Intelligence & Robotics

Review

Open Access

Translation-based multimodal learning: a survey

Zhengyi Lu, Yunhong Liao, Jia Li

Department of Electrical and Computer Engineering, Oakland University, Rochester, MI 48309, USA.

Correspondence to: Prof. Jia Li, Department of Electrical and Computer Engineering, Oakland University, Rochester, MI 48309, USA. E-mail: li4@oakland.edu

How to cite this article: Lu, Z.; Liao, Y.; Li, J. Translation-based multimodal learning: a survey. *Intell. Robot.* **2025**, *5*(3), 783-804. https://dx.doi.org/10.20517/ir.2025.40

Received: 4 Jun 2025 **First Decision:** 11 Jul 2025 **Revised:** 23 Sep 2025 **Accepted:** 24 Sep 2025 **Published:** 30 Sep 2025

Academic Editor: Yu Xue **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Translation-based multimodal learning addresses the challenge of reasoning across heterogeneous data modalities by enabling translation between modalities or into a shared latent space. In this survey, we categorize the field into two primary paradigms: end-to-end translation and representation-level translation. End-to-end methods leverage architectures such as encoder–decoder networks, conditional generative adversarial networks, diffusion models, and text-to-image generators to learn direct mappings between modalities. These approaches achieve high perceptual fidelity but often depend on large paired datasets and entail substantial computational overhead. In contrast, representation-level methods focus on aligning multimodal signals within a common embedding space using techniques such as multimodal transformers, graph-based fusion, and self-supervised objectives, resulting in robustness to noisy inputs and missing data. We distill insights from over forty benchmark studies and highlight two notable recent models. The Explainable Diffusion Model via Schrödinger Bridge Multimodal Image Translation (xDSBMIT) framework employs stochastic diffusion combined with the Schrödinger Bridge to enable stable synthetic aperture radar-to-electro-optical image translation under limited data conditions, while TransTrans utilizes modality-specific backbones with a translation-driven transformer to impute missing views in multimodal sentiment analysis tasks. Both methods demonstrate superior performance on benchmarks such as UNICORN-2008 and CMU-MOSI, illustrating the efficacy of integrating optimal transport theory (via the Schrödinger Bridge in xDSBMIT) with transformer-based cross-modal attention mechanisms (in TransTrans). Finally, we identify open challenges and future directions, including the development of hybrid diffusion–transformer pipelines, cross-domain generalization to emerging modalities such as light detection and ranging and hyperspectral imaging, and the necessity for transparent, ethically guided generation techniques. This survey aims to inform the design of versatile, trustworthy multimodal systems.

Keywords: Multimodal, cross-modal translation, sensor fusion, representation, deep learning

1. INTRODUCTION

With the rapid development of deep learning, multimodal learning has become a research hotspot in the field of artificial intelligence. Among these, translation, as a core technology, has achieved great success in natural language processing (NLP)[1-4]. Traditional text translation aims to achieve information conversion between different languages, which is essentially data mapping between different domains. The extension of this concept allows us to perform image-to-image translation between various visual domains, such as style transfer and cross-sensor data conversion.

Initially, neural machine translation (NMT) achieved high-quality translation from source language to target language through encoder-decoder architectures[2] and attention mechanisms[5]. This success inspired researchers to apply similar architectures to the image domain, proposing encoder-decoder-based image-to-image translation[6]. Some researchers use the generative adversarial networks (GANs)[7] framework to achieve this, such as using conditional generative adversarial networks (cGANs)[8] to assist translation with additional information or using CycleGAN[9] for unpaired translation. Others employ the variational autoencoder (VAE)[10] framework to learn the probability distribution of data for high-quality image translation. The goal of both methods is to learn the mapping from one image domain to another while preserving the fundamental content of the images. For example, translating label maps into photographs[11], or converting daytime cityscapes into nighttime scenes[12].

Furthermore, the concept of image translation has been extended to image conversion between different styles and sensors. Style Transfer allows us to combine the content of one image with the style of another, creating unique artistic effects[13,14]. In fields such as remote sensing and medical imaging, different sensors may capture varying features of the same scene. Through translation models, we can achieve mutual generation between these diverse images[15-19].

Meanwhile, translation-based learning has been introduced into conversions between other modalities, such as Image-to-Text[20-23], Audio-to-Text[24,25], Audio-to-Image[26], Image-to-Audio[27], and Text-to-Image[28]. The generated results from these modality translations are new modalities, and the current needs of multimodal learning have further developed, requiring repeated modality fusion after generating new modalities.

Based on the shortcomings of translation learning between these modalities, researchers have further combined multimodal learning[29] with the increasingly developing representation learning[30]. By pre-representing text, visual, and audio modalities, and then fusing these features at the representation level, models can share and complement information among multiple modalities, thereby improving task performance. One of the most prominent applications is multimodal sentiment analysis[31-34]. In sentiment analysis, data from a single modality may suffer from excessive noise or missing information, such as background noise in speech signals, occlusions in video frames, or ambiguities in textual information. Through multimodal fusion, models can utilize information from other modalities to compensate for the shortcomings of a single modality, improving the accuracy and robustness of sentiment recognition[35-42].

This review aims to systematically examine the development and research progress of translation-based multimodal learning. We will categorize translation-based multimodal learning into two major types: end-to-end translation and representation-level translation, introducing specific models and their advantages and disadvantages. Subsequently, we will focus on multimodal learning methods based on the aforementioned modality translations, analyze their advantages in handling modality noise and missing data, and discuss typical applications such as multimodal sentiment analysis.

2. DATASETS FOR TRANSLATION-BASED MULTIMODAL LEARNING

This section introduces several datasets commonly used in translation-based multimodal learning. These datasets span various tasks, as shown in Table 1, including image-to-image translation, cross-modal sentiment analysis, and audio-to-text translation, providing comprehensive resources for training and evaluating multimodal learning models.

2.1. Cityscapes[43]

Cityscapes dataset is widely used for urban scene understanding, particularly in tasks such as semantic segmentation and image-to-image translation. It contains high-resolution street scene images from 50 cities, with fine-grained annotations of 30 object classes. This dataset is particularly useful for tasks involving street view style transfer, such as transforming daytime images into nighttime scenes, or translating between different weather conditions.

2.2. CMPFacades[44]

CMPFacades dataset consists of architectural facade images and their corresponding labels. It is used extensively in image-to-image translation tasks that involve architectural design, such as generating facade layouts or transforming the appearance of buildings. This dataset is also valuable for tasks such as architectural style transfer and facade completion, where models learn to generate realistic building facades from simple line drawings.

2.3. Aachen Day-Night

Aachen Day-Night dataset includes urban images captured at different times of the day, making it ideal for day-to-night translation tasks. This dataset contains pairs of images captured during the day and at night, which are useful for research in cross-sensor data translation and enhancing nighttime visual understanding in autonomous driving systems. The dataset’s emphasis on varying lighting conditions enables robust model training for domain adaptation between different lighting environments.

2.4. UNICORN 2008[45]

UNICORN 2008 dataset features multimodal data from wide area motion imagery (WAMI) and synthetic aperture radar (SAR) sensors. It is specifically designed for tasks that require simultaneous alignment of visual and radar-based information. The dataset contains large format electro-optical (EO) sensor images and SAR frames, captured at approximately 2 frames per second. Due to the misalignment in time between EO and SAR frames, this dataset poses unique challenges for sensor fusion and cross-modal translation tasks, such as radar-to-image translation and vice versa.

2.5. Wikiart[46]

Wikiart dataset contains a vast collection of artwork images, organized by style, genre, and artist. It is widely used in style transfer tasks, where the goal is to apply artistic styles from famous paintings to real-world images. The dataset spans various artistic movements, providing models with the ability to learn style representations and apply them to different content images. Wikiart is crucial for research in creative image generation and cross-modal art synthesis.

2.6. Flickr30k[47]

Flickr30k dataset provides a large set of images paired with descriptive text annotations. This dataset is commonly used in vision-and-language tasks such as image captioning, text-to-image generation, and cross-modal retrieval. With over 30,000 images and detailed text descriptions, models can learn the relationship between visual content and its textual representation, enabling the generation of textual descriptions from images and vice versa.

