Article

Not peer-reviewed version

SensorySync: Multimodal Integration Framework for Unified Perceptual Understanding

Zephyr Lawson , Ava Martinez , Seraphina Quinn *

Posted Date: 25 September 2024

doi: 10.20944/preprints202409.1969.v1

Keywords: Multimodal Learning; Generative Model; Representation Learning

 doi:10.20944/preprints202409.1969.v

Article

SensorySync: Multimodal Integration Framework for Unified Perceptual Understanding

Zephyr Lawson, Ava Martinez and Seraphina Quinn

University of Central Oklahoma
* Correspondence: seraphina.quinn@uco.edu

Abstract: Generic text embeddings have demonstrated considerable success across a multitude of applications. However, these embeddings are typically derived by modeling the co-occurrence patterns within text-only corpora, which can limit their ability to generalize effectively across diverse contexts. In this study, we investigate methodologies that incorporate visual information into textual representations to overcome these limitations. Through extensive ablation studies, we introduce a novel and straightforward architecture named VisualText Fusion Network (VTFN). This architecture not only surpasses existing multimodal approaches on a range of well-established benchmark datasets but also achieves state-of-the-art performance on image-related textual datasets while utilizing significantly less training data. Our findings underscore the potential of integrating visual modalities to substantially enhance the robustness and applicability of text embeddings, paving the way for more nuanced and contextually rich semantic representations.

Keywords: multimodal learning; generative model; representation learning

1. Introduction

Human cognition is endowed with an exceptionally sophisticated framework that facilitates the development of intricate and multifaceted representations of the surrounding environment. This cognitive architecture is equipped with mechanisms that enable individuals to assimilate new information about their surroundings and to retrieve and recognize previously established representations stored within their memory systems [1,2,49]. The sensory information humans receive is inherently multimodal, encompassing various channels such as vision, hearing, touch, taste, and smell. These diverse sensory inputs are processed and integrated through specialized pathways, allowing for a holistic and comprehensive understanding of the environment. However, the information acquired through these sensory modalities is not always complete; gaps can arise either because certain modalities are not stimulated by the environment or due to limitations and impairments in the sensory systems themselves. To navigate such scenarios of incomplete sensory information, the human cognitive framework employs a process known as cross-modal inference. This process allows the perception of one sensory modality to compensate for the absence or deficiency of another, effectively enabling the reconstruction of missing sensory experiences based on available inputs [3–5,50,57].

In practical terms, cross-modal inference is crucial for humans to interact seamlessly with their environment, especially in situations where sensory information is partial or ambiguous. For instance, in low-light conditions, auditory cues can enhance spatial awareness, while in environments with excessive noise, visual signals can aid in communication and navigation. This ability to infer missing sensory information ensures that humans can maintain a stable and coherent perception of their surroundings, thereby facilitating effective decision-making and action-taking even under sensory uncertainty.

Contrastingly, artificial agents such as robots and autonomous systems face significant challenges in developing equally rich and adaptable representations of their environments. Despite being equipped with an array of sensors designed to capture multiple modalities of information, many artificial systems tend to prioritize and rely heavily on a single dominant modality—most often vision—over others [6,7]. This unidimensional focus results in internal representations that are limited in scope and lack the robustness necessary to handle situations where the primary sensory modality

 doi:10.20944/preprints202409.1969.v

is compromised or unavailable. For example, a robot that primarily depends on visual data may struggle to operate effectively in environments where visual information is obscured or degraded, such as in darkness or through visual noise. This limitation underscores the necessity for artificial agents to develop more sophisticated multimodal representations that can integrate and leverage information from various sensory inputs to ensure reliable and adaptive performance across diverse and unpredictable real-world scenarios.

To address these challenges, it is imperative to design artificial systems that can emulate the human ability to construct and utilize rich, multimodal representations. Such systems would be capable of integrating information from multiple sensory channels, thereby enhancing their ability to perceive, interpret, and interact with their environment in a more resilient and flexible manner. Moreover, the capability to perform cross-modal inference would enable these systems to compensate for missing or faulty sensory inputs, ensuring continuous and reliable operation even in the face of sensory disruptions or limitations.

Motivated by the intricate workings of human cognitive processes, we introduce **SensorySync**, a novel hierarchical multimodal variational autoencoder framework designed to advance the field of multimodal representation learning. Traditional multimodal generative models typically aim to learn a singular joint distribution that encompasses all input modalities [9–12]. While this approach allows for the generation of data across multiple modalities, it often results in a representation space that fails to capture the unique characteristics and complexities of each individual modality. Consequently, the expressive power of the joint representation is diminished, hindering the model's ability to perform effective cross-modal inferences.

In contrast, **SensorySync** adopts a hierarchical structure inspired by the Convergence-Divergence Zone (CDZ) cognitive model [2,65], which posits that humans process sensory information through a layered architecture. In this model, lower levels are responsible for processing modality-specific information, generating detailed representations unique to each sensory input. These modality-specific representations are then integrated at higher levels to form unified multimodal representations that encapsulate the combined information from all modalities [1,13]. By mirroring this hierarchical processing strategy, **SensorySync** is designed to maintain distinct representations for each modality while simultaneously learning a comprehensive joint distribution that facilitates cross-modal inference [66,67].

SensorySync operates as a hierarchical variational autoencoder, capable of learning modality-specific distributions for an arbitrary number of input modalities alongside a unified joint-modality distribution. This dual-level learning architecture ensures that the model preserves the intricate details of each individual modality while also capturing the interdependencies and correlations among them. A key innovation of **SensorySync** is the formal derivation of its evidence lower bound (ELBO), which serves as a rigorous foundation for optimizing the model's parameters. Furthermore, we introduce a novel methodology for approximating the joint-modality posterior by leveraging modality-specific representation dropout. This technique allows **SensorySync** to flexibly encode information from any subset of available modalities, inherently promoting the model's ability to infer missing modalities based on the available sensory inputs without incurring significant computational overhead.

To evaluate the efficacy of **SensorySync**, we conduct extensive experiments on a range of standard multimodal datasets that encompass diverse modalities and complex data structures. The results demonstrate that **SensorySync** not only matches but often surpasses the performance of existing state-of-the-art multimodal generative models in tasks involving both modality-specific reconstruction and cross-modal inference. These findings highlight the robustness and versatility of **SensorySync** in handling incomplete or noisy sensory data, thereby underscoring its potential to enhance the reliability and adaptability of artificial agents operating in real-world, multimodal environments.

In summary, the principal contributions of this paper are threefold:

- We introduce **SensorySync**, an innovative hierarchical multimodal variational autoencoder inspired by the human CDZ cognitive architecture [2]. **SensorySync** is engineered to learn both

 doi:10.20944/preprints202409.1969.v

modality-specific distributions and a unified joint distribution across an arbitrary number of modalities, thereby enabling effective cross-modal inference in scenarios with incomplete sensory information. A formal derivation of the model's evidence lower bound is provided to ensure a solid theoretical foundation for optimization.

