	3.530	0.448	3.453	0.419
20dB SNR Noise	3.382	0.427	3.343	0.384
Gold Transcripts*	3.326	0.410	3.186	0.308
AViD-SP w/o Image				
-2dB SNR Noise	1.496	0.605	1.484	0.628
0dB SNR Noise	1.564	0.590	1.474	0.615
20dB SNR Noise	1.510	0.575	1.539	0.589
Gold Transcripts*	1.503	0.563	1.454	0.545
AViD-SP w Incorrect Image**	1.510	0.575	1.540	0.589
AViD-SP w/o Prior Context***	3.313	0.583	4.110	0.509
AViD-SP w/o ASR Transcription	3.719	0.626	3.962	0.861

Table 3: Full results on the VG-SPICE test set for our AViD-SP model. All results in utilize the same pretrained model trained with vision, ASR, context, and audio components. AViD-SP was trained with CHiME5 noise augmentation sampled at 0, 5, 10, and 20dB SNR (-2dB is OOD from training, and 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 withing 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.

detracts from performance, likely due to the high cost of training robust audio processing models.

7 Limitations and Future Work

VG-SPICE and AViD-SP have limitations. 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 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.

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.

8 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 our novel Grouped Multimodal Attention Down Sampler (GMADS), demonstrates the potential for improving 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.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

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, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

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

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*.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders.

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 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.

B 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 [3] 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.

C Training and Inference Hyperparameters

The training regimen for AViD-SP spans two epochs across the dataset, using a combined batch

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

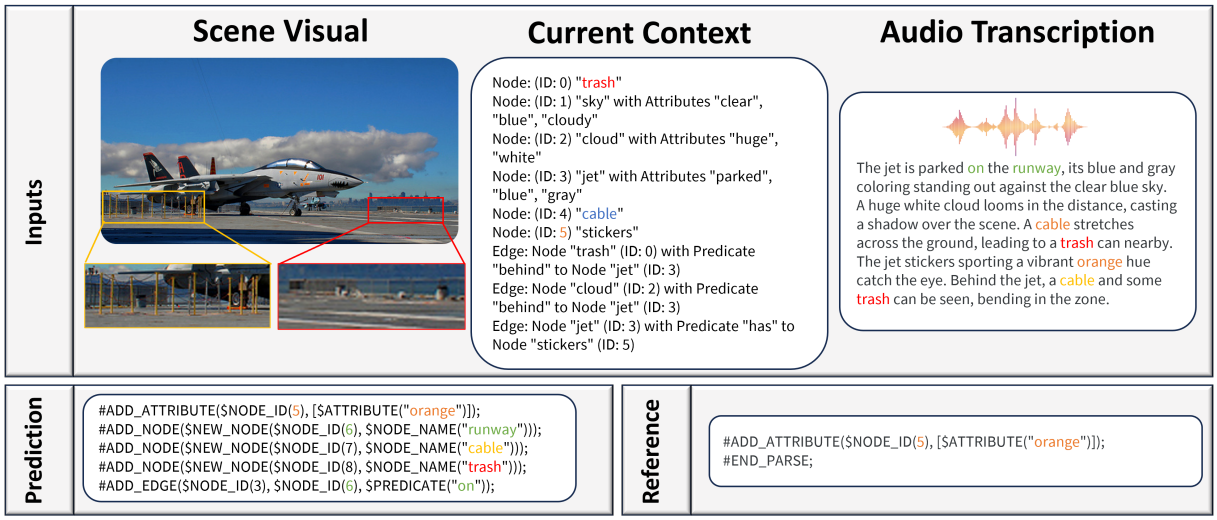

Figure 3: Sample generation output with corresponding inputs from AViD-SP. Scored a Soft-RED of 0.0 and Hard-RED of 6.727. Significant features highlighted in colors. Qualitative evaluation reveals that the majority of extraneous additions were either supported by the Audio Transcription, the scene image,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

size of 72 on six Nvidia L40 GPUs. An initial learning rate of 5×10^{-4} 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.

Inference leverages a group beam-search strategy, employing two beams across two beam groups. We impose a group diversity penalty of 5.0 and a repetition penalty of 1.2 to enhance the variety and uniqueness of the generated semantic parse tokens, with a cap set at 160 tokens. Note that the selection of max tokens is significant for AViD-SP, since more allowed tokens were generally observed to lower the Soft metrics (as more features were allowed to be added) while lower maximum generation tokens raised the Soft metrics but lowered the Hard ones.

D 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 E). We provide Nodes 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.

E 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_ATTRIBUTE* 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, fol-

lowed 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.

F 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.

G 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.