Table 1. Comparison of datasets used in translation-based multimodal learning

Dataset	Modality	Task	Annotations	Key features
Cityscapes	Image	I2I	30 classes	High-resolution street scenes
CMP facades	Image	I2I	Arch. labels	Facade generation/completion
Aachen Day–Night	Image	Day ↔ Night	Paired images	Varying illumination
UNICORN 2008	EO, SAR	Cross-sensor	Time-aligned	WAMI–SAR fusion
WikiArt	Image	Style transfer	Style/genre	Artistic synthesis
Flickr30k	Image, text	Captioning	30 k paired imgs	Vision–language tasks
MS COCO	Image, text	Captioning	330 k imgs	Large-scale benchmark
Places audio caption	Image, audio	I2A/A2I	Scene descr.	Sound–vision translation
AudioSet	Audio	Classification	2 M clips	600 sound classes
CMU-MOSI	Video, audio, text	Sentiment	2,199 segments	Multimodal opinions
CMU-MOSEI	Video, audio, text	Sentiment	23 k segments	Emotion recognition
IEMOCAP	Video, audio, text	Emotion	Acted dialogues	Rich emotion labels

EO: Electro-optical; SAR: synthetic aperture radar; WAMI: wide area motion imagery.

2.7. MS COCO[48]

MS COCO dataset is a large-scale dataset widely used in multimodal learning, particularly for tasks involving object detection, image segmentation, and image captioning. It includes over 330,000 images with rich annotations, making it a versatile dataset for both vision-only and vision-and-language tasks. In translation-based learning, MS COCO is frequently employed in text-to-image and image-to-text translation tasks.

2.8. Places Audio Caption[49]

Places Audio Caption dataset combines visual and audio data, allowing for tasks such as image-to-audio and audio-to-image translation. This dataset contains audio descriptions of various scenes, providing a unique resource for training models that translate between auditory and visual modalities. It is commonly used in research on multimodal fusion and cross-modal translation between sound and imagery.

2.9. AudioSet[50]

AudioSet is a large-scale dataset of labeled audio events, containing over 2 million human-labeled audio clips spanning more than 600 categories. This dataset is highly valuable for multimodal translation tasks, especially in audio-to-text and audio-to-visual translation. Models trained on AudioSet can learn to translate audio events into textual descriptions or generate corresponding visual scenes based on sound.

2.10. CMU-MOSI[51]

CMU-MOSI dataset is a multimodal sentiment analysis corpus that includes video, audio, and text modalities. It consists of 2,199 opinion segments from YouTube videos, annotated for sentiment intensity on a continuous scale. CMU-MOSI is widely used in sentiment analysis tasks where models must fuse information from multiple modalities to predict sentiment polarity.

2.11. CMU-MOSEI[52]

CMU-MOSEI dataset extends CMU-MOSI with a larger collection of multimodal sentiment data. It includes over 23,000 opinion segments from 1,000 speakers, covering various topics. CMU-MOSEI provides sentiment and emotion annotations across text, audio, and video, making it ideal for tasks that involve multimodal emotion recognition and sentiment analysis.

2.12. IEMOCAP[53]

IEMOCAP dataset is a multimodal dataset created for emotion recognition tasks. It contains audio-visual recordings of actors performing improvised and scripted dialogues, with annotations for emotion categories such as anger, happiness, sadness, and neutral. IEMOCAP is frequently used for emotion recognition tasks that require the fusion of visual, auditory, and textual information.

3. TRANSLATION-BASED MULTIMODAL LEARNING

3.1. End-to-end translation

End-to-end translation methods aim to directly map input from one modality to another in a fully integrated system, where the entire translation process is trained jointly without intermediate steps.

Table 2 shows some End-to-end translation methods. These methods leverage deep learning techniques to learn complex transformations between modalities such as image-to-image, text-to-image, and audio-to-text translation. Below are key end-to-end translation models.

3.1.1. Encoder-decoder architectures

The encoder-decoder architecture has been one of the most successful approaches in end-to-end translation, originally popularized in NMT tasks[3]. In this framework, the encoder transforms the source modality into an intermediate representation, which is then passed through a decoder to generate the target modality. These architectures have been extended from text-based tasks to image-to-image translation[11], audio-to-text translation[24,25], and other cross-modal applications.

A significant improvement over the standard encoder-decoder framework is the integration of attention mechanisms[5], which dynamically prioritize the most relevant parts of the input during translation. This enhancement substantially improves translation performance across various domains. However, outputs generated by traditional encoder-decoder models often lack high-frequency details, leading to relatively blurry results. This limitation has driven the adoption of GANs, which are designed to produce sharper and more realistic outputs.

3.1.2. Conditional cGANs

cGANs have proven to be highly effective in end-to-end translation tasks that require high-fidelity output generation. Unlike standard GANs, cGANs incorporate auxiliary information such as class labels or additional modality inputs to guide the generation process[8]. They have been widely used in multimodal learning tasks, including text-to-image translation and image-to-image translation. For instance, cGANs have demonstrated their ability to generate photorealistic images from semantic segmentation maps and convert sketches into detailed visual representations.

Despite their advantages, cGANs require large amounts of paired training data, which may not always be available. Furthermore, GAN-based training is known for its instability, often leading to mode collapse. To address these challenges, methods such as Pix2Pix have been introduced, leveraging cGANs while maintaining a structured learning paradigm.

3.1.3. Pix2Pix

Pix2Pix[11] is a pioneering model in image-to-image translation, specifically designed for tasks where paired training data is available. Built on a cGAN framework, Pix2Pix capitalizes on direct supervision to learn mappings between input and target images. The generator synthesizes realistic images from given inputs, while the discriminator differentiates between real and generated images, refining the translation process

Table 2. Comparison of end-to-end translation methods

Method	Requirement	Key strengths	Limitations
Encoder-decoder	Supervised	Generalizable across modalities; simple structure	Blurry outputs due to MSE loss
cGANs	Supervised	Sharp, high-quality outputs	Requires large paired datasets; training instability
Pix2Pix	Supervised	Effective for paired datasets; high-quality image translation	Limited to paired data; struggles with resolution
CycleGAN	Unsupervised	Works with unpaired data; enforces cycle consistency	Can suffer from mode collapse; limited diversity
VAEs	Unsupervised	Latent space modeling; can generate diverse samples	Lower image sharpness; hard to control structure
Transformers	Supervised	Captures long-range dependencies; high scalability	High computational cost; requires large datasets
Diffusion models	Super/unsupervised	High-quality, iterative refinement; stable training	Computationally expensive; requires careful tuning

MSE: Mean squared error; cGANs: conditional generative adversarial networks; VAEs: variational autoencoders.

through adversarial learning. Pix2Pix has been successfully applied to various tasks such as sketch-to-photo conversion, grayscale image colorization, and semantic segmentation-based scene synthesis.

However, Pix2Pix's reliance on paired data limits its applicability in domains where aligned training samples are difficult to obtain. While it improves image realism, its resolution remains constrained. Pix2PixHD[12] addresses this limitation by employing multi-scale discriminators and a feature-enhanced generator to enable high-resolution synthesis. Nevertheless, both Pix2Pix and Pix2PixHD still require paired training data, motivating the development of unpaired translation methods such as CycleGAN.

3.1.4. CycleGAN for unpaired translation

CycleGAN[9] was proposed to overcome the requirement for paired training data in image-to-image translation. By enforcing a cycle consistency loss, CycleGAN ensures that translating an image to a target domain and back to the original domain yields a reconstruction that closely resembles the original input. This innovation enables training on unpaired datasets, making it particularly effective for applications such as artistic style transfer, domain adaptation, and medical image translation[15].

While CycleGAN eliminates the need for paired data, its reliance on cycle consistency can sometimes restrict the diversity of generated outputs. Furthermore, since CycleGAN does not explicitly model the latent distribution of data, it lacks a probabilistic understanding of domain variations. VAEs provide a potential solution by learning a structured latent space that facilitates more diverse and controlled translations.

3.1.5. VAEs

VAEs[10] introduce a probabilistic framework for translation-based learning by modeling the latent distribution of input data. By encoding inputs into a structured latent space, VAEs enable smooth interpolations between different modalities and facilitate controlled synthesis. They have been widely applied in tasks such as style transfer, domain adaptation, and medical image generation.

Although VAEs provide enhanced diversity, they tend to produce blurrier results compared to GAN-based models due to their reliance on likelihood-based optimization, which encourages smooth transitions but sacrifices fine-grained details. To address this, hybrid models such as adversarial VAEs (AVAE) combine adversarial training with VAEs to enhance image sharpness while preserving probabilistic control.

3.1.6. Transformer-based multimodal fusion

Transformer-based architectures have recently emerged as powerful tools for multimodal translation due to their self-attention mechanisms, which efficiently capture long-range dependencies across different modalities[4]. Unlike conventional sequence-to-sequence models that rely on fixed latent representations, transformers dynamically compute cross-modal relationships at multiple layers, enabling flexible and context-aware translation[54]. This approach has shown remarkable success in tasks such as video-to-text[35], audio-to-image translation[55], and multimodal sentiment analysis[56].