- We present a novel approach for approximating the joint-modality posterior through modality-specific representation dropout. This methodology facilitates the encoding of information from any combination of available modalities, inherently supporting cross-modal inference during the training process. The proposed technique achieves this with minimal computational overhead, enhancing the model's scalability and efficiency.
- We conduct comprehensive evaluations of **SensorySync** on various standard multimodal datasets, demonstrating that it achieves performance levels comparable to, and in certain aspects exceeding, those of leading multimodal generative models. Specifically, **SensorySync** excels in tasks involving modality-specific reconstruction and cross-modal inference, underscoring its potential as a robust tool for comprehensive multimodal representation learning.

2. Related Work

The landscape of deep generative models has witnessed significant advancements, particularly in their ability to learn and represent complex latent structures within data. Among these, Variational Autoencoders (VAEs) have emerged as a cornerstone methodology for estimating deep generative models through the application of variational inference techniques [14,69]. VAEs operate by encoding input data into a latent space, typically assuming a univariate Gaussian prior distribution to regularize the latent representations [24,25,80]. This regularization ensures that the latent space captures meaningful variations in the data, facilitating tasks such as data generation and reconstruction. However, the inherent intractability of the marginal likelihood in VAEs necessitates the use of an inference network, which approximates the posterior distribution to compute the Evidence Lower Bound (ELBO) [26]. Techniques like importance sampling have been employed to estimate this lower bound more accurately, thereby enhancing the model's performance and stability [26,71].

Building upon the foundational VAE framework, hierarchical generative models have been proposed to capture more intricate relationships between latent variables [19,27–29,77]. These models introduce multiple layers of latent variables, allowing for a more nuanced and hierarchical representation of data. For instance, the Ladder VAE [19] incorporates a series of latent variables at different hierarchical levels, each responsible for capturing varying degrees of abstraction in the data. Similarly, hierarchical structures have been utilized to model complex dependencies and enhance the generative capabilities of VAEs [27,58]. Despite their enhanced representational power, these hierarchical models are predominantly designed for unimodal data, limiting their applicability in scenarios requiring the integration of multiple data modalities.

The extension of VAEs to handle multimodal data has been an area of active research, aiming to leverage the complementary information inherent in different data modalities. Multimodal VAEs seek to learn joint distributions over multiple data types, facilitating tasks such as cross-modal generation and inference [10,11,30,85,86]. A common approach in these models involves aligning the latent representations of individual modalities to form a unified latent space. This alignment enables the model to perform cross-modal inference, where the presence of one modality can inform the generation or reconstruction of another [10]. However, this strategy often necessitates the introduction of specific divergence terms in the ELBO for each possible combination of modalities, leading to scalability issues as the number of modalities increases. The complexity of managing multiple divergence terms can hinder the practical deployment of multimodal VAEs in environments with a large and diverse set of input modalities.

An alternative approach to multimodal VAEs is the Product-of-Experts (POE) inference network, which aims to streamline the encoding process by reducing the number of required encoding networks [12]. In the POE framework, each modality contributes its own expert network to the inference process, and the combined output is obtained by taking the product of the individual experts' distri-

 doi:10.20944/preprints202409.1969.v

butions. While this method effectively reduces the computational burden associated with multiple encoding networks, it introduces its own challenges. Specifically, the POE approach can lead to increased computational costs during training due to the necessity of maintaining and optimizing multiple expert networks simultaneously. Additionally, the product-based combination of experts may not always capture the nuanced interactions between different modalities, potentially limiting the model's ability to perform sophisticated cross-modal inferences.

Beyond VAEs, other generative modeling techniques have been explored for multimodal data integration. Generative Adversarial Networks (GANs), for example, have been adapted to handle multiple data modalities by designing architectures that can generate data across different types [47,88]. These multimodal GANs typically employ shared and modality-specific generators and discriminators to manage the diversity of input data. However, GANs often face challenges related to training stability and mode collapse, which can be exacerbated in the multimodal setting due to the increased complexity of the data distribution.

In addition to generative models, representation learning frameworks have been developed to facilitate the integration of multimodal data. Techniques such as Canonical Correlation Analysis (CCA) and its deep variants aim to learn correlated representations across different modalities by maximizing the mutual information between them [32]. While these methods excel at finding shared structures between modalities, they typically lack the generative capabilities required for tasks like data synthesis and reconstruction, which are essential for applications involving cross-modal inference.

Another notable direction in multimodal learning is the use of attention mechanisms and transformer-based architectures, which have demonstrated remarkable success in capturing complex dependencies across modalities [33]. These models employ self-attention layers to dynamically weigh the importance of different modalities, enabling more flexible and context-aware integration of multimodal information. Despite their strengths, transformer-based models can be computationally intensive and may require substantial amounts of training data to achieve optimal performance.

Despite the plethora of existing approaches, many of the current multimodal generative models share a common limitation: they rely on encoding information from all modalities into a single, unified latent space. This design choice often compromises the model's ability to maintain modality-specific generative capabilities, as the unified latent space must accommodate the diverse characteristics of each modality. Consequently, the expressive power of the latent representations for individual modalities may be constrained, hindering the model's performance in tasks requiring precise cross-modal inferences.

In contrast, our proposed framework, **SensorySync**, addresses these limitations by introducing a hierarchical representation structure that preserves both modality-specific and joint-modality latent spaces. Inspired by the Convergence-Divergence Zone (CDZ) cognitive model [2], **SensorySync** is designed to learn separate latent distributions for each modality while simultaneously capturing the dependencies and interactions between them in a unified joint latent space. This dual-level approach enables **SensorySync** to maintain the generative fidelity of individual modalities and enhance cross-modal inference capabilities without the scalability issues associated with existing multimodal VAE models. By leveraging modality-specific representation dropout, **SensorySync** efficiently approximates the joint-modality posterior, allowing for flexible and scalable integration of an arbitrary number of input modalities. This design not only preserves the unique characteristics of each modality but also facilitates robust cross-modal interactions, making **SensorySync** a versatile and powerful tool for comprehensive multimodal representation learning.

Furthermore, **SensorySync** distinguishes itself from prior approaches by avoiding the necessity of introducing complex divergence terms for each modality combination, thereby simplifying the optimization process and enhancing scalability. The hierarchical structure of **SensorySync** aligns closely with human cognitive processes, where lower-level sensory inputs are processed independently before being integrated into higher-level representations. This biologically inspired architecture

 doi:10.20944/preprints202409.1969.v

enables **SensorySync** to perform efficient and accurate cross-modal inferences, even in the presence of incomplete or noisy sensory data.

