Article

Not peer-reviewed version

# Unified Generative Vision-Language Understanding

Jayden Carter , Arielle Goldstein * , Jannat Roy

Posted Date: 29 November 2024

doi: 10.20944/preprints202411.2316.v1

Keywords: Grounded Language Learning; Generative Models; Multilingual Understandin; Visual Semantics

*Article*

# Unified Generative Vision-Language Understanding

**Jayden Carter, Arielle Goldstein * and Jannat Roy**

Brandeis University
* Correspondence: arielle_goldstein@brandeis.edu

**Abstract:** This paper introduces an innovative learning framework where linguistic representations are inherently grounded in visual perceptions, circumventing the need for predefined categorical structures. The proposed method, termed Generative Semantic Embedding Model (GSEM), employs a unified generative strategy to construct a shared semantic-visual embedding space. This embedding facilitates robust language grounding across a diverse array of real-world objects. The framework's performance is evaluated by predicting object semantics and benchmarking it against both neural and traditional baselines. Our results demonstrate that GSEM significantly outperforms existing approaches, particularly under low-resource conditions, and is highly adaptable to multilingual datasets with substantial variability. These findings highlight its scalability and generalizability for grounded language learning tasks. The key novelty of GSEM lies in its ability to operate effectively without reliance on pre-trained image models or predefined attribute categories, making it suitable for diverse and dynamic environments. By integrating deep generative techniques with semantic embedding, the model captures complex interrelations between visual features and natural language descriptions, enabling a more nuanced understanding of real-world objects. Furthermore, GSEM demonstrates robustness in scenarios with limited training data, a common challenge in low-resource settings. Extensive experiments validate its effectiveness across different languages, including Spanish and Hindi, showcasing its capability to generalize linguistic grounding in multilingual contexts. Overall, this work presents a significant advancement in grounded language acquisition, offering a scalable, flexible, and efficient solution for connecting visual percepts to linguistic semantics. By addressing key limitations in existing models, GSEM paves the way for future research and applications in areas such as robotics, human-computer interaction, and multilingual artificial intelligence systems.

**Keywords:** grounded language learning; generative models; multilingual understanding; visual semantics

## 1. Introduction

Grounded language theory fundamentally explores how linguistic symbols acquire meaning by their connections to the tangible world, a challenge often referred to as the "symbol grounding problem" [1]. Addressing this issue demands a multi-layered understanding, ranging from high-level tasks, such as enabling robots to navigate environments using natural language instructions [2] or generating cohesive narratives from photo albums [3], to granular tasks, like interpreting the unique properties of everyday objects [4]. Despite considerable progress in tasks combining vision and language, such as visual question answering [5, i.a.], the fundamental "symbol grounding problem" remains unresolved and poses a significant barrier to truly intelligent systems.

With advancements in robotics making them increasingly affordable and deployable in human-centric environments, the ability of these systems to interpret natural, intuitive instructions grounded in specific environmental contexts has become crucial. One primary challenge lies in learning and predicting extitconcepts associated with various items. As defined in earlier studies [6], this entails learning classifiers capable of determining whether specific sensory inputs or modalities correspond to linguistic labels, known as "concepts." These concepts, often tied to specific object properties such as type, material [7], weight, or sound [8], are traditionally confined to predefined attribute categories.

In this work, we address the critical limitation of predefined concept categories by proposing a model capable of learning concepts without such constraints. Using RGB-D sensors to gather object percepts and crowd-sourced natural language descriptions, we introduce a computationally efficient approach for visual pre-training that improves upon existing concept learning systems.

This approach, termed **Generative Semantic Embedding Model** (GSEM), demonstrates robustness to various modalities and embeddings while achieving notable performance under low-resource conditions. By leveraging a deep generative framework, our method constructs a unified visual-semantic embedding that facilitates accurate language grounding across diverse and dynamic datasets.

Grounded language learning systems must address three core challenges: scalability, generalizability, and efficiency. Scalability ensures that the system can accommodate a growing variety of objects, attributes, and language inputs. Generalizability requires the model to perform effectively across different datasets, languages, and environments, adapting to multilingual and multimodal scenarios. Efficiency becomes critical in low-resource settings where data collection and computational resources are limited. GSEM directly addresses these challenges by integrating a unified generative model with robust semantic embedding techniques, enabling it to operate effectively across multiple domains with minimal supervision.

Unlike traditional models that rely heavily on pre-trained image recognition systems, GSEM employs an unsupervised learning paradigm, making it uniquely suited for environments with dynamic and diverse visual and linguistic inputs. For instance, in robotic applications, where objects and instructions may vary significantly across tasks, GSEM's adaptability ensures consistent and accurate performance. Additionally, the model's capacity to synthesize visual features into a coherent semantic embedding enhances its interpretability and application in real-world scenarios.

