Article

Not peer-reviewed version

Variational Interpretable Framework for Multimodal Instruction Execution

Kay Morgan , Lobry Hsu , Zara Quinn *

Posted Date: 4 April 2025

doi: 10.20944/preprints202504.0394.v1

Keywords: instruction following; multimodal generative models; semi-supervised learning; multimodal variational autoencoders; language-guided navigation

Article

Variational Interpretable Framework for Multimodal Instruction Execution

Kay Morgan, Lobry Hsu and Zara Quinn *

Bond University
* Correspondence: zaraq@bond.edu.au

Abstract: Empowering agents with the ability to understand and follow complex language instructions in diverse environments is a crucial goal in both robotics and artificial intelligence. However, the substantial requirement of paired multimodal data, consisting of natural language commands and their corresponding trajectories, poses a significant challenge in real-world applications. In this work, we propose a novel generative learning framework, **IntraMIX** (Interpretable Multimodal Instruction eXecutor), tailored to semi-supervised instruction-following tasks. Our approach leverages a sequential multimodal generative mechanism to jointly encode and reconstruct both paired and unpaired data through shared latent representations. By extending traditional multimodal variational autoencoders into a sequential domain and introducing an attention-compatible latent structure, IntraMIX successfully addresses the limitations of prior models in sequence-to-sequence tasks. Moreover, we demonstrate how IntraMIX can be integrated into the prevalent speaker-follower pipeline by proposing a new regularization strategy that mitigates overfitting when leveraging unpaired trajectories. Experiments conducted in the BabyAI and Room-to-Room (R2R) environments confirm the effectiveness of our model, where IntraMIX improves instruction-following performance under limited supervision and enhances the speaker-follower framework by 2%–5%. Our results suggest that generative modeling presents a promising pathway toward more data-efficient instruction-following agents.

Keywords: instruction following; multimodal generative models; semi-supervised learning; multimodal variational autoencoders; language-guided navigation

---

1. Introduction

The capability of artificial agents to comprehend and execute natural language instructions in complex environments has remained a fundamental challenge in AI, gaining increasing relevance with advancements in robotics and embodied intelligence [4]. This task, often referred to as instruction following, demands an intricate fusion of linguistic comprehension and action-oriented planning. While recent developments in imitation learning (IL) and reinforcement learning (RL) have enabled agents to perform such tasks in visually and spatially grounded simulations, such as BabyAI [5] and Room-to-Room (R2R) [2], the high dependency on large-scale paired datasets continues to restrict scalability.

In particular, neural models that map linguistic inputs to navigation actions require tens or hundreds of thousands of annotated instruction-trajectory pairs, even in synthetic gridworlds [5]. This requirement is even more pronounced in real-world scenarios, where data collection is costlier and more error-prone. Thus, enabling agents to learn effectively from partially labeled or unlabeled data becomes crucial to improving data efficiency.

To address this issue, semi-supervised learning approaches have been introduced. One well-known framework is the speaker-follower model [7], which generates synthetic instructions for unpaired trajectories via a trained "speaker" model. The synthetic pairs are then used to train a "follower" policy, reducing the dependence on labeled data. Subsequent refinements of this approach, such as data augmentation [22] and confidence-based filtering [26], have shown promise. Nonetheless,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

these techniques are inherently limited, as they rely on a speaker trained only from the scarce paired data, which may not generalize well to diverse unseen instructions or environments.

To overcome these limitations, we propose a fundamentally different strategy rooted in probabilistic generative modeling. Our proposed solution, IntraMIX, reinterprets instruction following as a semi-supervised multimodal generation problem. Specifically, we extend the framework of multimodal variational autoencoders (M-VAEs) [25] to sequential settings, thereby enabling the learning of shared representations that capture the underlying semantics of both language and trajectory sequences.

Unlike conventional M-VAEs which are designed for static data modalities (e.g., images and captions), IntraMIX adopts a temporal generative formulation suitable for sequence-to-sequence tasks. A critical innovation is our introduction of a **bottleneck attention mechanism**, which aligns the variable-length sequences in both modalities by projecting them into a common latent space with fixed-length temporal representations. This structure facilitates compatibility with the attention mechanism [18], widely employed in prior work for modeling language-grounded planning [2,5].

Beyond pure generative modeling, IntraMIX is designed to be flexibly integrated with the speaker-follower architecture. Leveraging the bidirectionality of our generative formulation, we repurpose IntraMIX as both a trajectory generator (follower) and an instruction generator (speaker). This dual functionality allows for a closed-loop learning paradigm, in which unpaired data from either modality can be reconstructed, translated, or leveraged for regularization. To prevent overfitting and distributional mismatch when training with unpaired data, we introduce a novel regularization term in the loss function that minimizes the divergence between the latent distributions of paired and unpaired inputs. This cross-modal regularization ensures that unpaired trajectories and instructions are aligned in the latent space, promoting robustness and generalization across modalities.