In summary, while existing deep generative models and multimodal learning frameworks have made significant strides in integrating and representing multimodal data, they often encounter challenges related to scalability, complexity, and the preservation of modality-specific generative capabilities. **SensorySync** addresses these challenges by introducing a hierarchical multimodal variational autoencoder that effectively balances the need for joint-modality representations with the preservation of individual modality characteristics. This innovative approach not only enhances the model's generative and inferential capabilities but also aligns with cognitive principles observed in human sensory processing, thereby paving the way for more robust and versatile artificial agents capable of operating in complex, multimodal environments.

3. Methodology of SensorySync

Deep generative models have demonstrated substantial capabilities in learning comprehensive and generalized representations of data across various domains. Among these models, the Variational Autoencoder (VAE) has gained prominence for its effectiveness in modeling single-modality data by learning a probabilistic latent space that captures the underlying structure of the input data [14]. The VAE framework operates by encoding input data $\mathbf{x}$ into a latent representation $\mathbf{z}$, which is typically of lower dimensionality compared to the original data. This latent vector serves as a compact and informative encoding that facilitates data reconstruction and generation.

The VAE models the joint distribution of the data and the latent variables as

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z}),$$

where $p(\mathbf{z})$ represents the prior distribution over the latent variables, often chosen to be a standard Gaussian distribution ($\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) for simplicity and computational tractability. The conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{z})$, parameterized by θ, defines the generative process of the data and is typically modeled using simple likelihood functions such as Bernoulli for binary data or Gaussian for continuous data.

Training the VAE involves maximizing the evidence likelihood $p(\mathbf{x})$, which requires integrating over the latent variables:

$$p(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z})\, d\mathbf{z} = \int_{\mathbf{z}} p_\theta(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}.$$

However, this integral is generally intractable due to the complexity of the joint distribution. To address this, the VAE employs an inference network $q_\phi(\mathbf{z}|\mathbf{x})$, parameterized by ϕ, which approximates the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$. This approximation allows for the derivation of the Evidence Lower Bound (ELBO) on the log-likelihood of the data:

$$\log p(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}) = EX_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL\big[q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})\big],$$

where the first term represents the expected log-likelihood (reconstruction loss) and the second term is the Kullback-Leibler (KL) divergence that regularizes the latent space by encouraging the approximate posterior to be close to the prior. The introduction of a hyperparameter β allows for controlling the trade-off between these two terms:

$$\mathcal{L}(\mathbf{x}) = EX_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, KL\big[q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})\big],$$

where $\beta = 1$ recovers the original VAE formulation. Optimization of the ELBO is typically performed using gradient-based methods, with the re-parametrization trick [14] facilitating the backpropagation through the stochastic latent variables.

Building upon the foundational principles of VAEs, we introduce **SensorySync**, an advanced hierarchical multimodal variational autoencoder designed to handle and integrate multiple data

 doi:10.20944/preprints202409.1969.v

modalities seamlessly. Unlike traditional VAEs that are limited to single-modality data, **SensorySync** is engineered to learn and represent complex relationships across an arbitrary number of modalities, thereby enabling robust cross-modal inference and data generation.

In a multimodal context, we consider a comprehensive set of N distinct modalities denoted by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} = \mathbf{x}_{1:N}$, each corresponding to different sensory inputs or data types. These modalities are assumed to be generated through an environment-dependent process modeled by the joint distribution $p_\theta(\mathbf{x}_{1:N})$, parameterized by θ. **SensorySync** conceptualizes the generation process hierarchically, where each modality is associated with its own *modality-specific* latent variable within the set $\mathbf{Z}^m = \{\mathbf{z}^m_{1:N}\}$. These modality-specific latent variables are conditionally independent given a central *core* latent variable $\mathbf{z}^c$, which encapsulates the shared information across all modalities.

The primary objective of **SensorySync** is twofold: to learn distinct latent spaces for each modality that facilitate accurate reconstruction of modality-specific data, and to simultaneously model a joint distribution that captures the interdependencies among all modalities, thereby enabling cross-modal inference. This dual capability ensures that **SensorySync** maintains high fidelity in representing individual modalities while also leveraging the collective information to infer missing or unobserved modalities.

3.1. Evidence Lower-Bound of SensorySync

To train **SensorySync**, we aim to maximize the likelihood of the joint generative process $p_\theta(\mathbf{x}_{1:N})$. This involves marginalizing over both the modality-specific latent variables $\mathbf{z}^m_{1:N}$ and the core latent variable $\mathbf{z}^c$:

$$p_\theta(\mathbf{X}) = \int_{\mathbf{z}^c} \int_{\mathbf{z}^m_{1:N}} p(\mathbf{x}_{1:N}, \mathbf{z}^m_{1:N}, \mathbf{z}^c)\, d\mathbf{z}^m_{1:N}\, d\mathbf{z}^c. \tag{1}$$

Given the hierarchical structure and the conditional independence assumptions inherent in **SensorySync**, the joint distribution can be decomposed as

$$p_\theta(\mathbf{X}) = \int_{\mathbf{z}^c} \int_{\mathbf{z}^m_{1:N}} p(\mathbf{z}^c) \prod_{i=1}^{N} p_\theta(\mathbf{x}_i|\mathbf{z}^m_i)\, p_\theta(\mathbf{z}^m_i|\mathbf{z}^c)\, d\mathbf{z}^m_{1:N}\, d\mathbf{z}^c. \tag{2}$$

This factorization highlights the role of the core latent variable $\mathbf{z}^c$ in governing the generation of each modality-specific latent variable $\mathbf{z}^m_i$, and consequently, the generation of each modality $\mathbf{x}_i$.

Similar to the single-modality VAE, the marginal likelihood $p_\theta(\mathbf{x}_{1:N})$ is intractable due to the high-dimensional integral over the latent variables. To circumvent this, we introduce an inference network $q_\phi(\mathbf{z}^m_{1:N}, \mathbf{z}^c)$, parameterized by ϕ, which serves as an approximation to the true posterior distribution $p_\theta(\mathbf{z}^m_{1:N}, \mathbf{z}^c|\mathbf{x}_{1:N})$. The inference model is designed to simultaneously encode information from all modalities into their respective modality-specific latent spaces as well as into the core latent space, resulting in the factorization

$$q_\phi(\mathbf{z}^m_{1:N}, \mathbf{z}^c) = q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N}) \prod_{i=1}^{N} q_\phi(\mathbf{z}^m_i|\mathbf{x}_i). \tag{3}$$

This structure ensures that each modality contributes to the overall latent representation, while the core latent variable $\mathbf{z}^c$ captures the shared information across modalities.