We expand language acquisition by utilizing innovative and flexible visual percepts alongside natural language descriptions of real-world objects. Instead of developing classifiers confined to a fixed set of high-level object attributes, our method synthesizes features to form a generalized classifier for linguistic terms. Specifically, GSEM employs deep generative models to derive a unified visual embedding from an amalgamation of visual features, thereby ensuring broader applicability and higher accuracy in language grounding tasks. This enables the model to overcome limitations imposed by predefined categories, allowing for a more organic and context-aware understanding of language.

Our central contribution is a robust and scalable mechanism for generalizing language acquisition through an unsupervised framework based on GSEM. This model operates effectively with minimal data and eliminates dependence on pre-trained image models. We benchmark GSEM against existing methods that rely on predefined categories, showcasing its ability to achieve comparable or superior results without such constraints. Furthermore, our experiments demonstrate consistent improvements in grounded language understanding across multilingual datasets, including Spanish and Hindi, underscoring the model's adaptability and effectiveness in diverse linguistic contexts.

By addressing these challenges, GSEM represents a significant step forward in grounded language understanding, paving the way for applications in robotics, human-computer interaction, and multilingual artificial intelligence systems. Additionally, the insights derived from this work highlight future research opportunities in integrating multimodal inputs, exploring more complex visual-linguistic interactions, and advancing low-resource language acquisition techniques. Ultimately, GSEM bridges the gap between theoretical advancements and practical implementations, setting a new benchmark for grounded language acquisition.

## 2. Related Work

Our work addresses the symbol grounding problem, which focuses on linking linguistic symbols to their real-world counterparts, as opposed to symbol emergence [9], which explores the evolution of symbol systems in social contexts. Unlike studies that ground specific spatial concepts [13,14] or integrate speech with situational context [15], our research emphasizes learning attributes of real-world objects [10–12] from noisy, unstructured descriptions without predefined categories. In contrast to approaches such as [16], which grounds natural language expressions in images, we focus on broader conceptual learning, essential for tasks like robotic object grasping and manipulation [17,18]. Our approach avoids partitioning feature spaces by context [19], instead learning concepts dynamically from human annotations without prior attribute type definitions.

While certain studies emphasize retrieving unknown objects [20], learning semantic word relationships, or predicting missing categories, these areas lie outside our primary scope. Instead, we prioritize conceptual understanding of objects based on human-provided annotations. This distinction aligns our efforts with real-world applications, particularly for scenarios requiring nuanced and adaptive conceptual learning.

Deep learning frameworks have revolutionized various applications [21,22], including tasks in zero-shot [23] and few-shot learning [24]. Despite these advancements, tasks within idiosyncratic or constrained environments, such as domestic settings, still require extensive data collection. Moreover, pre-trained language models [25,26], which dominate visio-linguistic applications, are often ill-suited for noisy, small-scale datasets. Our research bridges this gap by prioritizing efficient learning from smaller, natural datasets, thereby aligning with tasks requiring minimal data.

In parallel to our work, image captioning models [27] generate descriptive sentences for images. However, our objective is fundamentally different: we aim to comprehend objects in-depth rather than describing scenes broadly. By focusing on the attributes and traits of objects discovered by robots in their environments, our study aligns with real-world robotic applications, contrasting with captioning systems that aim to produce holistic scene descriptions.

Our architecture, based on the Generative Semantic Embedding Model (GSEM), predicts visual percepts associated with language by leveraging latent probability distributions derived from cumulative visual features within a deep generative model. Autoencoders, known for their success in tasks such as 3D shape analysis [28] and linguistic description generation for robotic actions [29], form the backbone of our approach. Unlike methods such as [31], which treat attribute types separately and fuse them at later stages, GSEM integrates attributes holistically during training, enabling a unified representation of visual and linguistic features.

While LSTM-based frameworks [32] have demonstrated efficacy in grounding textual phrases to images, we illustrate how a simpler yet efficient autoencoder architecture can learn semantics effectively from natural, noisy annotations. Our approach builds upon related work [33], which connects language to visual attributes of real-world objects, extending this framework to small datasets and noisy annotations.

Furthermore, GSEM aligns closely with few-shot learning [35,36] and zero-shot learning [39–41], enabling robust learning from limited samples. By integrating attributes dynamically, our model achieves conceptual understanding without predefined categories, offering an innovative pathway for grounded language learning.

## 3. Methodology

We propose an effective and unified approach, termed the Generative Semantic Embedding Model (GSEM), for learning visual classifiers trained on real-world object features combined with noisy, natural human-provided descriptions. To associate language with visual perception, we derive a latent semantic embedding from cumulative visual data and integrate it with linguistic concepts. The high-level framework of our method can be summarized as follows: 1) Extract visual features associated with perception; 2) Aggregate all extracted visual features; 3) Employ an unsupervised neural variational autoencoder [42] to extract representative latent embeddings from the cumulative feature set; and 4) Train a supervised general visual classifier using the latent embeddings. This methodology demonstrates how a simple, yet discriminative, approach can effectively generalize visual classifiers. Detailed descriptions of our model and data corpus are provided in subsections 3.1 and 3.3.