We conduct extensive experiments in BabyAI and R2R environments to evaluate the efficacy of IntraMIX. The results demonstrate clear improvements over both purely supervised baselines and previous semi-supervised models. When used as a standalone follower model, IntraMIX outperforms traditional IL-trained agents by margins of 5.1% on BabyAI and 3.7% on R2R in task completion rate. When integrated with the speaker-follower pipeline, it further enhances instruction-following accuracy by an additional 2.5% to 4.9%. These results validate the dual utility of IntraMIX: as a generative learner for semi-supervised settings and as an enhancement module for existing pipelines.

- We propose **IntraMIX**, a novel generative framework for instruction-following agents that unifies multimodal sequence modeling with attention-aware representation learning.
- Our architecture introduces a bottleneck attention mechanism to handle variable-length sequence alignment, which is crucial for language-to-trajectory tasks.
- We demonstrate that IntraMIX can be seamlessly integrated into the speaker-follower paradigm, enhancing both components through latent alignment and a novel regularization objective.
- Extensive empirical results on two established benchmarks show that our model achieves competitive performance and outperforms prior semi-supervised approaches under limited supervision.

2. Related Work

Instruction-following tasks under semi-supervised settings have garnered extensive research attention, particularly with the advent of neural agents capable of interpreting natural language commands and executing them in grounded environments. A central paradigm in this domain is the *speaker-follower* model, which forms the foundational framework for leveraging unpaired trajectory data to mitigate the scarcity of annotated pairs. Pioneering works such as Fu *et al.* [8], Huang *et al.* [12], Tan *et al.* [22], Yu *et al.* [26] introduced various refinements and extensions to this architecture, enabling more robust utilization of synthetic or unpaired trajectories. Notably, Yu *et al.* [26] and Fu *et al.* [8] presented sophisticated strategies for collecting unpaired data with improved semantic fidelity and task relevance, thereby enhancing the downstream policy learning. Complementarily, Huang *et al.* [12] introduced a discriminator-based selection mechanism to filter out noisy or semantically inconsistent synthetic instructions generated by speakers. Such mechanisms effectively reduce the propagation of erroneous signals during training, particularly in large-scale or multi-instruction scenarios.

Beyond supervised imitation learning, variants of the speaker-follower model have also been adapted to reinforcement learning (RL) settings. For instance, Cideron *et al.* [6] proposed a data augmentation scheme aligned with the spirit of the speaker-follower model, wherein high-reward trajectories are reused with synthesized instructions to improve sample efficiency. However, a fundamental limitation persists across these models: the speaker component is often trained exclusively on the limited set of paired data, constraining its generalizability and rendering it brittle in out-of-distribution or long-horizon tasks.

In contrast, our proposed approach, **IntraMIX**, departs from the conventional reliance on deterministic back-translation and instead adopts a probabilistic generative modeling framework to unlock richer cross-modal interactions. While IntraMIX is architecturally distinct, it is inherently complementary to the speaker-follower framework. By enabling the learning of a shared latent representation space through semi-supervised training, IntraMIX directly augments the capabilities of both the speaker and follower components. This synergy permits mutual reinforcement between modalities, yielding more accurate synthetic data and more robust policies.

The conceptual underpinning of IntraMIX also draws connections to unsupervised machine translation literature, particularly the works of Artetxe *et al.* [3] and Lample *et al.* [16], which proposed effective frameworks for exploiting unaligned bilingual corpora. These methods typically employ two core strategies: (i) *back-translation*—generating a sentence in one language from its counterpart in another—and (ii) *shared latent reconstruction*—learning a joint representation that enables reconstruction of either modality. While the former is conceptually similar to the speaker-follower approach, the latter is more aligned with the generative mechanism adopted in IntraMIX. Specifically, the shared latent variable $\mathcal{V}$ in our formulation (see Eq. 4) corresponds to the cross-modal embedding used to jointly represent and reconstruct both language and trajectory data. However, unlike prior works that use static embeddings, IntraMIX introduces temporal structure via its attention-enabled sequence encoder, thus enhancing contextual coherence across modalities.

The novelty of IntraMIX also lies in its architectural innovations. For example, while Lee *et al.* [17] proposed a bottleneck-style module for extracting salient subsets from unordered sets, IntraMIX reinterprets this idea within a sequential attention context. Rather than filtering input features, our *bottleneck attention* serves as a cross-modality mediator, harmonizing variable-length linguistic and behavioral sequences into a temporally structured latent space. This not only facilitates sequence-to-sequence transformations but also aligns with the attention paradigms popularized by transformer models in vision-language navigation [11].