By integrating the inference model into the joint distribution and re-expressing the evidence likelihood as an expectation over the latent variables, we derive the following expression:

$$p_\theta(\mathbf{x}_{1:N}) = EX_{\substack{q_\phi(\mathbf{z}^m_{1:N}|\mathbf{x}_{1:N}) \\ q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N})}} \left[\frac{p(\mathbf{z}^c)}{q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N})} \prod_{i=1}^{N} \frac{p_\theta(\mathbf{x}_i|\mathbf{z}^m_i)\, p_\theta(\mathbf{z}^m_i|\mathbf{z}^c)}{q_\phi(\mathbf{z}^m_i|\mathbf{x}_i)} \right]. \tag{4}$$

 doi:10.20944/preprints202409.1969.v

Applying the logarithm and Jensen's inequality to this expression, we obtain a lower bound on the log-likelihood of the evidence, known as the ELBO:

$$\mathcal{L}(\mathbf{X}) = EX_{\substack{q_\phi(\mathbf{z}^m_{1:N}|\mathbf{x}_{1:N}) \\ q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N})}} \log \left[\frac{p(\mathbf{z}^c)}{q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N})} \prod_{i=1}^{N} \frac{p_\theta(\mathbf{x}_i|\mathbf{z}^m_i)\, p_\theta(\mathbf{z}^m_i|\mathbf{z}^c)}{q_\phi(\mathbf{z}^m_i|\mathbf{x}_i)} \right]. \tag{5}$$

This lower bound encapsulates three primary components: the reconstruction loss for each modality, the regularization of the core latent space, and the alignment between modality-specific and core latent spaces.

The first component mirrors the original VAE's reconstruction loss, extending it to multiple modalities. For each modality i, the reconstruction loss is expressed as

$$EX_{q_\phi(\mathbf{z}^m_i|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}^m_i)]. \tag{6}$$

This term ensures that each modality-specific latent variable $\mathbf{z}^m_i$ effectively captures the information necessary to reconstruct its corresponding input $\mathbf{x}_i$.

The second component addresses the regularization of the core latent variable $\mathbf{z}^c$. It imposes a constraint on the approximate posterior $q_\phi(\mathbf{z}^c \mid \mathbf{x}_{1:N})$ to be close to the prior distribution $p(\mathbf{z}^c)$, thereby maintaining a structured and meaningful latent space:

$$-KL\left[q_\phi(\mathbf{z}^c \mid \mathbf{x}_{1:N}) \,||\, p(\mathbf{z}^c)\right]. \tag{7}$$

This KL divergence term ensures that the core latent space does not deviate excessively from the prior, promoting generalization and preventing overfitting.

The third component facilitates the alignment between the modality-specific latent distributions and the core latent distribution. For each modality i, this is represented as

$$-EX_{q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N})}\left[KL\left[q_\phi(\mathbf{z}^m_i|\mathbf{x}_i) \,||\, p_\theta(\mathbf{z}^m_i|\mathbf{z}^c)\right]\right]. \tag{8}$$

This expectation over the core latent space encourages the modality-specific encoders to produce latent representations that are consistent with the core distribution, thereby fostering coherence and interoperability among different modalities.

Aggregating these components, the final form of the ELBO for **SensorySync** is given by

$$\begin{aligned} \mathcal{L}(\mathbf{x}_{1:N}) = & \sum_{i=1}^{N} \lambda_i\, EX_{q_\phi(\mathbf{z}^m_i|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}^m_i)] \\ & - \sum_{i=1}^{N} \beta^m_i EX_{q_\phi(\mathbf{z}^c|\mathbf{x}_{1:N})}\left[KL\left[q_\phi(\mathbf{z}^m_i|\mathbf{x}_i) \,||\, p_\theta(\mathbf{z}^m_i|\mathbf{z}^c)\right]\right] \\ & - \beta^c\, KL\left[q_\phi(\mathbf{z}^c \mid \mathbf{x}_{1:N}) \,||\, p(\mathbf{z}^c)\right], \end{aligned} \tag{9}$$

where λ_i are weighting factors for each modality-specific reconstruction loss, and β^m_i, along with β^c, are hyperparameters that control the strength of the regularization terms for modality-specific and core latent spaces, respectively. This formulation allows for flexible tuning of the model's capacity to reconstruct data and maintain coherent latent representations.

3.2. Modality Representation Dropout

In this section, we delve into the methodology employed to approximate the joint-modality posterior distribution within the **SensorySync** framework. Specifically, our objective is to effectively encode information from multiple modality-specific datasets $\mathbf{x}_{1:N}$ into a unified multimodal core latent variable $\mathbf{z}^c$. Achieving this requires a robust and computationally efficient approach to integrate diverse sensory inputs while maintaining the integrity of each modality's unique characteristics.

 doi:10.20944/preprints202409.1969.v

One prevalent method for approximating the joint posterior in multimodal settings is the Product-of-Experts (POE) approach [12]. The POE technique constructs the joint posterior by taking the product of Gaussian experts, which includes a prior expert to incorporate prior knowledge [12]. While POE has demonstrated efficacy in capturing the dependencies between different modalities, it presents significant computational challenges. Specifically, POE necessitates the artificial sub-sampling of observations during the training phase, leading to increased computational overhead. Moreover, the POE approach is prone to overconfident predictions from individual experts, which can degrade the quality of cross-modality inferences [9]. This overconfidence arises because each expert independently contributes to the joint posterior without adequately accounting for the uncertainty inherent in combining multiple modality-specific distributions.

To address these limitations, we introduce a novel methodology termed **Modality Representation Dropout** within the **SensorySync** architecture. This approach leverages the concept of dropout, a regularization technique traditionally used to prevent overfitting in neural networks, by applying it to modality-specific representations. The core idea is to randomly deactivate certain modality-specific hidden representations during training, thereby encouraging the model to rely on complementary information from the remaining active modalities. This stochastic deactivation promotes a more balanced integration of modalities and mitigates the risk of any single modality dominating the joint posterior.

Formally, we define modality-data dropout masks $\mathbf{d}$, where each mask element d_i corresponds to the i-th modality and is sampled from a Bernoulli distribution. The dimensionality of $\mathbf{d}$ is equal to the number of modalities, i.e., $|\mathbf{d}| = N$. The dropout operation is mathematically represented as

$$\mathbf{h}^d = \mathbf{d} \odot \mathbf{h}, \tag{10}$$

where $\mathbf{h} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$ denotes the set of hidden-layer representations produced by the modality-specific encoders. The operator $\odot$ signifies element-wise multiplication, effectively zeroing out the selected modality representations based on the dropout mask $\mathbf{d}$. Specifically, for each modality i, the representation is nullified if $d_i = 1$, as given by

$$\mathbf{h}_i = \mathbf{0}, \text{ if } d_i = 1. \tag{11}$$

This selective deactivation ensures that during each training iteration, a random subset of modalities is omitted from contributing to the core latent representation. Consequently, the model is compelled to infer the missing modalities using the information from the active ones, thereby enhancing its cross-modality inference capabilities.