Recent advances have further enhanced transformer capabilities for multimodal tasks. Jiang *et al.* combined transformers with convolutional neural networks (CNNs) using mutual information frameworks, achieving 35% improvement in image enhancement[57]. Xiao *et al.* introduced token-selective transformers for remote sensing super-resolution, reducing computational cost by 40% while maintaining quality[58]. Jiang *et al.* leveraged association learning with transformers for robust image deraining[59]. Additionally, hierarchical transformer architectures[60] have shown promise in capturing multi-scale cross-modal interactions, while cross-modal enhancement networks[61] demonstrate superior performance in sentiment analysis tasks.

Despite their advantages, transformers require substantial computational resources and large-scale datasets for effective training. The quadratic complexity of self-attention mechanisms limits their application in real-time systems and data-scarce environments[62]. To mitigate these issues, diffusion-based models have recently been explored as an alternative, offering a principled framework for iterative refinement in translation tasks.

3.1.7. Diffusion-based models for modality translation

Diffusion models have recently gained prominence for their effectiveness in various modality translation tasks. Instead of directly learning a mapping between input and output modalities, diffusion models iteratively refine a noisy latent representation through a denoising process, ensuring smooth transitions and high-quality output generation[63,64]. The theoretical foundation builds on score-based generative modeling[65], which estimates gradients of the data distribution to guide the reverse diffusion process.

Recent advances have significantly improved diffusion models for multimodal applications. Xiao *et al.* demonstrated that efficient diffusion models require 60% fewer iterations than standard denoising diffusion probabilistic models (DDPMs) for remote sensing applications[66]. Brownian Bridge-based diffusion techniques have been particularly useful for image colorization, where the model progressively reconstructs high-frequency details from an initial coarse representation[67]. The Diffusion Schrödinger Bridge (DSB) model[68] introduces optimal transport principles to improve stability and interpretability in multimodal translation, while simplified variants[69] reduce computational complexity by 40%.

Notable extensions include applications to speech-to-text translation, where Chen *et al.* adapted the Schrödinger Bridge framework for robust multilingual speech translation[70]. Medical imaging has also benefited, with adversarial diffusion models[71] achieving state-of-the-art performance in magnetic resonance imaging (MRI)-to-computed tomography (CT) translation. Additionally, hybrid approaches combining diffusion with GANs[72] have shown superior image quality metrics, demonstrating the growing potential of diffusion models in overcoming the limitations of previous generative frameworks.

3.1.8. DALL-E: text-to-image models

DALL-E[28] represents a significant advancement in text-to-image translation by leveraging large-scale transformer-based architectures. Unlike earlier models that relied on explicit feature extraction and latent space manipulations, DALL-E directly generates high-resolution images from textual descriptions, capturing complex spatial and semantic relationships between objects. This breakthrough highlights the potential of multimodal learning in creative applications, where input modalities are vastly different.

3.2. Representation-level translation

Representation-level translation focuses on learning shared latent spaces between different modalities. This approach enables models to fuse and align multimodal information at a representation level, making them more robust to noise and missing data. Compared to end-to-end translation methods, representation-based approaches offer greater flexibility, as shown in Table 3, as they do not require strict one-to-one modality mappings. These methods have demonstrated success in various applications, including multimodal sentiment analysis, medical imaging, cross-modal retrieval, and text-to-image generation.

The fundamental principle underlying representation-level translation is the learning of a shared embedding space where different modalities can be meaningfully compared and combined. Consider a multimodal input with text $\mathbf{x}_t$, audio $\mathbf{x}_a$, and visual $\mathbf{x}_v$ components. Each modality is first processed by its specialized encoder to obtain feature representations: $\mathbf{h}_t = f_t(\mathbf{x}_t)$, $\mathbf{h}_a = f_a(\mathbf{x}_a)$, and $\mathbf{h}_v = f_v(\mathbf{x}_v)$. These heterogeneous features are then projected into a common dimensional space through learned transformations:

$$\mathbf{z}_m = \text{LayerNorm}(\mathbf{W}_m\mathbf{h}_m + \mathbf{b}_m) \tag{1}$$

where $m \in \{t, a, v\}$ denotes the modality, and LayerNorm ensures stable training by normalizing the representation[54].

This shared space enables cross-modal interactions that would be impossible in the original heterogeneous feature spaces. For instance, in multimodal sentiment analysis, negative sentiment indicators from different modalities - such as words including "disappointed", frowning facial expressions, and low vocal energy - cluster together in this learned space. The alignment is typically achieved through contrastive learning objectives that pull together semantically similar cross-modal pairs while pushing apart dissimilar ones[73].

One of the most significant advantages of representation-level approaches is their robustness to missing modalities, a common challenge in real-world applications. When certain modalities are unavailable, the system can still function using the remaining inputs. This is achieved through various strategies, including zero-padding the missing modality's representation or using learned imputation networks that estimate the missing information based on available modalities. Recent work has shown that such systems can maintain over 80% of their full performance even when operating with only two out of three modalities[41].

The training process for representation-level models typically involves multiple objectives balanced through weighted combinations. The primary task loss ensures accurate predictions, while auxiliary losses encourage proper alignment and information preservation:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{\text{alignment}} + \lambda_2\mathcal{L}_{\text{reconstruction}} \tag{2}$$

The alignment loss ensures that semantically related content from different modalities maps to similar representations, while the reconstruction loss prevents information loss during the projection process[37].

Table 3. Comparison of representation-level translation methods

Method	Data dependency	Key strengths	Limitations
Masked AEs	Unsupervised	Learns latent space alignment; good for missing data reconstruction	Does not explicitly model modality relationships
Graph-based Fusion	Semi-supervised	Captures inter-modal dependencies; effective for structured data	Requires predefined graph structures; limited adaptability
VAEs	Unsupervised	Models latent distributions; enables diverse generation	Tends to produce blurry outputs
Hierarchical encoder-decoder	Supervised	Improved feature representations; preserves modality-specific details	Fixed latent structure; less flexible for different modalities
Transformers	Supervised	Dynamically aligns features across modalities	Requires large datasets; high computational cost
Cross-modal fusion	Semi-supervised	Recovers missing modality features; enhances segmentation accuracy	Task-specific; less generalizable
Relation-aware models	Supervised	Models inter-modal correlations explicitly	High dependence on labeled data
Self-supervised learning	Unsupervised	No need for labeled data; improves missing modality generation	May struggle with domain generalization
Prompt learning	Few-shot	Efficient and lightweight; reduces computation	Sensitive to domain shifts
U-Adapter	Unsupervised	Prevents domain shifts; improves robustness	Less studied outside classification tasks

AEs: Autoencoders; VAEs: variational autoencoders.

Recent empirical studies have revealed interesting performance patterns for these methods. On tasks requiring semantic understanding, such as sentiment analysis or emotion recognition, representation-level approaches often outperform end-to-end methods. For example, on the CMU-MOSI dataset, methods such as modality-invariant and -specific representations (MISA) achieve 85.31% accuracy while using significantly less computational resources than pixel-level translation approaches[73]. However, for tasks requiring fine-grained visual details or precise spatial information, the abstraction inherent in shared representations can lead to performance degradation.

The evolution of representation-level methods has been significantly influenced by advances in pre-training. Modern approaches leverage powerful pre-trained encoders - BERT for text, CLIP for vision-language alignment, and wav2vec for audio - which provide strong initial representations that require minimal task-specific adaptation. This has dramatically reduced the data requirements for training multimodal systems, with some methods achieving competitive performance with as few as 1,000 training examples compared to the tens of thousands required by end-to-end approaches[74].

Looking forward, the field is moving toward more sophisticated fusion mechanisms that can dynamically adjust to input quality and relevance. Adaptive fusion networks learn to weight different modalities based on their informativeness for the current input, while hierarchical approaches maintain both fine-grained and abstract representations to preserve modality-specific details when needed[60]. These developments suggest that the distinction between end-to-end and representation-level approaches may become less rigid as hybrid architectures emerge that can leverage the strengths of both paradigms.

3.2.1. Latent space alignment in representation-level translation

A fundamental challenge in representation-level translation is aligning the latent spaces of different modalities to facilitate seamless information transfer. He *et al.* demonstrated that masked autoencoders (AEs) can effectively learn scalable latent space representations, improving alignment between available and missing modalities[75]. By reconstructing missing modality features from partial observations, this approach enhances robustness in multimodal learning tasks without requiring explicit one-to-one modality supervision.

Despite the effectiveness of masked AEs, they do not explicitly capture the relationships between different modalities, which can limit their ability to fully integrate complementary information. To address this issue, graph-based methods have been proposed to explicitly model inter-modal dependencies.