Moreover, our framework relates to earlier work on variational sequence modeling. Kočiskỳ *et al.* [14], for instance, proposed a latent-variable model for structured prediction tasks, employing recurrent networks to generate latent sequences. While effective, this method introduces additional autoregressive components that limit parallelizability and are susceptible to vanishing gradients. IntraMIX circumvents these issues by relying exclusively on attention-based modules for latent variable inference, enabling parallel computation while maintaining expressive power and stable training dynamics.

Additionally, recent advances in large-scale pretrained language models such as BERT [28] have inspired applications of transfer learning to instruction following. These include leveraging textual embeddings for trajectory grounding [11] or using pretrained transformers as instruction encoders. Despite their success, these methods often overlook the bidirectional generative potential between language and action, which lies at the core of IntraMIX's architecture. Furthermore, many data augmentation strategies used in prior studies, including environment dropout [22] and reinforcement-driven auxiliary losses [24], can be viewed as orthogonal improvements that are complementary to the generative backbone introduced in our work.

 doi:10.20944/preprints202504.0394.v1

In summary, IntraMIX is positioned at the intersection of several research streams, including generative modeling, semi-supervised learning, instruction-grounded navigation, and sequence-to-sequence representation learning. Its capacity to unify these perspectives under a coherent probabilistic framework distinguishes it from prior approaches and opens up new avenues for multimodal agent training in low-resource or partially annotated environments.

3. Preliminary

3.1. Task Definition and Semi-Supervised Setup

We consider the task of visually grounded instruction following, in which an agent is required to interpret a natural language instruction and generate a corresponding action sequence that fulfills the task objectives. Let the instruction be represented as a variable-length sequence $y_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,L_i}]$, where L_i denotes the length of the instruction in episode i. The instruction typically conveys goal-directed behaviors, such as *"Turn right at the hallway and go to the red sofa"*.

At each discrete timestep t, the agent receives an observation $o_{i,t}$ and produces an action $a_{i,t}$ based on the accumulated sequence of past observations and the language instruction. Formally, the policy function is defined as:

$$\pi : (y_i, o_{i,1:t}) \mapsto a_{i,t}, \quad \forall t \in [1, T_i],$$

where T_i is the trajectory length in episode i.

We adopt a semi-supervised learning paradigm, following prior frameworks such as Fried *et al.* [7], where both labeled and unlabeled data are leveraged for training. The labeled (paired) dataset is defined as:

$$D_p = \{(\tau_i, y_i)\}_{i=1}^{M},$$

where $\tau_i = (o_i, a_i) = ([o_{i,1}, \ldots, o_{i,T_i}], [a_{i,1}, \ldots, a_{i,T_i}])$ is the interaction trajectory comprising both environment observations and the agent's actions. The unlabeled (unpaired) dataset consists only of trajectory data:

$$D_u = \{\tau_j\}_{j=1}^{N}, \quad \text{where } \tau_j = (o_j, a_j),$$

with M and N representing the number of paired and unpaired samples, respectively.

Throughout our formulation, we assume that the instruction lengths L_i and trajectory lengths T_i vary across episodes. For notational convenience, we drop the subscript i when the context is unambiguous. This setup presents a realistic and challenging scenario where annotated instructions are sparse, motivating the need for semi-supervised or generative learning mechanisms like IntraMIX to bridge the supervision gap.

3.2. Canonical Sequence-Based Follower Architecture

To establish a foundational comparison, we briefly revisit the standard attention-based sequence-to-sequence (seq2seq) follower architecture, which serves as the backbone for many instruction-following agents, including those used in environments like Room-to-Room (R2R) [2] and BabyAI [5]. Our proposed IntraMIX model adopts and extends this architectural baseline with significant improvements in generative inference and multimodal alignment.

The follower architecture consists of three primary components: an instruction encoder f_{enc}, a trajectory decoder f_{dec}, and an attention-based alignment module f_{att} that bridges them. Concretely, given an instruction $y = [y_1, \ldots, y_L]$, the encoder maps it to a sequence of contextual hidden representations:

$$[h_1, h_2, \ldots, h_L] = f_{\text{enc}}(y),$$

typically using an LSTM or Transformer encoder.

During interaction, at each timestep t, the decoder produces a hidden state based on prior actions and observations:

$$h'_t = f_{\text{dec}}(a_{1:t-1}, o_{1:t}),$$

 doi:10.20944/preprints202504.0394.v1

which summarizes the agent's partial trajectory and accumulated environment knowledge.

An attention mechanism then computes a context vector by aligning the decoder's current state with the instruction embeddings:

$$c_t = f_{\text{att}}(h'_t, [h_1, \ldots, h_L]),$$

where the attention output c_t captures the most relevant parts of the instruction at timestep t.