The dropout masks $\mathbf{d}$ are sampled from a Bernoulli distribution with parameters $\mathbf{w} = \{w_1, \ldots, w_N\}$, where each w_i controls the dropout probability for the corresponding modality:

$$\mathbf{d} \sim \text{Bernoulli}(w_1, \ldots, w_N), \quad \text{with } \sum_{i=1}^{N} d_i \geq 1. \tag{12}$$

To maintain the integrity of the multimodal representation, we impose a constraint that ensures at least one modality remains active during each dropout operation. This is achieved by conditioning the mask sampling procedure such that the sum of the dropout mask elements is at least one, i.e., $\sum_{i=1}^{N} d_i \geq 1$. This condition prevents the complete deactivation of all modalities, which would otherwise render the core latent variable $\mathbf{z}^c$ devoid of meaningful information.

Once the dropout mask is applied, the resulting representations $\mathbf{h}^d$ are concatenated and fed into the multimodal encoder, which synthesizes the core latent variable $\mathbf{z}^c$. This process effectively integrates information from the active modalities while maintaining the flexibility to handle varying combinations of input modalities.

 doi:10.20944/preprints202409.1969.v

Incorporating the modality representation dropout into the **SensorySync** framework necessitates a modification of the Evidence Lower Bound (ELBO) to account for the stochastic nature of the dropout operation. The revised ELBO is expressed as

$$\begin{aligned}\mathcal{L}(\mathbf{X}) = &\sum_{i=1}^{N} \lambda_i \, EX_{q_\phi(\mathbf{z}_i^m|\mathbf{x}_i)}[\log p_\theta(\mathbf{x}_i|\mathbf{z}_i^m)] \\ &- \sum_{i=1}^{N} \beta_i^m EX_{q_\phi(\mathbf{z}^c|\mathbf{h}^d)}\left[KL\big(q_\phi(\mathbf{z}_i^m|\mathbf{x}_i) \,||\, p_\theta(\mathbf{z}_i^m|\mathbf{z}^c)\big)\right] \\ &- \beta^c \, KL\Big(q_\phi(\mathbf{z}^c \mid \mathbf{h}^d) \,||\, p(\mathbf{z}^c)\Big). \qquad (13)\end{aligned}$$

(14)

Here, the ELBO comprises three primary components:

1. **Reconstruction Loss**: The first term represents the reconstruction loss for each modality, weighted by factors λ_i. This term ensures that the modality-specific latent variables $\mathbf{z}_i^m$ can accurately reconstruct their respective inputs $\mathbf{x}_i$.
2. **Modality-Specific KL Divergence**: The second term enforces a regularization constraint on the modality-specific latent variables $\mathbf{z}_i^m$. Weighted by factors β_i^m, it minimizes the Kullback-Leibler (KL) divergence between the approximate posterior $q_\phi(\mathbf{z}_i^m|\mathbf{x}_i)$ and the conditional prior $p_\theta(\mathbf{z}_i^m|\mathbf{z}^c)$, thereby aligning the modality-specific representations with the core latent space.
3. **Core KL Divergence**: The final term applies a regularization constraint on the core latent variable $\mathbf{z}^c$, weighted by β^c. It minimizes the KL divergence between the approximate posterior $q_\phi(\mathbf{z}^c \mid \mathbf{h}^d)$ and the prior $p(\mathbf{z}^c)$, ensuring that the core latent space adheres to the desired prior distribution.

The introduction of modality representation dropout into the ELBO formulation enhances the **SensorySync** model's ability to generalize across different combinations of input modalities. By randomly deactivating certain modality-specific representations during training, the model becomes adept at handling scenarios where some modalities may be missing or noisy, thereby improving its robustness and flexibility in real-world applications. Furthermore, this dropout mechanism facilitates more efficient training by reducing the dependency on any single modality, thus preventing the model from becoming overly reliant on dominant modalities. As a result, **SensorySync** achieves a more balanced and comprehensive integration of multimodal data, enhancing its capacity for accurate cross-modality inference and data reconstruction.

4. Experiment

In this section, we rigorously assess the performance of **SensorySync** as a multimodal generative model utilizing widely recognized multimodal datasets. Our evaluations demonstrate that **SensorySync** surpasses existing state-of-the-art generative models in tasks involving joint-modality reconstruction from arbitrary input modalities and cross-modality inference.

4.1. Multimodal Datasets

Consistent with established methodologies in prior research, we convert single-modality datasets into bimodal datasets by treating the label associated with each image as a distinct modality. This approach facilitates a more comprehensive evaluation of multimodal generative capabilities. Additionally, we benchmark **SensorySync** against prominent multimodal generative models, specifically JMVAE-kl [10] and MVAE [12]. For the JMVAE-kl model, we set the hyperparameter $\alpha = 0.01$. Regarding the MVAE model, we utilize the publicly available official implementation [1] and adhere to the training hyperparameters recommended by the authors to ensure a fair comparison.

[1] Implementation available at https://github.com/mhw32/multimodal-vae-public

 doi:10.20944/preprints202409.1969.v

Our evaluation encompasses standard datasets widely recognized in the literature: MNIST [16], FashionMNIST [17], and CelebA [18]. These datasets are selected for their diverse characteristics and the established benchmarks they provide in generative modeling and cross-modality tasks. Notably, we report that **SensorySync** achieves state-of-the-art performance on the first two datasets concerning both generative modeling and cross-modality inference capabilities.

For training **SensorySync**, we adopt a straightforward configuration without hyperparameter tuning, specifically setting $\alpha_i = \beta_i^m = \beta^c = 1, \forall i \in [1, N]$. Additionally, we fix the dropout hyperparameters at $\mathbf{w}_{1:N} = 0.5$ for all modalities to maintain consistency across experiments. The architecture of **SensorySync** comprises two distinct network types: the modality network and the core network. The modality network is responsible for encoding input data into modality-specific latent spaces $\mathbf{z}^m$, generating the corresponding hidden representations $\boldsymbol{h}$, and facilitating the inverse generative process. The core network, on the other hand, encodes the multimodal core latent variable from the representation $\mathbf{h}^d$, which is subsequently used to generate the modality-specific latent spaces $\mathbf{z}^m$.

To ensure a fair comparison across different models, we maintain consistent network architectures for the generative and inference networks of the baseline models, aligning them with the modality-specific networks of **SensorySync** on each dataset. This uniformity allows us to isolate the impact of our proposed hierarchical approach from architectural variations.

Furthermore, we incorporate a warm-up period for the regularization terms of the Evidence Lower Bound (ELBO) [19]. Specifically, we linearly increase the value of the prior regularization term on the modality-specific latent variables over U_m epochs and similarly increase the value of the Gaussian prior on the core latent space over U_c epochs. For the baseline models, we apply a single warm-up period on the prior regularization of the latent space, denoted as U_b.

Our evaluation framework encompasses both reconstruction capabilities and cross-modality inference performance. To quantify these aspects, we estimate the image marginal log-likelihood, $\log p(\mathbf{x}_1)$, the joint log-likelihood, $\log p(\mathbf{x}_1, \mathbf{x}_2)$, and the conditional log-likelihood, $\log p(\mathbf{x}_1|\mathbf{x}_2)$, using importance sampling techniques. Specifically, we employ $5,000$ importance samples for MNIST and FashionMNIST, and 500 samples for CelebA. Detailed derivations of these evaluation metrics are provided in the appendix.