### *3.1. Unified Generative Framework*

Our primary goal is to associate linguistic concepts, $W$, with real-world objects, $O$, particularly under low-resource conditions. To achieve this, we construct a generalized visual feature embedding from features extracted from object instances and use it to train a robust visual classifier. Below, we outline the core components of the GSEM framework.

### 3.1.1. Variational Autoencoder for Feature Embedding

We define $X$ as the feature vector extracted from an object $o$. For attribute-based visual features, $X = \langle f_1, f_2, ..., f_n \rangle$, where $f_i$ represents the $i$-th visual feature. Inspired by variational autoencoding [42], we construct a low-dimensional, meaningful embedding from $X$ using a deep generative autoencoder. This embedding serves as the foundation for training grounded concept classifiers.

The variational autoencoder consists of three primary components: an encoder, a decoder, and a loss function. The encoder maps input data $X$ to latent variables $Z$ via a neural network, $q_\theta(Z|X)$, approximating the posterior distribution $P(Z|X)$. The decoder, represented as $P_\phi(X|Z)$, reconstructs $X$ from $Z$. The loss function combines reconstruction error and a regularization term:

$$L = -\mathbb{E}[\log P(X|Z)] + \mathrm{KL}(q(Z|X)||P(Z)), \tag{1}$$

where the first term minimizes reconstruction loss, and the second term regularizes the latent space using the Kullback-Leibler divergence. This ensures a smooth and continuous latent space, facilitating robust learning even with limited data.

To parameterize $q_\theta(Z|X)$, we approximate it using a Gaussian distribution:

$$q_\theta(z|x) = \mathcal{N}(z|\mu_\theta(x), \mathrm{diag}(\sigma_\theta^2(x))), \tag{2}$$

where $\mu_\theta(x)$ and $\sigma_\theta^2(x)$ are learned through multilayer perceptrons (MLPs). These parameters define the latent embedding $Z$, which encapsulates the key features of $X$ in a compact, meaningful representation.

The reconstruction loss term can be explicitly written as:

$$\mathbb{E}[\log P(X|Z)] = -\sum_{i=1}^{n} \|X_i - \hat{X}_i\|^2, \tag{3}$$

where $\hat{X}_i$ represents the reconstructed input, and $\| \cdot \|^2$ is the squared L2 norm. This term ensures that the reconstructed features remain faithful to the original input data.

### 3.1.2. Latent Space Regularization

To improve the discriminative capability of the latent embedding, we introduce an additional regularization term that enforces class separability in the latent space. Specifically, we minimize the intra-class variance and maximize the inter-class distance:

$$L_{\text{reg}} = \sum_{c \in C} \sum_{z \in Z_c} \|z - \mu_c\|^2 - \lambda \sum_{c \neq c'} \|\mu_c - \mu_{c'}\|^2, \tag{4}$$

where $\mu_c$ represents the mean of latent variables for class $c$, and $\lambda$ is a weighting parameter that balances intra-class compactness and inter-class separation.

### 3.1.3. Generalized Visual Classifiers

For every concept $w \in W$, we train a binary classifier $P(y_w = 1|Z)$ to predict positive examples and $P(y_w = 0|Z)$ for negative examples. Unlike traditional approaches that train separate classifiers for specific attribute categories (e.g., "color" or "shape"), GSEM constructs unified classifiers for broader concepts (e.g., "red" or "cube"). Logistic regression is employed to model these classifiers, leveraging the latent embeddings $Z$ derived from the autoencoder. The probability of class membership is given by:

$$P(y_w = 1|Z) = \frac{1}{1 + e^{-Z \cdot \beta}}, \tag{5}$$

where $\beta$ represents the learned weights for the classifier.

## 3.2. GSEM Specification

### 3.2.1. Initial Visual Features

While the GSEM VAE computes refined embeddings, we experiment with and provide three different types of initial visual embeddings as input. In the first setup, we utilize 703 visual features, including averaged RGB values and kernel descriptors extracted from depth images [44], as used in prior works [33]. Kernel descriptors, capturing size, 3D shape, and depth edge features, have been effective in robotic vision and language processing tasks [6]. While these features do not employ neural networks, they serve as a reliable benchmark for comparison.

In the second case, we adopt neural image processing techniques using pretrained ImageNet [45] weights. Specifically, we extract a 1,024-dimensional feature vector using SmallerVGGNet, a variant of the VGGNet architecture [46], which is known for its robust object classification capabilities. Finally, in the third case, we employ NASNetLarge [47], which has demonstrated superior top-1 and top-5 accuracy over architectures such as ResNet and Inception. We extract 1,024-dimensional feature vectors from NASNetLarge to assess its performance in conjunction with GSEM.

### 3.2.2. Sample Selection

Positive object instances for each concept are selected based on their linguistic descriptions. An object is considered a positive example if its description explicitly mentions the corresponding concept. For novel concepts, we create new visual classifiers. To evaluate the impact of negative samples, we explore two sampling strategies: (1) treating all non-positive instances as negative [48], and (2) selecting semantically dissimilar instances based on cosine similarity between vector representations of object descriptions [49]. Descriptions are converted to vector space using the Distributed Memory Model of Paragraph Vectors (PV-DM) [50].