Finally, the action at timestep t is predicted via an action predictor module f_{act}, which integrates the decoder and context vectors:

$$a_t = f_{\text{act}}(h'_t, c_t).$$

This architecture facilitates effective sequence grounding and supports long-horizon goal execution. However, it relies heavily on paired data for training and lacks mechanisms to incorporate unpaired trajectories or leverage latent semantic structures—limitations that our proposed IntraMIX model is designed to overcome.

3.3. Data Augmentation with Speaker-Follower Paradigm

The *speaker-follower* framework is a widely adopted strategy for semi-supervised instruction following. It introduces an auxiliary module—the **speaker**—which is responsible for generating natural language descriptions from observed trajectories. The core idea is to use this component to synthetically annotate unpaired trajectories, thereby expanding the effective training data for the follower.

The speaker is architecturally similar to the follower: it follows an attention-based seq2seq model, but operates in reverse. Given a trajectory $\tau = (o, a)$, it encodes the sequence into latent representations and autoregressively decodes them into a textual instruction $\hat{y} = [\hat{y}_1, \ldots, \hat{y}_L]$. More formally:

$$\hat{y}_t \sim p_{\text{spk}}(\hat{y}_t | \hat{y}_{1:t-1}, \tau),$$

where p_{spk} denotes the speaker's conditional language model, trained using teacher-forced supervision on paired samples (τ, y).

Once trained, the speaker is applied to each unpaired trajectory $\tau_j \in D_u$ to produce a pseudo-instruction $\hat{y}_j$. This creates an auxiliary paired dataset:

$$\hat{D}_p = \{(\tau_j, \hat{y}_j)\}_{j=1}^{N},$$

which is then used to train or fine-tune the follower model using standard supervised objectives.

Despite its empirical success, this framework suffers from several inherent limitations. First, the speaker is trained only on the small paired dataset D_p, which limits its generalization capability in unseen environments or trajectory styles. Second, the generation process is unidirectional and deterministic, lacking uncertainty quantification or latent modeling of cross-modal semantics. Third, the synthetic instructions may exhibit semantic drift or compositional inconsistencies, introducing noise into the training process.

In this work, our proposed IntraMIX model provides a principled generative alternative to the speaker, capable of modeling uncertainty via latent variables and supporting reconstruction from either modality. Moreover, IntraMIX can act as both a speaker and follower simultaneously, with shared latent grounding. In Section 4.1, we detail how IntraMIX not only subsumes the speaker-follower paradigm but also improves its effectiveness through probabilistic cross-modal alignment and latent-space regularization.

4. IntraMIX: A Generative Probabilistic Perspective

4.1. Overview

In this section, we introduce the underlying probabilistic formulation of our proposed framework, **IntraMIX**, which stands for *Interpretable Multimodal Instruction eXecutor*. The core idea is to model

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

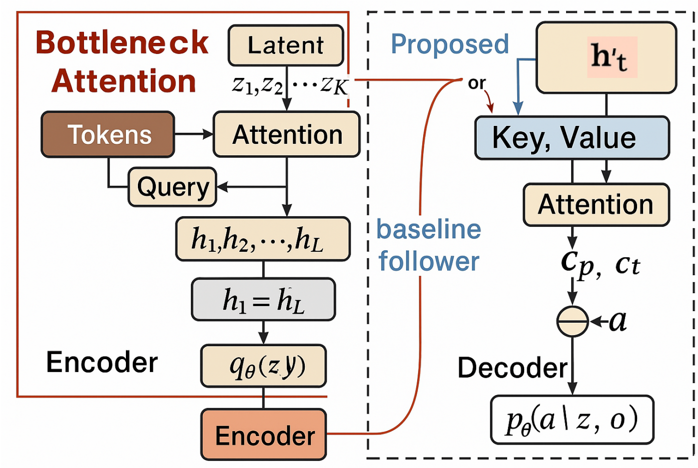

Figure 1. Overview architecture of the DIFERNet framework.

instruction following as a multimodal generative task that captures the joint distribution of language instructions, trajectories, and environmental observations through a latent variable framework. Our approach builds upon the foundations of Multimodal Variational Autoencoders (M-VAE) [25], extending them to sequential decision-making settings with attention and autoregressive decoding.

We begin by defining the generative process over a latent representation z, a language instruction y, and an action sequence a given a sequence of environment observations o. The generative model is structured as:

$$p_\theta(z, y, a|o) = p_\theta(z) \cdot p_\theta(y|z) \cdot p_\theta(a|z, o), \tag{1}$$

where $p_\theta(z)$ denotes the prior over latent task embeddings, $p_\theta(y|z)$ represents the language generation process, and $p_\theta(a|z, o)$ models the trajectory conditioned on both the environment context and the latent semantics.