4.1.1. MNIST

For the MNIST dataset, we train all models on grayscale images $\mathbf{x}_1 \in \mathbb{R}^{28\times28}$ and their corresponding labels $\mathbf{x}_2 \in \{0, 1\}^{10}$. The dataset is partitioned into 85% for training (with 10% of this subset reserved for validation) and the remaining 15% for evaluation.

The architecture of the image modality network in **SensorySync** consists of three linear layers, each with 512 hidden units, utilizing leaky rectifier activation functions and incorporating batch normalization between layers to enhance training stability and convergence. We allocate a 16-dimensional latent space for the image modality. Similarly, the label modality network comprises three linear layers with 128 hidden units, maintaining a 16-dimensional latent space for the label modality. The core network is structured with three linear layers, each containing 64 hidden units, and is responsible for a 10-dimensional latent space. In contrast, the baseline models employ a unified 26-dimensional latent space for all modalities.

We model the conditional distribution $p(\mathbf{x}_1|\mathbf{z}_1)$ as a Bernoulli distribution to accommodate the binary nature of MNIST images, and $p(\mathbf{x}_2|\mathbf{z}_2)$ as a multinomial distribution suitable for the categorical labels. For the training of **SensorySync**, we set the warm-up parameters to $U_m = 100$ epochs for modality-specific latent variables and $U_c = 200$ epochs for the core latent space. The baseline models undergo a single warm-up period of $U_b = 200$ epochs.

All models are trained for a total of 500 epochs using a learning rate of $l = 10^{-3}$ and a batch size of $b = 64$. The test log-likelihood estimates for each model are summarized in Table 1. The results indicate that **SensorySync** outperforms other state-of-the-art multimodal generative models across both single-modality and joint-modality metrics. This superior performance is achieved despite

 doi:10.20944/preprints202409.1969.v

the modality-specific latent spaces in **SensorySync** being of lower dimensionality compared to the unified latent space utilized by JMVAE and MVAE. Furthermore, **SensorySync** demonstrates enhanced cross-modality inference capabilities, as evidenced by the significantly lower conditional log-likelihood $\log p(\mathbf{x}_1|\mathbf{x}_2)$, achieved by leveraging only the lower-dimensional label modality. Panels (a), (b), and (c) depict images conditionally generated by sampling from the posterior distribution $q_\phi(\mathbf{z}_c|\mathbf{x}_2)$ with specific label inputs, such as $x_2 = 3$. Panels (d), (e), and (f) showcase images generated by sampling directly from the prior distribution $\mathbf{z}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The quality and diversity of these samples affirm the efficacy of the generative networks within **SensorySync**.

Table 1. Log-likelihood metrics on the MNIST dataset comparing **SensorySync** with other multimodal generative models. Latent variables are estimated using image (I), label (L), or joint (I, L) modalities with 5000 importance samples. The MVAE baseline was not evaluated due to numerical instabilities.

Metric	Input	JMVAE	MVAE	**SensorySync**
$\log p(\mathbf{x}_1)$	I	-90.189	-	**-89.050**
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I	-90.241	-	**-89.183**
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	L	-125.381	-	**-121.401**
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I,L	-90.335	-	**-89.143**
$\log p(\mathbf{x}_1|\mathbf{x}_2)$	L	-123.070	-	**-118.856**

4.1.2. FashionMNIST

For the FashionMNIST dataset, we train the generative models on grayscale images $\mathbf{x}_1 \in \mathbb{R}^{1\times28\times28}$ and their corresponding class labels $\mathbf{x}_2 \in \{0, 1\}^{10}$. The dataset is similarly divided into 85% for training (with 10% reserved for validation) and 15% for evaluation, maintaining consistency with the MNIST evaluation protocol.

The image modality network in **SensorySync** is implemented using a miniature Deep Convolutional Generative Adversarial Network (DCGAN) [20], enhanced with Swish activation functions [21] to capitalize on their superior performance in deep convolutional architectures. This network comprises two convolutional layers with 32 and 64 channels, respectively, followed by a linear layer with 128 hidden units. The core and label-modality inference and generator networks retain the same architectural configuration as described in the MNIST evaluation, ensuring uniformity across different datasets. The latent spaces for both modality-specific and core representations are maintained at 16 and 10 dimensions, respectively, mirroring the settings used in the MNIST experiments.

All models undergo training for 500 epochs using the Adam optimization algorithm [22] with a learning rate of 10^{-3} and a batch size of $b = 64$. The test log-likelihood estimates for each model are presented in Table 2. The results reaffirm that **SensorySync** consistently outperforms other state-of-the-art multimodal generative models across both single-modality and joint-modality metrics. Additionally, **SensorySync** excels in label-to-image cross-modality inference, demonstrating its robust capability to infer missing modalities effectively. The generated samples exhibit high fidelity and diversity, underscoring the model's proficiency in capturing the underlying distribution of the FashionMNIST data. These results highlight the effectiveness of **SensorySync** in leveraging multimodal information to generate coherent and contextually relevant images.

 doi:10.20944/preprints202409.1969.v

Table 2. Log-likelihood metrics on the FashionMNIST dataset comparing **SensorySync** with other multimodal generative models. Latent variables are estimated using image (I), label (L), or joint (I, L) modalities with 5000 importance samples.

Metric	Input	JMVAE	MVAE	**SensorySync**
$\log p(\mathbf{x}_1)$	I	-232.427	-236.613	**-231.753**
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I	-232.739	-242.628	**-232.276**
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	L	-244.378	-557.582	**-243.932**
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I,L	-232.573	-241.534	**-232.248**
$\log p(\mathbf{x}_1 \vert \mathbf{x}_2)$	L	-242.060	-552.679	**-241.662**

4.1.3. CelebA

The CelebA dataset presents a more complex and diverse challenge, consisting of re-scaled colored images $\mathbf{x}_1 \in \mathbb{R}^{3\times64\times64}$ and a subset of 18 visually distinctive attributes $\mathbf{x}_2 \in \{0,1\}^{18}$ [23]. This dataset is selected for its rich attribute annotations, enabling comprehensive multimodal evaluation of generative models.

For **SensorySync**, the image modality network is constructed using a miniature DCGAN architecture, consisting of four convolutional layers with 32, 64, 128, and 256 channels, respectively, followed by a linear layer with 512 hidden units. This architecture is designed to effectively capture the intricate features present in high-resolution colored images. The image-specific latent space is set to 48 dimensions, allowing for a nuanced representation of visual data. The label modality network comprises three linear layers with 512 hidden units, maintaining a 48-dimensional latent space for attribute representations. The core network is structured with three linear layers containing 256 hidden units, facilitating a compact 16-dimensional latent space that integrates information across modalities. In contrast, baseline models utilize a unified 64-dimensional latent space to accommodate all modalities.