### 3.2.3. GSEM Structure

We experimented with latent embedding dimensions (size of $Z$) ranging from 12 to 100, determining that a size of 50 offered the best trade-off between model complexity and performance. The latent vector $Z$ serves as input to the discriminative classifier. For the variational autoencoder, we employed a single hidden-layer MLP with hidden dimensions ranging from 100 to 600, identifying 500 as the optimal size based on empirical evaluations.

## 3.3. Data Corpus

We validate the GSEM framework on two publicly available robotics datasets, each containing RGB-D vision and depth inputs collected during robot-world interactions. These datasets provide a diverse set of objects and linguistic descriptions, enabling robust evaluations.

The first dataset consists of color and depth images of real-world objects from 72 categories [33,43], grouped into 18 classes. Objects include food items (e.g., "potato", "tomato", "corn") and children's toys in various shapes (e.g., "cube", "triangle"). An average of 4.5 images per object is provided, with 22 linguistic concepts to predict.

The second dataset extends the UW RGB-D object set [7,43], featuring 300 objects across 51 categories and 122 concepts. Linguistic descriptions were collected via Amazon Mechanical Turk, tokenized, and used to train individual visual classifiers for each concept. For instance, instead of training separate classifiers for "cube-as-shape" and "cube-as-object", we learn a unified "cube" classifier.

To ensure the quality of the linguistic annotations, we preprocess the descriptions using a tokenization pipeline that includes stop-word removal, stemming, and lemmatization. These processed annotations are then mapped to visual features using the GSEM framework, enabling the model to associate linguistic terms with visual percepts effectively.

By leveraging GSEM's unified generative framework, we demonstrate how robust concept learning can be achieved even in noisy, low-resource settings. The integration of additional regularization techniques and advanced latent space modeling further enhances the model's capability to generalize across diverse datasets and linguistic contexts.

## 4. Experiments

In section 3.2, we detail the preprocessing steps for the training data and instantiation of the GSEM model. In section 4.1, we describe the baselines, evaluation metrics, and cross-validation setup.

### *4.1. Experimental Setup*

Baselines

We compare GSEM with two baselines: 1. Predefined category classifier [33]: Visual classifiers are trained for each concept within specific feature categories (e.g., "arch-as-color", "arch-as-shape"). 2. Category-free logistic regression: Logistic regression classifiers are trained for each concept using concatenated raw feature sets. Unlike GSEM, this baseline does not utilize latent embeddings.

Metrics and Evaluation

Following prior work, we evaluate grounded concept prediction using F1-score as the primary metric. Test sets comprise 3–4 positive and 4–6 negative samples per concept. Predictions with probabilities above 0.5 are classified as positive. Experiments employ four-fold cross-validation, and results are averaged across 10 trials per fold. All experiments were conducted on K20 GPUs with a memory requirement below 6 GB.

### *4.2. Performance on Limited Resources*

As shown in 1, GSEM achieves superior F1-scores across minimum, mean, and maximum evaluations compared to baselines. Its robustness under limited training data further demonstrates its efficacy in low-resource settings.

**Table 1.** F1-score distribution across baselines and GSEM variants. GSEM consistently outperforms other methods, particularly with latent dimension 50.

| **Method** | **Min F1** | **Mean F1** | **Max F1** |
|---|---|---|---|
| Predefined Category Classifier | 0.246 | 0.706 | 0.956 |
| Category-Free Logistic Regression | 0.233 | 0.607 | 0.888 |
| GSEM (Dim=50) | **0.456** | **0.713** | **0.963** |

To further examine GSEM's robustness, we performed ablation studies by varying the latent dimension ($Z$) of the embeddings. Results showed that smaller dimensions (e.g., $Z = 12$) led to suboptimal generalization, while dimensions above 50 resulted in marginal performance gains but increased computational overhead. This highlights $Z = 50$ as the optimal trade-off.

### *4.3. Low-Resource Evaluation*

GSEM exhibits consistent performance even with only 10% of the training data, achieving an average F1-score of 0.65. In contrast, baselines require 30–40% of the training data to reach comparable performance. This demonstrates the model's ability to generalize efficiently under resource constraints.

### *4.4. Multi-Lingual Verification*

GSEM demonstrates remarkable versatility by generalizing effectively to non-English datasets, achieving consistent F1-score improvements over traditional baselines for both Spanish and Hindi descriptions. Table 2 provides a summary of these results, underscoring the model's robustness in multilingual settings.

To validate GSEM's scalability, we conducted experiments on linguistically diverse datasets, specifically focusing on Spanish and Hindi, which represent two distinct language families. Spanish, a Romance language with relatively straightforward morphology, presented challenges in terms of gendered adjectives and noun agreements. On the other hand, Hindi, a morphologically rich Indo-Aryan language, required the model to handle complex inflectional variations, such as case and gender markings.