Analogous to standard M-VAE configurations, this factorization allows us to interpret the generation of paired modalities—language and action—from a shared semantic factor z. That is, given $x_{(1)} = a|o$ and $x_{(2)} = y$, this model can be viewed as a structured extension of the canonical multimodal generation process.

However, as the true posterior distribution $p_\theta(z|x_{(1)}, x_{(2)})$ is intractable, we introduce a variational approximation:

$$q_\phi(z|y, a, o) = \frac{1}{2} q_\phi(z|y) + \frac{1}{2} q_\phi(z|a, o), \tag{2}$$

following the mixture-of-experts (MoE) strategy shown by Shi *et al.* [20] to be particularly effective for cross-modal latent alignment. This balanced fusion enforces a symmetrical contribution from each modality, helping to mitigate mode collapse and ensure mutual consistency.

To learn the model, we maximize the evidence lower bound (ELBO) of the marginal log-likelihood under both paired and unpaired settings. For paired data D_p, the ELBO becomes:

$$\begin{aligned}\mathcal{J} = &\frac{1}{2}\mathbb{E}_{q_\phi(z|y)}\left[\log \frac{p_\theta(z)p_\theta(y|z)p_\theta(a|z,o)}{q_\phi(z|y,a,o)}\right] \\ &+\frac{1}{2}\mathbb{E}_{q_\phi(z|a,o)}\left[\log \frac{p_\theta(z)p_\theta(y|z)p_\theta(a|z,o)}{q_\phi(z|y,a,o)}\right].\end{aligned} \tag{3}$$

In contrast, for unpaired trajectory-only data D_u, we derive a reduced ELBO over the marginal $p_\theta(a|o)$ as:

$$\mathcal{V} = \mathbb{E}_{q_\phi(z|a,o)}[\log p_\theta(a|z, o)] + D_{\mathrm{KL}}[q_\phi(z|a, o) \parallel p_\theta(z)]. \tag{4}$$

This component allows the model to learn from unannotated sequences by reconstructing plausible actions based solely on contextual observations and inferred task embeddings.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

Combining both settings, we formulate the overall training objective as a weighted combination:

$$\max \mathbb{E}_{\{\tau,y\}\in D_p}[\bar{\mathcal{J}}] + \gamma \mathbb{E}_{\tau\in D_u}[\mathcal{V}], \tag{5}$$

where γ modulates the influence of unpaired supervision. This dual-objective framework enables IntraMIX to benefit from large quantities of unpaired data while maintaining alignment with high-quality paired annotations.

To operationalize action generation, we incorporate a timestep variable t and reformulate the trajectory decoder in an autoregressive manner:

$$p_\theta(a|z,o) = \prod_{t=1}^{T} p_\theta(a_t|z, o_{1:t}, a_{1:t-1}). \tag{6}$$

This allows the agent to produce actions sequentially, in alignment with how agents interact in real-world environments.

Similarly, the language decoder is also defined as:

$$p_\theta(y|z) = \prod_{t=1}^{L} p_\theta(y_t|z, y_{1:t-1}). \tag{7}$$

The inference processes for follower and speaker usage are then defined respectively as:

$$a_t \sim p_\theta(a_t|z, o_{1:t}, a_{1:t-1}), \quad z \sim q_\phi(z|y), \tag{8}$$

$$y_t \sim p_\theta(y_t|z, y_{1:t-1}), \quad z \sim q_\phi(z|a,o). \tag{9}$$

We refer to the two usage modes of IntraMIX as **IntraMIX-Follower** (direct rollout) and **IntraMIX-SpeakerFollower** (integration into the speaker-follower framework).

4.2. Cross-Modal Compression via Bottleneck Attention

To support alignment across modalities with potentially differing sequence lengths, we introduce a novel **Bottleneck Attention Module**, designed to extract fixed-length latent sequences from variable-length modality streams. Unlike naive approaches that produce sequence-level latent variables from modality-specific encoders and risk mismatch in temporal alignment, our mechanism projects variable-length hidden states into a common K-dimensional sequence via trainable token queries $e = [e_1, ..., e_K]$.

Let $h = [h_1, ..., h_L]$ denote the encoded hidden states (e.g., of a language instruction). The module performs multi-head attention with the bottleneck tokens e as queries, and h as both keys and values:

$$z = \text{Attention}(e, h, h).$$

Each resulting z_i is modeled as a Gaussian latent variable with its own mean and variance, parameterized via the attention output. These variables form the latent sequence used downstream:

$$q_\phi(z|y) = \prod_{i=1}^{K} \mathcal{N}(\mu_i, \sigma_i^2).$$

This structure ensures consistent dimensionality across modalities and improves compatibility with attention-based decoding mechanisms. Importantly, the bottleneck attention is modular and can be seamlessly integrated into any Transformer-style encoder or decoder without requiring architectural overhauls.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

4.3. Domain Alignment with Latent Distribution Regularization

Despite careful design, domain mismatch between paired and unpaired data poses a risk to generalization. Specifically, encodings of trajectories from D_p and D_u may drift apart due to their involvement in different loss terms. To mitigate this, we introduce **Domain Distance Regularization**, a penalty that encourages alignment of the latent distributions over z induced by the two domains.