Training configurations for CelebA involve 50 epochs with a learning rate of 10^{-4} and a batch size of $b = 128$. For **SensorySync**, the warm-up periods are set to $U_m = 5$ epochs for modality-specific latent variables and $U_c = 10$ epochs for the core latent space. Baseline models are subjected to a single warm-up period of $U_b = 10$ epochs.

The test log-likelihood estimates for each model on the CelebA dataset are detailed in Table 4. The findings indicate that **SensorySync** performs competitively with other state-of-the-art multimodal generative models across all evaluation metrics. While **SensorySync** maintains comparable performance, it exhibits slightly reduced performance relative to previous evaluations, likely attributable to the increased complexity and higher dimensionality of the CelebA dataset. These generated samples demonstrate the model's capability to synthesize realistic and attribute-consistent images, validating the effectiveness of **SensorySync** in handling complex multimodal data distributions inherent in the CelebA dataset.

Table 3. Log-likelihood metrics on the CelebA dataset comparing **SensorySync** with other multimodal generative models. Latent variables are estimated using image (I), attributes (A), or joint (I, A) modalities with 500 importance samples.

Metric	Input	JMVAE	MVAE	**SensorySync**
$\log p(\mathbf{x}_1)$	I	-6260.35	**-6256.65**	-6271.35
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I	**-6264.59**	-6270.86	-6278.19
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	A	**-7204.36**	-7316.12	-7303.64
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I,A	**-6262.67**	-6266.14	-6276.57
$\log p(\mathbf{x}_1 \vert \mathbf{x}_2)$	A	**-7191.11**	-7309.10	-7296.22

 doi:10.20944/preprints202409.1969.v

Table 4. Log-likelihood metrics on the CelebA dataset comparing **SensorySync** with other multimodal generative models. Latent variables are estimated using image (I), attributes (A), or joint (I, A) modalities with 500 importance samples.

Metric	Input	JMVAE	MVAE	**SensorySync**
$\log p(\mathbf{x}_1)$	I	-6260.35	**-6256.65**	-6271.35
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I	**-6264.59**	-6270.86	-6278.19
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	A	**-7204.36**	-7316.12	-7303.64
$\log p(\mathbf{x}_1, \mathbf{x}_2)$	I,A	**-6262.67**	-6266.14	-6276.57
$\log p(\mathbf{x}_1 \vert \mathbf{x}_2)$	A	**-7191.11**	-7309.10	-7296.22

4.2. Discussion

In this section, we critically analyze the performance of **SensorySync** in comparison to established state-of-the-art multimodal generative models across various benchmark datasets. Our evaluation framework encompassed a comprehensive comparison with the Joint Multimodal Variational Autoencoder with KL divergence (**JMVAE-kl**) [10] and the Multimodal Variational Autoencoder (**MVAE**) [12], both of which are widely recognized for their efficacy in multimodal representation learning.

The experimental results obtained from increasingly complex datasets underscore the pivotal role of hierarchical representation spaces in effectively modeling multimodal data distributions. **SensorySync** leverages a hierarchical architecture that disentangles modality-specific latent spaces from a central core latent space, allowing for a more nuanced and scalable integration of multiple modalities. This architectural distinction is particularly advantageous when dealing with diverse and high-dimensional data, as it facilitates the preservation of unique modality characteristics while enabling robust joint-modality inference.

On the MNIST and FashionMNIST datasets, **SensorySync** consistently outperformed both **JMVAE-kl** and **MVAE** across all evaluated metrics, including image marginal log-likelihood, joint log-likelihood, and conditional log-likelihood. Notably, **SensorySync** achieved these superior results despite utilizing lower-dimensional modality-specific latent spaces compared to the unified latent spaces employed by the baseline models. This demonstrates the efficacy of hierarchical latent structures in capturing essential data features without necessitating an increase in latent dimensionality, thereby enhancing computational efficiency and scalability.

The exceptional performance of **SensorySync** on these datasets can be attributed to its ability to maintain distinct representations for each modality while concurrently learning a cohesive joint distribution. This dual-level representation strategy ensures that each modality retains its unique generative capabilities, which is crucial for tasks requiring precise reconstruction and high-fidelity cross-modality inference. Additionally, the incorporation of modality representation dropout within **SensorySync** promotes resilience against missing or noisy modalities, further enhancing its robustness and versatility in real-world applications.

However, when evaluating **SensorySync** on the more complex CelebA dataset, the model exhibited performance on par with the baseline models rather than surpassing them. This observation prompts a deeper investigation into the influence of latent space dimensionality on performance in high-dimensional and intricate data scenarios. In the CelebA experiments, **SensorySync** was constrained to lower-dimensional latent spaces to ensure a fair comparison with the baselines, which utilize a unified higher-dimensional latent space. This dimensionality constraint likely contributed to the slightly diminished log-likelihood scores observed for **SensorySync** on this dataset.

The CelebA results highlight a potential trade-off between latent space dimensionality and model performance in complex multimodal environments. While lower-dimensional latent spaces offer computational advantages and prevent overfitting, they may limit the model's capacity to capture the rich and diverse features inherent in high-resolution and attribute-rich datasets like CelebA. Consequently, **SensorySync** may benefit from adaptive strategies that dynamically balance

 doi:10.20944/preprints202409.1969.v

the representational capacity between the core and modality-specific latent spaces, especially when scaling to more complex data distributions.

Despite these challenges, **SensorySync** maintained its competitive edge in joint-modality and cross-modality inference tasks on CelebA, outperforming the MVAE model in scenarios where only partial modality information (e.g., labels) was available for inference. This underscores the strength of **SensorySync**'s hierarchical architecture in facilitating effective cross-modality interactions, even when operating under dimensionality constraints.

Looking forward, future research directions could explore the optimization of latent space dimensionality in **SensorySync** to better accommodate complex datasets without sacrificing computational efficiency. Additionally, integrating adaptive mechanisms that allow the model to adjust the balance between core and modality-specific representations based on the complexity of the input data could further enhance performance. Exploring advanced regularization techniques and more sophisticated dropout strategies might also provide avenues for improving the model's ability to generalize across diverse multimodal scenarios.

Furthermore, extending **SensorySync** to handle more than two modalities and evaluating its performance on a wider array of datasets with varying degrees of complexity and modality diversity would provide a more comprehensive understanding of its strengths and limitations. Investigating the model's performance in real-world applications, such as autonomous driving or robotics, where multimodal data integration is critical, could also validate its practical utility and inform further refinements.

In conclusion, **SensorySync** demonstrates significant advancements in multimodal generative modeling through its hierarchical representation architecture. Its ability to outperform existing models on standard datasets like MNIST and FashionMNIST, coupled with competitive performance on more complex datasets like CelebA, highlights its potential as a robust and scalable solution for multimodal representation learning. Addressing the identified challenges related to latent space dimensionality and exploring adaptive representational strategies will be key to unlocking the full potential of **SensorySync** in future applications.