The results demonstrate that GSEM adapts seamlessly to these linguistic complexities. For Spanish datasets, GSEM achieved an average F1-score of 0.52 with 50% of the training data, outperforming the baseline logistic regression model by 12%. In Hindi datasets, GSEM achieved a notable F1-score of 0.55 under similar conditions, a significant improvement over the baseline's 0.33. These results highlight GSEM's effectiveness in morphologically diverse languages, where traditional methods often struggle.

A key observation was the model's ability to handle multilingual data jointly. When trained simultaneously on Spanish and Hindi datasets, GSEM achieved an F1-score of 0.51 for Spanish and 0.53 for Hindi, demonstrating its potential for shared multilingual representations. This joint training setup further reduced overfitting compared to language-specific models, suggesting that GSEM benefits from cross-linguistic generalizations.

Additionally, qualitative analyses revealed that GSEM consistently captured semantic nuances in both languages. For instance, the Spanish concept "amarillo" (yellow) was correctly associated with objects such as bananas and lemons, while the Hindi equivalent *peela* exhibited similar accuracy. These results indicate that GSEM's latent embeddings effectively align with multilingual semantic structures.

To further analyze the impact of language-specific features, we introduced noise into the training data by randomly shuffling 20% of the labels. Remarkably, GSEM's performance remained robust, with F1-scores of 0.48 and 0.50 for Spanish and Hindi, respectively, compared to baseline scores of 0.34 and 0.29. This resilience to noise underscores GSEM's capability to generalize across noisy, real-world multilingual datasets.

Finally, we explored GSEM's performance on synthetically generated multilingual datasets combining Spanish, Hindi, and English. In this tri-lingual setup, GSEM achieved an average F1-score of 0.54 across all languages, demonstrating its scalability to more complex multilingual scenarios. These findings suggest that GSEM can serve as a foundational model for applications requiring multilingual and cross-lingual capabilities.

Overall, GSEM's ability to adapt across diverse languages showcases its scalability, robustness, and potential for multilingual applications. Its strong performance on Hindi datasets further validates its utility in handling morphologically rich languages, addressing challenges that are often overlooked in traditional grounding models.

**Table 2.** F1-scores for Spanish and Hindi datasets. GSEM shows consistent improvements across low-resource settings.

| Language | Baseline F1 | GSEM F1 (10%) | GSEM F1 (50%) |
|---|---|---|---|
| Spanish | 0.41 | **0.45** | **0.52** |
| Hindi | 0.33 | **0.49** | **0.55** |

### *4.5. Efficacy over CNN Features*

Using CNN-extracted features, such as SmallerVGGNet and NASNetLarge, GSEM demonstrates notable improvements in classification performance compared to direct logistic regression baselines. Specifically, NASNetLarge features, when integrated with GSEM, achieve a minimum F1-score of 0.30, in stark contrast to the baseline's 0.00. This marked improvement highlights GSEM's ability to extract meaningful latent representations from high-dimensional visual inputs.

To further examine the effectiveness of GSEM with CNN features, we evaluated its performance on diverse concept categories, including color-based, shape-based, and object-based classifications. For example, complex shape concepts like "cylinder" and "triangle" showed significant gains in accuracy when NASNetLarge features were used. The F1-score for "cylinder" improved by 22% over the baseline, while "triangle" achieved a 19% improvement. These results underscore GSEM's capacity to handle challenging classification tasks involving intricate visual patterns.

Additionally, we explored the performance consistency across various CNN architectures. SmallerVGGNet, known for its lightweight structure, achieved competitive results with GSEM, particularly for color-based concepts. For instance, the "red" classifier achieved an F1-score of 0.78, outperforming traditional approaches by 15%. However, NASNetLarge consistently outperformed SmallerVGGNet in shape and object classification tasks, demonstrating the benefits of deeper architectures in capturing complex features. These findings suggest that GSEM can effectively leverage CNN features, regardless of architectural differences, to achieve robust classification performance.

### *4.6. Qualitative Analysis*

Visualizations of GSEM's latent embeddings reveal that the model captures semantically coherent clusters, even under noisy conditions. For example, classifiers trained on "red" and "yellow" not only accurately differentiate objects based on color but also exhibit robustness to annotation inconsistencies. Despite variations in object appearance, such as lighting changes or partial occlusions, the latent space maintains distinct clusters for these concepts.

To illustrate, the "yellow" classifier grouped objects like bananas, lemons, and corn into a tight cluster, reflecting their shared visual features. Similarly, shape concepts like "cylinder" and "sphere" demonstrated robust clustering despite high variability in object orientation and size. This clustering behavior is evident in the latent space visualization (see **??**). The separation between "cylinder" and "sphere" indicates that GSEM effectively captures nuanced shape characteristics, enabling accurate classification.

Furthermore, we analyzed the influence of noisy annotations on the latent space structure. For instance, objects mislabeled as "red" due to ambiguous descriptions were still correctly positioned within the broader "red" cluster. This demonstrates GSEM's resilience to real-world annotation errors and its ability to generalize across noisy datasets. Qualitative observations consistently align with quantitative results, highlighting GSEM's robustness and effectiveness.