Let:

$$\rho = \mathbb{E}_{\tau \in D_p}[q_\phi(z|\tau)], \quad v = \mathbb{E}_{\tau' \in D_u}[q_\phi(z|\tau')].$$

We define a regularization penalty $D(\rho, v)$, which measures the divergence between these distributions. While any divergence metric can be used, we adopt the *Sliced Wasserstein Distance (SWD)* [15] for its efficiency and empirical robustness.

The final objective becomes:

$$\max \mathbb{E}_{\{\tau,y\} \in D_p}[\bar{\mathcal{J}}] + \gamma \mathbb{E}_{\tau \in D_u}[\mathcal{V}] - \alpha D(\rho, v), \tag{10}$$

with α as the domain alignment strength coefficient.

Through this combination of generative modeling, bottleneck attention, and domain alignment regularization, IntraMIX provides a principled, extensible foundation for instruction-following in multimodal semi-supervised environments.

5. Experiments

5.1. Benchmarks and Setup Overview

To systematically assess the effectiveness of our proposed model **IntraMIX**, we conducted experiments across two widely-used instruction-following environments: **BabyAI** [5] and **Room-to-Room (R2R)** [2]. These environments offer complementary characteristics—BabyAI focuses on discrete, symbolic reasoning in a gridworld, while R2R emphasizes grounded vision-language understanding in photorealistic 3D environments.

Within BabyAI, we evaluate our model across four tasks: **GoToSeq**, **GoToSeqLocal**, **BossLevel**, and **BossLocal**. The latter two (GoToSeqLocal and BossLocal) are introduced in our study as simplified but semantically aligned variants of their original counterparts, specifically tailored to evaluate generalization with limited input complexity. In GoTo-style tasks, the agent must reach a target object described in natural language, whereas Boss-level tasks require completing multiple sub-instructions, including object manipulation (e.g., "pick up the red key").

Performance on BabyAI is primarily evaluated using the mean **Success Rate (SR)**, which measures whether the agent successfully completes all subgoals in the instruction. Additionally, to evaluate the quality of the speaker (language generation from trajectories), we adopt the **BLEU-4** metric, commonly used in text generation tasks.

In the R2R dataset, the agent is required to navigate a photorealistic environment based on natural language instructions describing the intended path. The dataset contains 7,189 human-annotated trajectory-instruction pairs, with an additional 178,000 unpaired trajectories provided for semi-supervised learning [7]. R2R uses three evaluation environments—*validation seen*, *validation unseen*, and *test*. Performance is measured using:

- **Success Rate (SR)**: Whether the agent stops within 3 meters of the goal.
- **Oracle Success Rate (OSR)**: Whether any position along the path is within 3 meters of the goal.
- **Navigation Error (NE)**: The final distance from the target location (lower is better).

Further experimental configurations, including architectural specifics, optimization hyperparameters, and validation strategies, are detailed in Appendix D.

5.2. Ablation: Impact of Architectural Design

Evaluating Bottleneck Attention We first analyze the architectural contributions of IntraMIX by isolating the effects of attention mechanisms. As shown in Table 1, a conventional seq2seq model

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

without attention underperforms significantly in complex environments such as BossLevel, highlighting its limitations in handling long-horizon or compositional instructions. By incorporating standard attention, performance improves drastically, particularly in semantically dense tasks.

Table 1. Comparative SR (%) across different architectural variants on BabyAI. Bold denotes top-2 performance.

Architecture	K	GoToSeqLocal	GoToSeq	BossLocal	BossLevel
seq2seq		98.7	**96.3**	86.5	47.2
w/ attention		**99.6**	95.1	**99.1**	**88.4**
w/ bottleneck	4	**99.4**	93.2	96.4	80.8
	16	99.2	**96.0**	**98.9**	**86.5**

Our bottleneck attention module achieves competitive performance, even surpassing attention baselines in GoToSeqLocal. Notably, with $K = 16$, IntraMIX nearly matches the attention-based seq2seq model across all tasks, while offering the added benefit of latent interpretability and alignment. This suggests that bottleneck attention serves as an effective substitute for standard attention in settings where symbolic reasoning and latent sequence alignment are essential.