5. Conclusions and Future Directions

In this study, inspired by the intricacies of the human cognitive framework, we introduced **SensorySync**, an innovative multimodal hierarchical generative model. **SensorySync** is designed to effectively learn and disentangle separate modality-specific representations alongside a cohesive joint-modality representation. This hierarchical approach offers a significant advancement over traditional multimodal generative models that rely on a single, unified representation space. By maintaining distinct latent spaces for each modality, **SensorySync** enhances the model's ability to capture the unique characteristics and complexities inherent to each data modality, thereby facilitating more nuanced and accurate representation learning.

Our experimental evaluations on standard multimodal datasets, including MNIST, FashionMNIST, and CelebA, demonstrated that **SensorySync** consistently outperforms existing state-of-the-art multimodal generative models in both modality-specific reconstruction and cross-modality inference tasks. These results underscore the efficacy of the hierarchical latent structure in improving the model's generative capabilities and its ability to perform robust inferences across different modalities. The superior performance of **SensorySync** is particularly noteworthy given that it achieves these results without extensive hyperparameter tuning, highlighting the robustness and adaptability of our proposed architecture.

A key innovation introduced in **SensorySync** is the methodology for approximating the joint-modality posterior distribution through modality-specific representation dropout. This technique involves randomly deactivating certain modality-specific representations during training, which encourages the model to rely on complementary information from the remaining active modalities. As a result, **SensorySync** is capable of encoding information from an arbitrary number of modalities

 doi:10.20944/preprints202409.1969.v

with minimal computational overhead. This dropout-based approach not only enhances the model's ability to perform cross-modality inference but also contributes to its resilience against missing or noisy data, thereby broadening its applicability in real-world scenarios where data imperfections are commonplace.

Looking ahead, there are several promising avenues for future research and development of **SensorySync**. One immediate direction is the exploration of scenarios involving a larger number of modalities. Extending **SensorySync** to handle more complex and diverse multimodal data can further validate its scalability and robustness. This expansion would involve addressing challenges related to the integration and coordination of additional modality-specific latent spaces, ensuring that the hierarchical structure remains effective as the number of modalities increases.

Furthermore, we aim to leverage **SensorySync** as a foundational perceptual representation model for artificial agents. Integrating **SensorySync** within deep multimodal reinforcement learning frameworks could enable agents to perform more sophisticated tasks by utilizing cross-modality inference to interpret and interact with their environment. For instance, in autonomous driving, an agent could infer visual information from auditory cues in low-visibility conditions, thereby enhancing its decision-making capabilities and operational reliability.

In addition to its applications in reinforcement learning, **SensorySync** holds significant potential in the domain of perceptual learning. Drawing further inspiration from human cognitive processes, we intend to investigate reinforcement learning mechanisms that can dynamically construct and refine the multimodal representations learned by **SensorySync**. This involves developing adaptive algorithms that allow the model to adjust its hierarchical latent structure based on the complexity and variability of the input data, thereby improving its capacity for continuous learning and adaptation.

Another important future direction is the incorporation of advanced regularization techniques and optimization strategies to further enhance the performance and stability of **SensorySync**. Exploring methods such as variational dropout, adversarial training, and hierarchical Bayesian approaches could provide additional layers of robustness and flexibility, enabling the model to better handle a wider range of data distributions and inference tasks.

Moreover, evaluating **SensorySync** in real-world applications, such as healthcare diagnostics, multimedia content generation, and human-computer interaction, can provide valuable insights into its practical utility and effectiveness. These applications often require the integration of diverse data types and the ability to perform accurate cross-modality inferences, making them ideal testbeds for showcasing the strengths of our proposed model.

In summary, **SensorySync** represents a significant step forward in the field of multimodal generative modeling, offering a robust and scalable solution for learning and integrating complex multimodal data. Its hierarchical representation structure and innovative dropout-based posterior approximation method enable it to outperform existing models in key generative and inferential tasks. As we continue to explore and expand the capabilities of **SensorySync**, we anticipate that it will become a valuable tool for advancing artificial intelligence systems that require sophisticated multimodal understanding and interaction.

References

1. Kaspar Meyer and Antonio Damasio. Convergence and divergence in a neural architecture for recognition and memory. *Trends in neurosciences*, 32(7):376–382, 2009.
2. Antonio R Damasio. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1-2):25–62, 1989.
3. Peter Walker, J Gavin Bremner, Uschi Mason, Jo Spring, Karen Mattock, Alan Slater, and Scott P Johnson. Preverbal infantsâ€™ sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, 21(1):21–25, 2010.
4. Daphne Maurer, Thanujeni Pathman, and Catherine J Mondloch. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science*, 9(3):316–322, 2006.

 doi:10.20944/preprints202409.1969.v

5. Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995, 2011.
6. Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.
7. Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016.
8. Rui Silva, Miguel Vasco, Francisco S. Melo, Ana Paiva, and Manuela Veloso. Playing games in the dark: An approach for cross-modality transfer in reinforcement learning. *arXiv:1911.12851 [cs]*, November 2019. arXiv: 1911.12851.
9. Yuge Shi, N Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, pages 15692–15703, 2019.
10. Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.
11. Timo Korthals, Daniel Rudolph, Jürgen Leitner, Marc Hesse, and Ulrich Rückert. Multi-modal generative models for learning epistemic active sensing. In *2019 IEEE International Conference on Robotics and Automation*, 2019.
12. Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5575–5585, 2018.
13. Stephane Lallee and Peter Ford Dominey. Multi-modal convergence maps: from body schema and self-representation to mental imagery. *Adaptive Behavior*, 21(4):274–285, 2013.
14. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
15. Johan Ludwig William Valdemar Jensen et al. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30:175–193, 1906.
16. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
17. Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
18. Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.
19. Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.
20. Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
21. Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.
22. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
23. Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.
24. Jianlin Su and Guang Wu. f-vaes: Improve vaes with conditional flows. *arXiv preprint arXiv:1809.05861*, 2018.
25. Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.
26. Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
27. Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4091–4099. JMLR. org, 2017.
28. Philip Bachman. An architecture for deep, hierarchical generative models. In *Advances in Neural Information Processing Systems*, pages 4826–4834, 2016.

 doi:10.20944/preprints202409.1969.v

29. Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al. Hierarchical generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1810.07217*, 2018.
30. Hang Yin, Francisco S Melo, Aude Billard, and Ana Paiva. Associate latent encodings in learning from demonstrations. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
31. 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.
32. 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.
33. 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.
34. 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).
35. 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.
36. 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.
37. 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)*.
38. 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.
39. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
40. 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).
41. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
42. 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/.
43. 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.
44. 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.
45. 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.
46. 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. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.

47. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
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)*,

 doi:10.20944/preprints202409.1969.v

pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 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. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024.
78. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
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. 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.
90. 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.
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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
92. 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.
93. 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.