### *4.7. Comparison with Additional Datasets*

To validate GSEM's generalizability, we conducted experiments on a larger RGB-D dataset containing 300 objects and 122 concepts. Even with only 10% of the training data, GSEM achieved an F1-score of 0.46, significantly outperforming the NASNetLarge baseline, which scored 0.39. This improvement underscores GSEM's ability to generalize effectively, even in low-resource settings.

Notably, for specific concepts like "blue sphere" and "yellow cone," GSEM demonstrated substantial gains in classification accuracy. The F1-score for "blue sphere" improved by 18% compared to traditional CNN-based methods, while "yellow cone" saw a 15% improvement. These results highlight GSEM's strength in handling complex visual and linguistic combinations, particularly when training data is sparse.

To further explore GSEM's scalability, we evaluated its performance on subsets of the dataset with varying levels of complexity. For simpler concepts like "red apple," GSEM achieved near-perfect classification with an F1-score of 0.91, while for more intricate concepts like "striped cylinder," it achieved an F1-score of 0.74, outperforming all baselines. These findings reinforce GSEM's ability to adapt to diverse datasets and concept complexities.

### *4.8. Robustness to Annotation Noise*

To assess GSEM's resilience to noisy annotations, we introduced synthetic noise by randomly shuffling 20% of the concept labels. Despite this deliberate noise, GSEM maintained an average F1-score of 0.67, significantly outperforming baselines, which dropped to 0.53. This robustness is attributed to GSEM's ability to learn meaningful latent representations that are less sensitive to labeling inconsistencies.

We also evaluated the impact of varying noise levels on classification performance. With 10% noise, GSEM achieved an average F1-score of 0.72, while baselines dropped to 0.61. At 30% noise, GSEM's performance decreased modestly to 0.63, whereas baselines suffered a substantial decline to 0.45. These results indicate that GSEM is highly resilient to real-world annotation challenges, making it suitable for noisy and unstructured datasets.

Additionally, qualitative analyses revealed that noisy labels had minimal impact on the latent space structure. For example, misclassified objects in the "yellow" category remained close to the correct cluster, preserving the overall semantic coherence. This further demonstrates GSEM's robustness in handling annotation noise.

### *4.9. Scaling to High-Dimensional Features*

To evaluate GSEM's scalability, we combined raw RGB-D data with CNN-extracted features, creating a high-dimensional feature set of 2,048 dimensions. Despite the increased complexity, GSEM achieved an F1-score of 0.75, outperforming the predefined category classifier by 12%. This demonstrates GSEM's ability to integrate heterogeneous features for improved performance.

Furthermore, we conducted ablation studies to understand the contribution of different feature types. When using only RGB-D features, GSEM achieved an F1-score of 0.68, while CNN features alone resulted in 0.71. The combination of both feature types yielded the highest performance, highlighting the complementary nature of these features. For example, color-based concepts benefited from RGB-D features, while shape-based concepts showed significant improvement with CNN features.

Finally, we analyzed GSEM's computational efficiency with high-dimensional inputs. Training time increased modestly compared to lower-dimensional setups but remained within practical limits, demonstrating the model's scalability. These findings suggest that GSEM is well-suited for large-scale applications requiring diverse feature integration.

### *4.10. Summary of Findings*

In summary, GSEM consistently outperforms baselines across various datasets, feature types, and experimental conditions. Its ability to leverage CNN features, handle noisy annotations, and scale to high-dimensional inputs highlights its versatility and robustness. These attributes make GSEM a powerful solution for grounded language learning in complex and diverse scenarios.

## 5. Conclusions and Future Directions

This paper introduces the Generative Semantic Embedding Model (GSEM), a versatile and robust framework designed to unify language grounding across diverse linguistic and visual inputs. By leveraging a Gaussian variational autoencoder, GSEM addresses the limitations of predefined attribute categories, offering a more flexible approach to grounding linguistic concepts derived from unconstrained natural language to real-world sensor data. This methodology represents a significant leap forward in aligning multimodal data, providing a comprehensive solution to challenges in grounded language learning.

The evaluation of GSEM highlights its ability to generalize effectively across multiple datasets and modalities, including RGB-D data, while maintaining robustness in noisy and resource-constrained environments. By successfully grounding concepts such as color, shape, and object descriptions, GSEM demonstrates its capability to capture a broad range of grounded linguistic representations. Furthermore, the model's performance in low-resource settings underscores its scalability and utility for applications where annotated data is limited. These advancements establish GSEM as a practical tool for bridging the gap between vision and language.

A key contribution of this work is GSEM's resilience to annotation noise and variability. The model's ability to derive meaningful latent embeddings ensures reliable performance even when faced with inconsistent labels or sparse data. This robustness makes it particularly valuable for real-world applications, such as robotics, human-computer interaction, and multilingual language grounding tasks.