5.3. Ablation: Training Objective Decomposition

Effect of Loss Components To assess the contribution of each loss component in IntraMIX, Table 2 compares three settings: (1) supervised-only, (2) unsupervised generative modeling without regularization, and (3) the full IntraMIX objective (with both $\mathcal{V}$ and domain distance regularization $D(\rho, v)$).

Table 2. Ablation study on IntraMIX's objectives: SR and BLEU for different loss configurations on BabyAI.

		Task	GoToSeqLocal		BossLocal	
			SR	BLEU	SR	BLEU
Method	$\mathcal{V}$	$D(\rho, v)$				
supervised			54.8	10.32	45.6	4.83
IntraMIX			49.9	11.01	41.3	6.31
	✓		66.5	10.78	66.1	6.09
(full)	✓	✓	**70.8**	**11.61**	**74.5**	**7.21**

Results indicate that incorporating unpaired trajectory data via $\mathcal{V}$ significantly improves both SR and BLEU over the baseline. The addition of the regularization term further boosts speaker performance, mitigating overfitting caused by domain drift. This validates the importance of aligning embedding distributions across paired and unpaired domains, particularly for speaker generalization.

5.4. Comparison: Alternative Semi-Supervised Strategies

IntraMIX as Follower vs. Speaker-Follower As shown in Table 3, IntraMIX consistently outperforms conventional follower and speaker-follower models across both BabyAI and R2R. Notably, IntraMIX used as a speaker-follower leads to the highest success rate, demonstrating the synergistic value of a generative speaker integrated within a semi-supervised framework. These findings confirm that IntraMIX not only functions as a standalone follower but also strengthens traditional augmentation pipelines when used as a generative speaker.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

Table 3. Performance of IntraMIX under two usage modes compared to baseline methods.

Method	Follower w/ D_u	Speaker w/ D_u	SR BabyAI -BossLocal	SR R2R -unseen
follower			45.3	31.2*
speaker-follower	✓		75.6	35.5*
IntraMIX-follower	✓		76.1	34.2
IntraMIX-speaker-follower	✓	✓	**82.3**	**40.5**

5.5. Benchmarking Against State-of-the-Art Approaches

Performance on R2R Table 4 presents a comprehensive comparison between IntraMIX and multiple state-of-the-art semi-supervised methods on R2R. When used with greedy decoding, IntraMIX consistently surpasses baseline methods in OSR and performs competitively on SR, despite using fewer training annotations. When equipped with pragmatic inference, IntraMIX significantly outperforms all competitors across every metric and split, confirming its robustness and generalization capabilities.

Table 4. Comparison of IntraMIX with prior semi-supervised methods on R2R. ↓/↑ indicates better-lower/better-higher.

Decoding	Method	Validation Seen NE↓	Validation Seen SR↑	Validation Seen OSR↑	Validation Unseen NE↓	Validation Unseen SR↑	Validation Unseen OSR↑	Test NE↓	Test SR↑	Test OSR↑
greedy	speaker-follower [7]	3.36	66.4	73.8	6.62	35.5	45.0	-	-	-
	Tan *et al.* [22]	3.99	62.1	-	**5.22**	**52.2**	-	-	**51.5**	-
	Huang *et al.* [12]	5.00	50.4	-	5.90	39.1	-	-	-	-
	Yu *et al.* [26]	5.03	53.0	61.6	6.29	38.9	46.7	-	-	-
	Fu *et al.* [8]	**3.30**	**68.2**	**74.9**	6.10	38.8	46.7	**5.90**	37.6	**46.4**
	IntraMIX	3.72	64.4	73.2	6.35	40.5	**49.0**	6.35	38.2	46.2
pragmatic	speaker-follower [7]	3.08	70.1	78.3	4.83	54.6	65.2	4.87	53.5	63.9
	IntraMIX	**2.70**	**74.3**	**80.3**	**4.44**	**56.9**	**66.4**	**4.52**	**56.1**	**64.2**

In particular, IntraMIX's performance gains in unseen environments underscore its ability to learn transferable semantic representations—a result attributed to its generative modeling of task semantics and its latent space regularization.

6. Conclusions

In this paper, we presented **IntraMIX**, a novel generative modeling framework for semi-supervised learning in sequence-to-sequence multimodal tasks, with a specific focus on instruction-following agents. Building upon the foundation of Multimodal Variational Autoencoders (M-VAE), our method introduces two key innovations: the bottleneck attention module, which enables alignment-aware latent representation learning, and a domain distance regularization term that ensures consistent cross-domain generalization.

Compared with prior M-VAE architectures, IntraMIX has the distinct advantage of supporting attention mechanisms, making it more compatible with complex sequential tasks where information must be selectively attended over time. This architectural benefit was quantitatively validated through the results shown in Table 1, where attention-enabled models significantly outperformed vanilla seq2seq baselines, particularly on semantically demanding tasks like BossLevel.