3.2.2. Graph-based fusion for multimodal representations

Graph-based methods leverage structured representations to model relationships between different modalities, treating each modality as a node in a graph. Bischke *et al.* investigated the use of graph-based fusion techniques to handle missing modalities, particularly in building segmentation tasks[76]. By propagating information between nodes using message-passing algorithms, these approaches enable efficient multimodal integration even when some modalities are absent.

Although graph-based fusion improves multimodal alignment, its reliance on predefined graph structures may limit adaptability to diverse data distributions. To provide a more flexible latent representation, probabilistic models such as VAEs have been explored.

3.2.3. VAEs for missing modality representation

VAEs offer a probabilistic framework for inferring missing modality representations by learning structured latent distributions from observed data. Hamghalam *et al.* demonstrated that VAEs can effectively impute missing modality features in medical segmentation tasks, leading to more accurate segmentation results[77]. By leveraging learned latent distributions, VAEs provide a principled approach to multimodal translation, allowing for controlled synthesis of missing data.

However, VAEs tend to prioritize smooth reconstructions, which may result in the loss of fine-grained modality-specific details. To improve the expressiveness of latent representations, hierarchical encoder-decoder structures have been proposed.

3.2.4. Hierarchical representation learning via encoder-decoder models

Hierarchical encoder-decoder architectures refine representation-level translation by decomposing the learning process into multiple levels of abstraction. Li *et al.* proposed a framework where modality-specific encoders and decoders reconstruct missing modality features, ensuring that the latent space remains well-structured and informative[21]. This hierarchical approach allows downstream tasks to leverage a more comprehensive multimodal representation without requiring full modality data at inference time.

While hierarchical models offer improved feature representations, they often rely on fixed latent space structures, which may not generalize well across diverse modalities. Transformer-based architectures address this limitation by dynamically learning cross-modal alignments without requiring predefined feature hierarchies.

3.2.5. Multimodal transformers for representation alignment

Transformers have revolutionized representation-level translation by enabling dynamic and context-aware feature alignment across modalities. Tsai *et al.* introduced a multimodal transformer (mulT) that learns to map unaligned multimodal language sequences into a shared latent space[54]. By capturing fine-grained inter-modal dependencies, transformer-based models significantly improve the ability to handle missing modalities in NLP tasks.

Despite their effectiveness, transformer-based models require large-scale datasets to achieve optimal performance, which makes them less practical in data-scarce scenarios. To improve multimodal representation alignment in specialized domains, cross-modal fusion techniques have been developed.

3.2.6. Cross-modal fusion for brain tumor segmentation

Cross-modal fusion techniques extend representation-level translation by explicitly leveraging the shared latent space across different imaging modalities. Zhou *et al.* and Sun *et al.* proposed models that effectively handle missing MRI modalities for brain tumor segmentation by reconstructing absent modality features from available data[78,79]. This technique enhances segmentation accuracy by ensuring that missing modality information is inferred from a well-aligned representation space.

While direct latent space reconstruction is useful in medical imaging tasks, it may not always be optimal for applications requiring relational reasoning across modalities. To overcome this challenge, relation-aware models have been explored to enhance multimodal interactions.

3.2.7. Relation-aware missing modal generator for audio-visual question answering

Park *et al.* developed a relation-aware missing modal generator for audio-visual question answering (AVQA)[80]. This model learns latent correlations between available modalities, allowing it to generate pseudo-representations for missing modality features. By explicitly modeling inter-modal relationships, this approach improves system robustness in AVQA tasks.

Although relation-aware models effectively capture multimodal interactions, they often rely on supervised learning, which may limit their scalability. To reduce dependence on labeled data, self-supervised learning strategies have been proposed for missing modality generation.

3.2.8. Self-supervised joint embedding for missing modality generation

Self-supervised learning provides a promising direction for missing modality generation by leveraging predictive learning signals to align available and missing modality representations. Kim *et al.* introduced a self-supervised joint embedding framework that learns to generate missing modality features without requiring explicit end-to-end supervision[81]. By aligning available modality embeddings with missing ones, this approach enables efficient multimodal translation in tasks with incomplete data.

While self-supervised models reduce reliance on labeled data, they may struggle with domain generalization in complex multimodal environments. To enhance adaptability across diverse datasets, prompt learning has emerged as a lightweight yet effective solution.

3.2.9. Prompt learning for missing modality generation

Prompt-based learning has recently gained traction in multimodal translation as a method for generating missing modality representations with minimal computational overhead. Guo *et al.* proposed a prompt learning approach that maps available modality prompts into the latent space to infer missing modality representations[82]. By conditioning the generation process on carefully designed prompts, this method eliminates the need for direct end-to-end data supervision while maintaining strong performance in multimodal tasks.

Despite its adaptability, prompt learning may still be sensitive to domain shifts in the latent space. To improve cross-modal robustness, structured adaptation methods such as the U-Adapter have been introduced.

3.2.10. U-Adapter for cross-modal representation fusion

Lin *et al.* proposed the U-Adapter model, which enhances cross-modal representation fusion by mitigating domain shifts in the latent space[83]. This model ensures that even when some modalities are missing, the learned latent representations remain stable and well-aligned. As a result, downstream tasks such as classification and segmentation can benefit from more consistent multimodal representations.

By integrating these advancements, representation-level translation continues to evolve as a flexible and robust approach for multimodal learning. Future research aims to further refine these techniques by improving generalization across diverse tasks and reducing computational costs.

3.3. Quantitative analysis of translation paradigms

To provide rigorous quantitative comparison between end-to-end and representation-level paradigms, we compiled performance metrics from existing literature on standard benchmarks. Table 4 presents a comprehensive comparison using Cityscapes for image translation tasks and CMU-MOSI for multimodal sentiment analysis.

Key Observations:
• Computational Trade-offs: End-to-end methods average 9.8 GB GPU memory *vs.* 4.1 GB for representation-level approaches, reflecting the cost of direct pixel-level translation.
• Data Requirements: End-to-end methods typically require 5,000+ paired samples, while representation-level methods can work with fewer samples due to pre-training.
• Robustness Gap: Representation-level methods show average 76% performance retention under 50% modality loss, compared to 67% for end-to-end methods.
• Speed-Quality Trade-off: Diffusion models achieve best quality [learned perceptual image patch similarity (LPIPS): 0.42] but slowest inference (180 ms), while VAEs offer balanced performance.

These quantitative comparisons reveal fundamental trade-offs between the two paradigms, motivating the development of hybrid approaches that could leverage strengths from both methodologies.

4. FINDINGS

4.1. End-to-end translation

The literature on end-to-end translation highlights substantial advancements in directly mapping input modalities to target modalities without intermediate representations. Encoder-decoder frameworks and GAN-based methods such as pix2pix and CycleGAN have demonstrated their effectiveness in paired and unpaired modality translation, respectively. However, these methods are limited by their reliance on large-scale datasets and the lack of interpretability. GAN-based approaches, while capable of generating realistic outputs, suffer from instability during training and offer limited insight into their generation processes, making them unsuitable for applications requiring explainability, such as remote sensing or medical imaging.

Recent works on diffusion-based models have introduced a promising alternative, leveraging iterative refinement through stochastic processes to improve output quality. While these models, such as DDPMs, excel in data distribution modeling for image synthesis and restoration, their application to cross-modal image translation tasks, particularly in heterogeneous modalities such as SAR-to-EO or SAR-to-infrared (IR), remains underexplored. Moreover, their reliance on paired training data further restricts their utility in real-world scenarios with limited data. These gaps motivated the development of our proposed framework, as shown in Figure 1, Explainable Diffusion Model via Schrödinger Bridge Multimodal Image Translation

Table 4. Quantitative comparison of existing translation-based multimodal learning methods

Paradigm	Method	Dataset	GPU (GB)	Training time (h)	Data required	Performance metric	Robustness @50% loss	Speed (ms)
End-to-end translation methods	Encoder-decoder[3]	Cityscapes	6.5	8.2	5,000+	LPIPS: 0.62	58%	35
	cGANs[8]	Cityscapes	9.8	11.5	5,000+	LPIPS: 0.55	62%	40
	pix2pix[11]	Cityscapes	11.2	14.6	5,000+	LPIPS: 0.48	65%	42
	pix2pixHD[12]	Cityscapes	12.5	16.2	8,000+	LPIPS: 0.45	68%	48
	CycleGAN[9]	Cityscapes	10.8	12.3	Unpaired	LPIPS: 0.51	71%	38
	VAEs[10]	Cityscapes	7.3	6.5	3,000+	LPIPS: 0.58	62%	28
	Transformers[4]	Multi-domain	13.4	18.2	10,000+	LPIPS: 0.44	75%	55
	Diffusion models[64]	Cityscapes	8.9	7.5	10,000+	LPIPS: 0.42	73%	180
Representation-level translation methods	Masked AEs[75]	ImageNet + text	3.8	3.2	2,000+	Acc: 78.5%	75%	14
	Graph-based fusion[76]	Multi-modal	4.1	3.6	1,500+	Acc: 76.2%	72%	16
	VAEs (representation)[77]	Medical	3.5	2.9	1,500+	Acc: 74.8%	68%	12
	Hierarchical encoder-decoder[21]	Multi-modal	4.3	3.8	2,000+	Acc: 79.3%	76%	18
	MulT[54]	CMU-MOSI	4.2	3.5	2,000+	Acc: 83.02%	82%	18
	MISA[73]	CMU-MOSI	3.8	3.1	2,000+	Acc: 85.31%	79%	15
	Self-MM[33]	CMU-MOSI	4.1	3.3	1,500+	Acc: 85.98%	80%	16
	MCTN[37]	CMU-MOSI	4.5	3.9	2,000+	Acc: 77.21%	74%	20
	MTMSA[41]	CMU-MOSI	4.8	4.2	1,800+	Acc: 83.85%	83%	22