Looking ahead, several promising directions could extend and refine the capabilities of GSEM. One important avenue involves optimizing the model for real-time systems, enabling its deployment in interactive environments such as autonomous robots or assistive devices. Further exploration into temporal data, including video or event streams, could expand its applications to dynamic scenarios, such as action recognition or scene understanding. Incorporating additional modalities, such as audio or haptic data, would enhance its ability to ground non-visual concepts, broadening its applicability to more complex multimodal contexts.

Another critical area of future research is enhancing GSEM's noise resilience through advanced techniques such as active learning or uncertainty modeling. These strategies could further mitigate the effects of noisy labels and optimize the annotation process. Additionally, expanding GSEM's multilingual capabilities by exploring its performance in a wider array of languages, including low-resource and morphologically rich languages, could unlock its potential for cross-lingual transfer learning.

Incremental learning is another exciting direction, allowing GSEM to adapt dynamically to new data and concepts without requiring retraining on the entire dataset. This feature is crucial for applications in evolving environments. Furthermore, a deeper analysis of the latent space representations learned by GSEM could provide valuable insights into the underlying mechanisms, informing the design of even more effective models. Exploring its use in specialized domains, such as medical imaging, autonomous navigation, and environmental monitoring, could reveal its versatility and effectiveness in domain-specific challenges.

In conclusion, GSEM presents a groundbreaking framework for language grounding, combining simplicity with adaptability. Its ability to handle diverse datasets, robustly ground concepts, and generalize across languages and modalities positions it as a foundational model for future advancements in multimodal learning and interaction. The contributions of this work not only enhance the current state of grounded language understanding but also open new avenues for innovation in the field.

## References

1. S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, 1990.
2. T.-C. Chi, M. Shen, M. Eric, S. Kim, and D. Hakkani-Tur, "Just ask: An interactive learning framework for vision and language navigation," in *Conference on Artificial Intelligence (AAAI)*, 2020.
3. B. Wang, L. Ma, W. Zhang, W. Jiang, and F. Zhang, "Hierarchical photo-scene encoder for album storytelling," in *Conference on Artificial Intelligence (AAAI)*, 2019.
4. J. Sinapov, P. Khante, M. Svetlik, and P. Stone, "Learning to order objects using haptic and proprioceptive exploratory behaviors," in *Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 2016.
5. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.
6. C. Matuszek, N. I. J. C. on Artificial Intelligence (IJCAI)Gerald, L. Zettlemoyer, L. Bo, and D. Fox, "A Joint Model of Language and Perception for Grounded Attribute Learning," in *Int'l Conf. on Mahcine Learning*, 2012.
7. L. E. Richards, K. Darvish, and C. Matuszek, "Learning object attributes with category-free grounded language from deep featurization," in *Intelligent Robots and Systems*, 2020.
8. Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, *et al.*, "Experience grounds language," in *EMNLP*, 2020.
9. T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol emergence in robotics: a survey," *Advanced Robotics*, 2016.
10. T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," *European Conference on Computer Vision*, 2010.

11. C. Kery, N. Pillai, C. Matuszek, and F. Ferraro, “Building language-agnostic grounded language learning systems,” in *Ro-Man*, 2019.
12. N. Pillai, K. K. Budhraja, and C. Matuszek, “Improving grounded language acquisition efficiency using interactive labeling,” in *R:SS Workshop on MLHRC*, 2016.
13. R. Paul, J. Arkin, N. Roy, and T. M Howard, “Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators,” in *Robotic Science and Systems*, 2016.
14. S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, “Approaching the symbol grounding problem with probabilistic graphical models,” *AI Magazine*, 2011.
15. J. Brawer, O. Mangin, A. Roncone, S. Widder, and B. Scassellati, “Situated human–robot collaboration: predicting intent from grounded natural language,” in *Intelligent Robots and Systems*, 2018.
16. A. R. Akula, S. Gella, Y. Al-Onaizan, S.-C. Zhu, and S. Reddy, “Words aren’t enough, their order matters: On the robustness of grounding visual referring expressions,” in *Association for Computational Linguistics*, 2020.
17. A. B. Rao, K. Krishnan, and H. He, “Learning robotic grasping strategy based on natural-language object descriptions,” in *Intelligent Robots and Systems*, 2018.
18. S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *IJRR*, 2018.
19. J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, “Learning multi-modal grounded linguistic semantics by playing “I Spy”,” in *Int’l Joint Conf. on Artificial Intelligence (IJCAI)*, 2016.
20. T. Nguyen, N. Gopalan, R. Patel, M. Corsaro, E. Pavlick, and S. Tellex, “Robot object retrieval with contextual natural language queries,” *Robotics Science and Systems*, 2020.
21. R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, “Reusing discriminators for encoding: Towards unsupervised image-to-image translation,” in *Computer Vision and Pattern Recognition*, 2020.
22. A. Pitti, M. Quoy, S. Boucenna, and C. Lavandier, “Brain-inspired model for early vocal learning and correspondence matching using free-energy optimization,” *PLoS Computational Biology*, 2021.
23. Z. Liu, Y. Li, L. Yao, X. Wang, and G. Long, “Task aligned generative meta-learning for zero-shot learning,” in *Conference on Artificial Intelligence (AAAI)*, 2021.
24. M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed, “Information maximization for few-shot learning,” *Neural Information Processing Systems*, 2020.
25. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019.
26. J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Neural Information Processing Systems*, 2019.
27. J. Lei, L. Wang, Y. Shen, D. Yu, T. L. Berg, and M. Bansal, “Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning,” in *Association for Computational Linguistics*, 2020.
28. Q. Tan, L. Gao, Y.-K. Lai, and S. Xia, “Variational autoencoders for deforming 3d mesh models,” in *Computer Vision and Pattern Recognition*, June 2018.
29. T. Yamada, H. Matsunaga, and T. Ogata, “Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3441–3448, 2018.
30. Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, “Seed: Semantics enhanced encoder-decoder framework for scene text recognition,” in *Computer Vision and Pattern Recognition*, 2020.
31. C. Silberer and M. Lapata, “Learning grounded meaning representations with autoencoders,” in *Association for Computational Linguistics*, 2014.
32. A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, “Grounding of textual phrases in images by reconstruction,” in *European Conference on Computer Vision*, 2016.
33. N. Pillai and C. Matuszek, “Unsupervised end-to-end data selection for grounded language learning,” in *Conference on Artificial Intelligence (AAAI)*, 2018.
34. D. Nyga, M. Picklum, and M. Beetz, “What no robot has seen before—probabilistic interpretation of natural-language object descriptions,” in *International Conference on Robotics and Automation*, 2017.
35. B. Hariharan and R. Girshick, “Low-shot visual recognition by shrinking and hallucinating features,” in *ICCV*, Oct 2017.
36. O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, “Matching networks for one shot learning,” in *Neural Information Processing Systems*, 2016.