Moreover, our ablation results in Table 2 highlight the substantial contribution of each component of our method. The use of unpaired trajectories via the variational ELBO term ($\mathcal{V}$) clearly enhances both policy success rate and speaker performance, while the inclusion of domain distance regularization $D(\rho, v)$ further improves language generation fidelity by mitigating embedding drift across data domains.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

From a practical perspective, IntraMIX demonstrated dual utility. As shown in Tables 2 and 3, it significantly improves the performance of the follower when used as a standalone model. Simultaneously, when acting as a speaker within the speaker-follower paradigm, it leads to further gains, as evidenced by the improvements in Table 4. This dual-role capability underscores the complementary nature of our generative approach: it strengthens both components of traditional semi-supervised systems and offers greater flexibility in deployment.

In essence, IntraMIX contributes to the field in two critical ways. First, it offers a powerful framework for exploiting unpaired trajectory data, thereby alleviating the heavy reliance on expensive paired annotation. Second, it introduces a modular design compatible with existing architectures and applicable across modalities, thereby supporting integration into broader multimodal learning scenarios.

Looking ahead, several promising directions remain for future research. One natural extension is to leverage unpaired language data by incorporating a generative model over instructions, i.e., replacing Eq. 4 with a symmetric objective over $p_\theta(z, y)$. This would allow IntraMIX to benefit from large-scale, unannotated corpora—a critical resource in instruction-heavy domains such as robotics or AR/VR navigation.

Additionally, given the general nature of our model's formulation, IntraMIX can be readily applied to other tasks involving sequential multimodal generation, including video captioning, text-to-speech synthesis, dialogue agents with memory, or even vision-language pretraining setups. Since our bottleneck attention mechanism is orthogonal to the specific modality design, its integration with recent large multimodal pre-trained transformers presents an exciting avenue for further investigation.

In summary, this work introduces a flexible, principled, and empirically validated generative learning strategy that improves upon existing semi-supervised frameworks for instruction-following. We hope that IntraMIX will serve as a foundation for future explorations at the intersection of multimodal sequence modeling and semi-supervised reasoning.

References

1. Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.
2. Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.
3. Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations*, 2018.
4. Michael Beetz, Ulrich Klank, Ingo Kresse, Alexis Maldonado, Lorenz Mösenlechner, Dejan Pangercic, Thomas Rühr, and Moritz Tenorth. Robotic roommates making pancakes. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pages 529–536, 2011.
5. Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*, 2019.
6. Geoffrey Cideron, Mathieu Seurin, Florian Strub, and Olivier Pietquin. Higher: Improving instruction following with hindsight generation for experience replay. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 225–232. IEEE, 2020.
7. Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
8. Tsu-Jui Fu, Xin Eric Wang, Matthew F Peterson, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. Counterfactual vision-and-language navigation via adversarial path sampler. In *European Conference on Computer Vision*, pages 71–86. Springer, 2020.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2025 doi:10.20944/preprints202504.0394.v1

9. Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.
10. Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
11. Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021.
12. Haoshuo Huang, Vihan Jain, Harsh Mehta, Jason Baldridge, and Eugene Ie. Multi-modal discriminative model for vision-and-language navigation. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 40–49, 2019.
13. David Yu-Tung Hui, Maxime Chevalier-Boisvert, Dzmitry Bahdanau, and Yoshua Bengio. Babyai 1.1. *arXiv preprint arXiv:2007.12770*, 2020.
14. Tomáš Kočiskỳ, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. Semantic parsing with semi-supervised sequential autoencoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1078–1087, 2016.
15. Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019.
16. Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018.
17. Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.
18. Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.
19. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.
20. Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32:15718–15729, 2019.
21. Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.
22. Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. 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 2610–2621, 2019.
23. George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. In *International Conference on Learning Representations*, 2018.
24. Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.
25. Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5580–5590, 2018.
26. Felix Yu, Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. Take the scenic route: Improving generalization in vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 920–921, 2020.
27. Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 5885–5892, 2019.
28. 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.

29. 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).
30. 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.
31. 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.
32. 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)*.
33. 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.
34. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
35. 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).
36. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. https://doi.org/10.1007/s00530-010-0182-0.
37. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
38. 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/.
39. 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.
40. 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.
41. 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. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
42. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
43. 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.
44. 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.
45. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
46. 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.
47. 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.

48. 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.
49. 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.
50. 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.
51. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
61. 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.
62. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
63. 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.
64. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
65. 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.
66. 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. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
67. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
68. 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.
69. 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.

70. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. S. Banerjee and A. Lavie, “METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,” in *IEEMMT*, 2005, pp. 65–72.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.

 doi:10.20944/preprints202504.0394.v1

90. 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.
91. 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.
92. 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.
93. 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.
94. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
95. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
96. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.