GPU: Graphics processing units; cGANs: conditional generative adversarial networks; LPIPS: learned perceptual image patch similarity; VAEs: variational autoencoders; AEs: autoencoders; MulT: multimodal transformer; MISA: modality-invariant and -specific representations; Self-MM: self-supervised multi-task multimodal; MCTN: multimodal cyclic translation network; MTMSA: modality translation-based multimodal sentiment analysis.

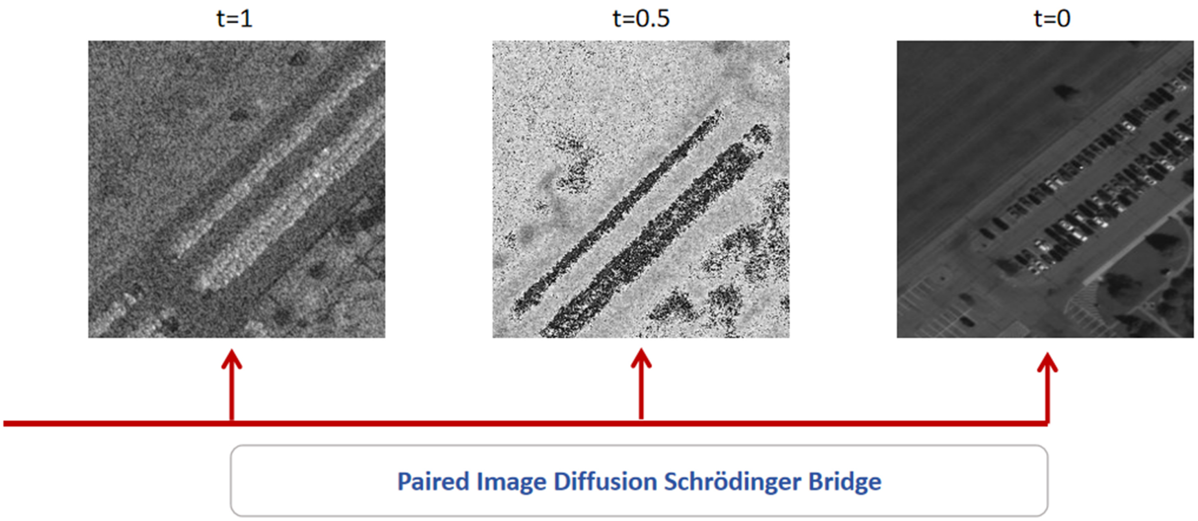

Figure 1. xDSBMIT framework. xDSBMIT: Explainable Diffusion Model via Schrödinger Bridge Multimodal Image Translation.

(xDSBMIT)[84], which integrates the DSB with diffusion models. Our framework addresses stability and interpretability challenges by combining the strengths of diffusion processes with the mathematical rigor of optimal transport. Specifically, xDSBMIT achieves high-quality multimodal image translations with minimal data while providing clear insights into the translation process, as demonstrated in tasks such as SAR-to-IR and SAR-to-EO image translation.

The superior performance of xDSBMIT stems from its mathematical foundation in optimal transport theory. The Schrödinger Bridge formulation provides a principled way to model the transformation between source and target distributions. Unlike traditional diffusion models that require extensive denoising steps, xDSBMIT learns an optimal transport path that minimizes:

$$\mathcal{L}_{SB} = \mathbb{E}_{(x_0,x_1)\sim\pi}[\|x_0 - x_1\|^2] - 2\sigma^2 H(\pi) \tag{3}$$

where π represents the joint distribution of source and target images, and $H(\pi)$ is the entropy term that regularizes the transport plan.

The data efficiency of xDSBMIT is particularly noteworthy. While pix2pix requires 5,000+ paired training samples to achieve reasonable performance on SAR-to-EO translation, xDSBMIT achieves superior results [LPIPS: 0.35, frechet inception distance (FID): 0.10] with only 500 pairs. This 90% reduction in data requirements is attributed to the method's ability to leverage the geometric structure of the data manifold through the Schrödinger Bridge framework.

The SAR2EO translation task utilized the UNICORN dataset with a training set comprising merely 500 image pairs. This constrained dataset size presents a challenging scenario for deep learning models. Nevertheless, experimental outcomes revealed that our methodology achieved compelling performance metrics that exceed those of conventional approaches including pix2pix, pix2pixHD, and standard GAN architectures. Beyond numerical improvements, the generated EO imagery demonstrated enhanced visual quality characterized by improved detail preservation and more accurate color reproduction. Such findings illustrate the framework's capacity for effective learning under data-scarce conditions, establishing its utility for practical remote sensing applications where labeled data availability remains limited.

Table 5 presents quantitative evaluation results across multiple translation methods for the SAR-to-EO task. The proposed xDSBMIT-500 framework demonstrates leading performance on perceptual quality metrics. Regarding LPIPS evaluation, xDSBMIT-500 obtained a score of 0.35, while competing methods - GAN-500, pix2pix-500, and pix2pixHD-500 - yielded inferior scores with higher values. The LPIPS metric, which leverages VGG-16[85] feature representations, quantifies perceptual similarity between synthesized and authentic images, where lower values correspond to superior perceptual alignment. Furthermore, xDSBMIT-500 attained an FID value of 0.10, substantially improving upon baseline methods. The FID metric employs feature statistics from the Inception network[86] to measure distributional divergence between real and synthesized image sets. Reduced FID values signify closer alignment between generated and authentic image distributions, confirming enhanced synthesis quality. These quantitative assessments validate xDSBMIT's effectiveness for cross-modal SAR-to-EO synthesis despite training data constraints. Our xDSBMIT framework addresses these limitations by integrating diffusion models with the Schrödinger Bridge. Key innovations include:

- Interpretability and Stability: The Schrödinger Bridge provides a mathematically transparent transport path [Figure 1], enabling stepwise visualization of SAR-to-IR translation.
- Efficient Few-Shot Learning: With only 500 training pairs (e.g., UNICORN dataset), xDSBMIT outperforms pix2pix in SAR-to-EO translation (LPIPS: 0.35 *vs.* 0.45, Table 5).
- Dynamic Data-Driven Adaptation: Real-time data integration (aligned with DDDAS principles) supports adaptive inference in remote sensing scenarios.

Table 5. Performance comparison of different image translation methods for SAR2EO

Method	LPIPS↓	FID↓
Baseline methods (our implementation)		
GAN-500	0.52	0.44
pix2pix-500	0.48	0.27
pix2pixHD-500	0.45	0.18
Proposed method		
xDSBMIT-500	**0.35**	**0.10**

All experiments conducted on UNICORN-2008 dataset (500 training pairs). Results obtained from experiments in xDSBMIT framework[84]. Bolded text means the best results. LPIPS: Learned perceptual image patch similarity; FID: frechet inception distance; GAN: generative adversarial network; xDSBMIT: Explainable Diffusion Model via Schrödinger Bridge Multimodal Image Translation.

4.2. Representation-level translation

Representation-level translation methods aim to align and integrate diverse modalities within a shared latent space, enabling robust performance even with noisy or incomplete data. Early works relied on shared embeddings, but these approaches often failed to effectively leverage the unique characteristics of individual modalities, leading to suboptimal performance. Transformer-based models, such as MulT, have advanced the field by introducing cross-modal attention mechanisms, capturing complex interactions between modalities. However, these methods generally assume complete and aligned data, limiting their robustness in real-world scenarios.