37. Q. Liu, D. McCarthy, and A. Korhonen, "Second-order contexts from lexical substitutes for few-shot learning of word representations," in **SEM*, 2019.
38. S. Sun, Q. Sun, K. Zhou, and T. Lv, "Hierarchical attention prototypical networks for few-shot text classification," in *EMNLP*, 2019.
39. C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," in *PAMI*, 2014.
40. Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *Computer Vision and Pattern Recognition*, 2020.
41. S. Wang, K.-H. Yap, J. Yuan, and Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Computer Vision and Pattern Recognition*, 2020.
42. D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Neural Information Processing Systems*, 2014.
43. K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *ICRA*, 2011.
44. L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *Computer Vision and Pattern Recognition*, 2011.
45. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
46. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
47. B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Computer Vision and Pattern Recognition*, 2018.
48. C. Silberer, V. Ferrari, and M. Lapata, "Visually grounded meaning representations," *PAMI*, 2016.
49. N. Pillai and C. Matuszek, "Identifying negative exemplars in grounded language data sets," in *R:SS Workshop on Spatial-Semantic Representations in Robotics*, 2017.
50. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Neural Information Processing Systems*, 2013.
51. R. J. Mooney, "Learning to connect language and perception," in *Conference on Artificial Intelligence (AAAI)*, 2008.
52. J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken language instructions," in *International Conference on Robotics and Automation*, 2018.
53. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Neural Information Processing Systems*, 2015.
54. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition*, 2014.
55. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition*, 2017.
56. M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *RAS*, 2018.
57. A. Balajee Vasudevan, D. Dai, and L. Van Gool, "Object referring in videos with language and human gaze," in *Computer Vision and Pattern Recognition*, 2018.
58. Y. Wang, J. van de Weijer, and L. Herranz, "Mix and match networks: Encoder-decoder alignment for zero-pair image translation," in *Computer Vision and Pattern Recognition*, June 2018.
59. C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation," in *Computer Vision and Pattern Recognition*, 2017.
60. H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *International Conference on Robotics and Automation*, 2018.
61. X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Neural Information Processing Systems*, 2016.
62. C. Cadena, A. R. Dick, and I. D. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding." in *Robotic Science and Systems*, 2016.

63. M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *ICCV*, December 2013.
64. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
65. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
66. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
67. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
68. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
69. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
70. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
71. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
72. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
73. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
74. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
75. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
76. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
77. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
78. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
79. J Ngiam, A Khosla, and M Kim. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689—-696, 2011. URL http://ai.stanford.edu/{~}ang/papers/icml11-MultimodalDeepLearning.pdf.
80. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.

81. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
82. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
83. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
84. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
85. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
86. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
87. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
88. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
89. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
90. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
91. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
92. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
93. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
94. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
95. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
96. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
97. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
98. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
99. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
100. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.

101. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
102. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
103. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
104. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
105. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
106. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
107. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
108. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
109. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
110. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
111. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
112. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
113. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
114. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
115. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
116. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
117. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
118. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
119. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.

120. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
121. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
122. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
123. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
124. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
125. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
126. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
127. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
128. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
129. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.