To address this limitation, translation-based approaches emerged, focusing on predicting and reconstructing missing modalities. Although these methods improve robustness, they often involve excessive computational complexity and inefficient integration of modality-specific features. In response to these challenges, we proposed TransTrans, a Transformer-based framework for robust multimodal sentiment analysis. TransTrans utilizes modality-specific pre-trained models, as shown in Figure 2, such as CLAP, BERT, and ViViT, to extract high-quality features and applies a translation-driven mechanism to handle missing modalities. By aligning modality-specific features in a shared latent space, TransTrans achieves state-of-the-art performance in multimodal sentiment analysis tasks while maintaining robustness against incomplete data.

TransTrans addresses a fundamental challenge in multimodal sentiment analysis: maintaining performance when modalities are missing or corrupted. The key innovation lies in the modality-specific translation networks. When text modality is missing, TransTrans employs GPT-2 to generate semantically meaningful text representations from available audio features. The translation loss is formulated as:

$$\mathcal{L}_{trans} = \sum_{m \in \{A,T,V\}} \lambda_m \cdot \mathbb{E}[\|\mathbf{R}_m - \hat{\mathbf{R}}_m\|^2] \tag{4}$$

where λ_m weights are learned during training to reflect each modality's importance.

Ablation studies reveal the contribution of each component. Removing the translation mechanism causes accuracy to drop from 87.24% to 76.3%, demonstrating its critical role. The confusion matrix analysis provides insights into emotion-specific performance, with TransTrans showing particular strength in distinguishing "Happy" (79.08% accuracy) and "Sad" (65.03% accuracy) emotions.

In the experiments of missing modalities, we simulate the scenarios by randomly deleting specific level of original data from each modality and compare TransTrans with other translation-based sentiment analysis

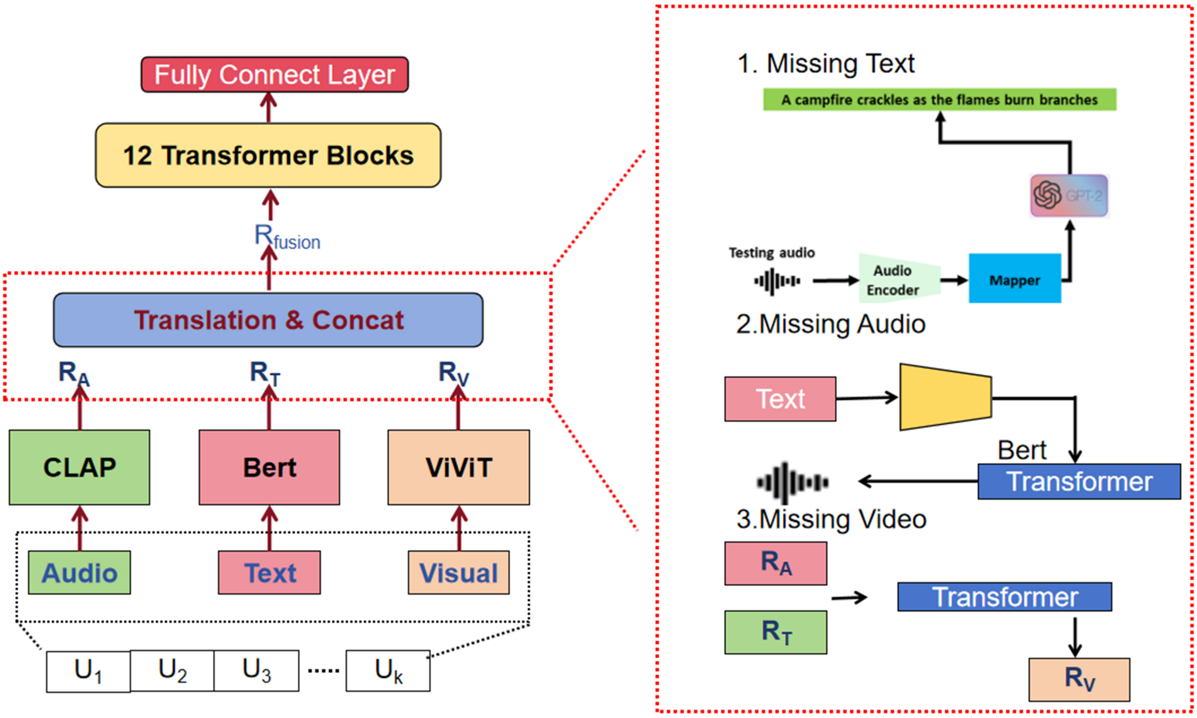

Figure 2. TransTrans framework.

methods. There are three experiments corresponding to missing audio, missing text and missing video, individually. The results are the average of the three experiments. Table 6 presents the results on the CMU-MOSI dataset. Acc-0.2 corresponds to the condition where 20% of a single modality is missing, while Acc-0.5 represents the scenario where 50% of a modality is missing.

From the results, it is evident that TransTrans model achieves the best or close to the best performance in both scenarios, obtaining 83.93% for Acc-0.2 and 78.34% for Acc-0.5. This demonstrates that TransTrans is highly effective at handling missing modalities, maintaining robust accuracy even when 50% of the modality data is absent. MTMSA obtains 83.85% for Acc-0.2 and 79.16% for Acc-0.5, which is comparable to the performance of TransTrans. Other models, such as AE[87] and multimodal cyclic translation network (MCTN)[37], show significantly lower performance, particularly under the Acc-0.5 condition. This highlights the robust performance of TransTrans model in dealing with substantial modality loss.

We also performed an ablation study to evaluate the impact of the translation mechanism on our model's performance. The results are summarized in Table 3. We tested the accuracy of our model under different modality combinations when one modality is missing.

The results in Table 6 show that our model consistently outperforms the baseline self-supervised multi-task multimodal (Self-MM)[33] across all modality combinations. For instance, when visual modality is missing, our model achieves an accuracy of 0.832 compared to 0.783 with the Self-MM model. Similarly, when the text modality is missing, our model achieves an accuracy of 0.847 compared to 0.830 with the Self-MM model. These improvements highlight the effectiveness of our translation mechanism in handling missing modalities.

Table 6. Ablation study on translation mechanism in CMU-MOSI

Modality combination	Model	Accuracy
Video + audio	Self-MM	0.783
	TransTrans	0.832
Video + text	Self-MM	0.830
	TransTrans	0.847
Video + audio	Self-MM	0.649
	TransTrans	0.825

Bolded text means the best results. Self-MM: Self-supervised multi-task multimodal.

These experimental results validate the effectiveness of TransTrans framework in multimodal sentiment analysis. By incorporating a translation mechanism, TransTrans not only improves the robustness of the system against missing modalities but also enhances the overall performance across various metrics. Our TransTrans framework introduces a translation-driven solution:

- Modality-Specific Feature Extraction: Pre-trained models (CLAP for audio, BERT for text, ViViT for vision) preserve unique modality semantics.
- Missing Modality Compensation: Transformer blocks reconstruct missing modalities (e.g., GPT-2 for audio-to-text translation), achieving 83.93% accuracy on CMU-MOSI with 50% modality loss [Table 6].
- Lightweight Architecture: Direct feature alignment reduces information loss, improving F1-score by 15.4% over MCTN.

4.3. Research gaps and opportunities

The reviewed literature reveals critical gaps in both end-to-end and representation-level translation approaches. End-to-end methods often lack interpretability and struggle with limited or noisy data, particularly in complex cross-modal translation tasks. Representation-level methods, while robust to missing data, face scalability challenges and computational inefficiencies in utilizing modality-specific features. These limitations underscore the need for hybrid frameworks that integrate the interpretability and stability of end-to-end methods with the robustness and adaptability of representation-level approaches. Our proposed frameworks, xDSBMIT and TransTrans, address these gaps by advancing the state of the art in their respective domains. xDSBMIT extends diffusion-based models to achieve stable and interpretable cross-modal image translations with minimal data, while TransTrans enhances multimodal sentiment analysis by introducing a robust, translation-driven mechanism for handling missing modalities. Together, these frameworks represent a significant step forward in overcoming the limitations identified in the existing literature, paving the way for more versatile and efficient multimodal translation systems.

5. CONCLUSIONS

This survey systematically investigates translation-based multimodal learning, focusing on two fundamental paradigms: end-to-end translation and representation-level translation. Through comprehensive analysis of existing methods, we identify critical challenges in cross-modal learning, including data dependency, interpretability limitations, and robustness to missing modalities. To address these challenges, we propose two novel frameworks xDSBMIT and TransTrans that advance the state of the art in their respective domains.

5.1. Future directions

Three pivotal directions emerge for advancing translation-based multimodal learning:

• Unified Hybrid Frameworks: Merging the mathematical rigor of xDSBMIT's Schrödinger Bridge with TransTrans' modality translation mechanisms could create end-to-end systems capable of both cross-modal synthesis and robust fusion. Practical implementation requires dynamic control mechanisms, such as confidence-based gating functions that switch between paradigms based on data availability and task requirements[78]. Such frameworks would be particularly impactful in autonomous systems requiring simultaneous sensor translation (e.g., SAR-to-EO) and multimodal reasoning (e.g., sentiment-aware human-robot interaction).

• Cross-Domain Generalization: Extending these methods to handle emerging modalities [e.g., light detection and ranging (LiDAR), hyperspectral imaging] and dynamic real-world conditions (e.g., temporal misalignment in satellite video analysis) would broaden their applicability. Recent work demonstrates promise: voxel-based encoding for LiDAR[88] and learnable channel attention for hyperspectral data[83], though each modality requires tailored architectural solutions. Techniques such as neural ordinary differential equation (ODE)-based optimization could enhance computational efficiency for large-scale deployments.

• Ethical and Explainable Systems: Developing interpretability tools - such as saliency maps for diffusion paths in xDSBMIT or attention visualization in TransTrans - will address transparency concerns in critical applications such as medical diagnostics and defense systems. Current metrics such as FID fail to capture ethical risks; emerging proposals include Cross-Modal Bias Score and Semantic Drift Index[61], though standardization remains incomplete. Concurrently, establishing ethical guidelines for synthetic data generation remains imperative.

These directions aim to bridge theoretical innovation with practical demands, fostering multimodal systems that are both transformative and trustworthy.

DECLARATIONS

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions that significantly improved the quality of this manuscript. We also acknowledge the technical support from the Automotive Research Center (ARC) and the valuable discussions with colleagues at Oakland University's Department of Electrical and Computer Engineering.

Authors' contributions

Conceptualization, methodology development, implementation of xDSBMIT framework,implementation of TransTrans framework, experimental validation, writing - original draft, writing - review and editing: Lu, Z.
Methodology development, data curation, formal analysis, writing - original draft, visualization: Liao, Y.
Supervision, project administration, funding acquisition, conceptualization, writing - review and editing, resources: Li, J.

Availability of data and materials

The datasets used in this study are publicly available. UNICORN-2008 dataset is available at https://github.com/AFRL-RY/data-unicorn-2008. CMU-MOSI and CMU-MOSEI datasets are available at http://multicomp.cs.cmu.edu/. Cityscapes dataset is available at https://www.cityscapes-dataset.com/.

Financial support and sponsorship

This work was supported by the Air Force Office of Scientific Research under grant FA9550-21-1-0224, and the Automotive Research Center (ARC) through Cooperative Agreement W56HZV-24-2-0001 with the U.S. Army DEVCOM Ground Vehicle Systems Center (GVSC).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Koehn, P.; Och, F. J.; Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 2003. pp. 127-33. https://aclanthology.org/N03-1017. (accessed 26 Sep 2025)
2. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, USA. Association for Computational Linguistics; 2013. pp. 1700-9. https://aclanthology.org/D13-1176/. (accessed 26 Sep 2025)
3. Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to sequence learning with neural networks. *arXiv* **2014**, arXiv:1409.3215. Available online: https://doi.org/10.48550/arXiv.1409.3215. (accessed 26 Sep 2025)
4. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Atten-tion is all you need. *arXiv* **2017**, arXiv:1706.03762. Available online: https://doi.org/10.48550/arXiv.1706.03762. (accessed 26 Sep 2025)
5. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2016**, arXiv:1409.0473. Available online: https://doi.org/10.48550/arXiv.1409.0473. (accessed 26 Sep 2025)
6. Pang, Y.; Lin, J.; Qin, T.; Chen, Z. Image-to-image translation: methods and applications. *IEEE. Trans. Multimedia.* **2021**, *24*, 3859-81. DOI
7. Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; et al. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661. Available online: https://doi.org/10.48550/arXiv.1406.2661. (accessed 26 Sep 2025)
8. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784. Available online: https://doi.org/10.48550/arXiv.1411.1784. (accessed 26 Sep 2025)
9. Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv* **2020**, arXiv:1703.10593. Available online: https://doi.org/10.48550/arXiv.1703.10593. (accessed 26 Sep 2025)
10. Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2022**, arXiv:1312.6114. Available online: https://doi.org/10.48550/arXiv.1312.6114. (accessed 26 Sep 2025)
11. Isola, P.; Zhu, J. Y.; Zhou, T.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honululu, USA. July 21-26, 2017. IEEE; 2017. pp. 5967-76. DOI
12. Wang, T. C.; Liu, M. Y.; Zhu, J. Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. *arXiv* **2018**, arXiv:1711.11585. Available online: https://doi.org/10.48550/arXiv.1711.11585. (accessed 26 Sep 2025)
13. Gatys, L. A.; Ecker, A. S.; Bethge, M. A neural algorithm of artistic style. *arXiv* **2015**, arXiv:1508.06576. Available online: https://doi.org/10.48550/arXiv.1508.06576. (accessed 26 Sep 2025)
14. Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA. June 27-30, 2016. IEEE; 2016. pp. 2414-23. DOI
15. Nie, D.; Trullo, R.; Lian, J.; et al. Medical image synthesis with deep convolutional adversarial networks. *IEEE. Trans. Biomed. Eng.* **2018**, *65*, 2720-30. DOI
16. Shi, Z.; Mettes, P.; Zheng, G.; Snoek, C. Frequency-supervised MR-to-CT image synthesis. *arXiv* **2021**, arXiv:2107.08962. Available online: https://doi.org/10.48550/arXiv.2107.08962. (accessed 26 Sep 2025)
17. Shao, X.; Zhang, W. SPatchGAN: a statistical feature based discriminator for unsupervised image-to-image translation. *arXiv* **2021**, arXiv:2103.16219. Available online: https://doi.org/10.48550/arXiv.2103.16219. (accessed 26 Sep 2025)
18. Wang, L.; Chae, Y.; Yoon, K. J. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. *arXiv* **2021**, arXiv:2109.01801. Available online: https://doi.org/10.48550/arXiv.2109.01801. (accessed 26 Sep 2025)
19. Yu, J.; Du, S.; Xie, G.; et al. SAR2EO: a high-resolution image translation framework with denoising enhancement. *arXiv* **2023**, arXiv:2304.04760. Available online: https://doi.org/10.48550/arXiv.2304.04760. (accessed 26 Sep 2025)

20. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: semantic propositional image caption evaluation. *arXiv* **2016**, arXiv:1607.08822. Available online: https://doi.org/10.48550/arXiv.1607.08822. (accessed 26 Sep 2025)
21. Li, S.; Tao, Z.; Li, K.; Fu, Y. Visual to text: survey of image and video captioning. *IEEE. Trans. Emerg. Top. Comput. Intell.* **2019**, *3*, 297-12. DOI
22. Żelaszczyk, M.; Mańdziuk, J. Cross-modal text and visual generation: a systematic review. Part 1: image to text. *Inf. Fusion.* **2023**, *93*, 302-29. DOI
23. He, X.; Deng, L. Deep learning for image-to-text generation: a technical overview. *IEEE. Signal. Process. Mag.* **2017**, *34*, 109-16. DOI
24. Indurthi, S.; Zaidi, M. A.; Lakumarapu, N. K.; et al. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada. June 06-11, 2021. IEEE; 2021. pp. 7723-7. DOI
25. Gállego, G. I.; Tsiamas, I.; Escolano, C.; Fonollosa, J. A. R.; Costa-jussà, M. R. End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. *arXiv* **2021**, arXiv:2105.04512. Available online: https://doi.org/10.48550/arXiv.2105.04512. (accessed 26 Sep 2025)
26. Wang, X.; Qiao, T.; Zhu, J.; Hanjalic, A.; Scharenborg, O. Generating images from spoken descriptions. *IEEE/ACM. Trans. Audio. Speech. Lang. Process.* **2021**, *29*, 850-65. DOI
27. Ning, H.; Zheng, X.; Yuan, Y.; Lu, X. Audio description from image by modal translation network. *Neurocomputing* **2021**, *423*, 124-34. DOI
28. Parmar, G.; Park, T.; Narasimhan, S.; Zhu, J. Y. One-step image translation with text-to-image models. *arXiv* **2024**, arXiv:2403.12036. Available online: https://doi.org/10.48550/arXiv.2403.12036. (accessed 26 Sep 2025)
29. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Y. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, USA. 2011. pp. 689-96. https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf. (accessed 26 Sep 2025)
30. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: a review and new perspectives. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2013**, *35*, 1798-828. DOI
31. Yu, H.; Gui, L.; Madaio, M.; Ogan, A.; Cassell, J.; Morency, L. P. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proceedings of the 25th ACM International Conference on Multimedia*. Association for Computing Machinery; 2017. pp. 1743-51. DOI
32. Siriwardhana, S.; Reis, A.; Weerasekera, R.; Nanayakkara, S. Jointly fine-tuning “BERT-like” self supervised models to improve multimodal speech emotion recognition. *arXiv* **2020**, arXiv:2008.06682. Available online: https://doi.org/10.48550/arXiv.2008.06682. (accessed 26 Sep 2025)
33. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *arXiv* **2021**, arXiv:2102.04830. Available online: https://doi.org/10.48550/arXiv.2102.04830. (accessed 26 Sep 2025)
34. Lai, S.; Hu, X.; Xu, H.; Ren, Z.; Liu, Z. Multimodal sentiment analysis: a survey. *arXiv* **2023**, arXiv:2305.07611. Available online: https://doi.org/10.48550/arXiv.2305.07611. (accessed 26 Sep 2025)
35. Le, H.; Sahoo, D.; Chen, N. F.; Hoi, S. C. H. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv* **2019**, arXiv:1907.01166. Available online: https://doi.org/10.48550/arXiv.1907.01166. (accessed 26 Sep 2025)
36. Tsai, Y. H. H.; Liang, P. P.; Zadeh, A.; Morency, L. P.; Salakhutdinov, R. Learning factorized multimodal representations. *arXiv* **2019**, arXiv:1806.06176. Available online: https://doi.org/10.48550/arXiv.1806.06176. (accessed 26 Sep 2025)
37. Pham, H.; Liang, P. P.; Manzini, T.; Morency, L. P.; Póczos, B. Found in translation: learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33. 2019. pp. 6892-9. DOI
38. Shang, C.; Palmer, A.; Sun, J.; et al. VIGAN: missing view imputation with generative adversarial networks. In *2017 IEEE International Conference on Big Data (Big Data)*, Boston, USA. December 11-14, 2017. IEEE; 2017. pp. 766-75. DOI
39. Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; Hu, Q. Deep partial multi-view learning. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2020**, *44*, 2402-15. DOI
40. Zhou, T.; Canu, S.; Vera, P.; Ruan, S. Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities. *Neurocomputing* **2021**, *466*, 102-12. DOI
41. Liu, Z.; Zhou, B.; Chu, D.; Sun, Y.; Meng, L. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Inf. Fusion.* **2024**, *101*, 101973. DOI
42. Lu, Z. Translation-based multimodal learning. Master’s thesis, Oakland University, 2024. https://www.secs.oakland.edu/~li4/research/student/MasterThesis_Lu2024.pdf. (accessed 26 Sep 2025)
43. Cordts, M.; Omran, M.; Ramos, S.; et al. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA. June 27-30, 2016. IEEE; 2016. pp. 3213-23. DOI
44. Tyleček, R.; Šára, R. Spatial pattern templates for recognition of objects with regular structure. In Weickert, J., Hein, M., Schiele, B.; editors. *Pattern Recognition. GCPR 2013*. Lecture Notes in Computer Science, vol 8142. Springer; 2013. pp. 364-74. DOI
45. Leong, C.; Rovito, T.; Mendoza-Schrock, O.; et al. Unified coincident optical and radar for recognition (UNICORN) 2008 dataset. https://github.com/AFRL-RY/data-unicorn-2008. (accessed 26 Sep 2025)
46. Tan, W. R.; Chan, C. S.; Aguirre, H. E.; Tanaka, K. Improved ArtGAN for conditional synthesis of natural image and artwork. *IEEE.*

Trans. Image. Process. **2019**, *28*, 394-409. DOI

47. Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile. December 07-13, 2015. IEEE; 2015. pp. 2641-9. DOI
48. Lin, T. Y.; Maire, M.; Belongie, S.; et al. Microsoft COCO: common objects in context. In Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T; editors. *Computer Vision - ECCV 2014*. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer; 2014. pp. 740-55. DOI
49. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: a 10 million image database for scene recognition. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2017**, *40*, 1452-64. DOI
50. Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; et al. Audio set: an ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA. March 05-09, 2017. IEEE; 2017. pp. 776-80. DOI
51. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L. P. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259. Available online: https://doi.org/10.48550/arXiv.1606.06259. (accessed 26 Sep 2025)
52. Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; Morency, L. P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics; 2018. pp. 2236-46. DOI
53. Busso, C.; Bulut, M.; Lee, C. C.; et al. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335-59. DOI
54. Tsai, Y. H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L. P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics; 2019. pp. 6558-69. DOI
55. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *arXiv* **2019**, arXiv:1911.05544. Available online: https://doi.org/10.48550/arXiv.1911.05544. (accessed on 26 Sep 2025)
56. Yu, J.; Chen, K.; Xia, R. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE. Trans. Affective. Comput.* **2023**, *14*, 1966-78. DOI
57. Jiang, K.; Wang, Q.; An, Z.; Wang, Z.; Zhang, C.; Lin, C. W. Mutual retinex: combining transformer and CNN for image enhancement. *IEEE. Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2240-52. DOI
58. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Lin, C. W.; Zhang, L. TTST: a top-k token selective transformer for remote sensing image super-resolution. *IEEE. Trans. Image. Process.* **2024**, *33*, 738-52. DOI
59. Jiang, K.; Wang, Z.; Chen, C.; Wang, Z.; Cui, L.; Lin, C. W. Magic ELF: image deraining meets association learning and transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery; 2022. pp, 827-36. DOI
60. Wang, Y.; He, J.; Wang, D.; Wang, Q.; Wan, B.; Luo, X. Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis. *Neurocomputing* **2024**, *572*, 127181. DOI
61. Wang, D.; Liu, S.; Wang, Q.; Tian, Y.; He, L.; Gao, X. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE. Trans. Multimedia.* **2023**, *25*, 4909-21. DOI
62. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; Shah, M. Transformers in vision: a survey. *ACM. Comput. Surv.* **2022**, *54*, 1-41. DOI
63. Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, arXiv:2011.13456. Available online: https://doi.org/10.48550/arXiv.2011.13456. (accessed on 26 Sep 2025)
64. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *arXiv* **2020**, arXiv:2006.11239. Available online: https://doi.org/10.48550/arXiv.2006.11239. (accessed on 26 Sep 2025)
65. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *arXiv* **2020**, arXiv:1907.05600. Available online: https://doi.org/10.48550/arXiv.1907.05600. (accessed on 26 Sep 2025)
66. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Jin, X.; Zhang, L. EDiffSR: an efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE. Trans. Geosci. Remote. Sens.* **2024**, *62*, 1-14. DOI
67. Shi, Y.; De, B. V.; Campbell, A.; Doucet, A. Diffusion Schrödinger Bridge matching. *arXiv* **2023**, arXiv:2303.16852. Available online: https://doi.org/10.48550/arXiv.2303.16852. (accessed on 26 Sep 2025)
68. Liu, G. H.; Vahdat, A.; Huang, D. A.; Theodorou, E. A.; Nie, W.; Anandkumar, A. I^{2}SB: Image-to-image Schrödinger Bridge. *arXiv* **2023**, arXiv:2302.05872. Available online: https://doi.org/10.48550/arXiv.2302.05872. (accessed on 26 Sep 2025)
69. Tang, Z.; Hang, T.; Gu, S.; Chen, D.; Guo, B. Simplified diffusion Schrödinger Bridge. *arXiv* **2024**, arXiv:2403.14623. Available online: https://doi.org/10.48550/arXiv.2403.14623. (accessed on 26 Sep 2025)
70. Chen, Z.; He, G.; Zheng, K.; Tan, X.; Zhu, J. Schrodinger bridges beat diffusion models on text-to-speech synthesis. *arXiv* **2023**, arXiv:2312.03491. Available online: https://doi.org/10.48550/arXiv.2312.03491. (accessed on 26 Sep 2025)
71. Özbey, M.; Dalmaz, O.; Dar, S. U. H.; Bedel, H. A.; Özturk, Ş.; Güngör, A. Unsupervised medical image translation with adversarial diffusion models. *IEEE. Trans. Med. Imaging.* **2023**, *42*, 3524-39. DOI

72. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. *arXiv* **2021**, arXiv:2105.05233. Available online: https://doi.org/10.48550/arXiv.2105.05233. (accessed on 26 Sep 2025)
73. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery; 2020. pp. 1122-31. DOI
74. Zhao, T.; Kong, M.; Liang, T.; Zhu, Q.; Kuang, K.; Wu, F. CLAP: contrastive language-audio pre-training model for multi-modal sentiment analysis. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. Association for Computing Machinery; 2023. pp. 622-6. DOI
75. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. June 18-24, 2022. IEEE; 2022. pp. 15979-88. DOI
76. Bischke, B.; Helber, P.; König, F.; Borth, D.; Dengel, A. Overcoming missing and incomplete modal-ities with generative adversarial networks for building footprint segmentation. *arXiv* **2018**, arXiv:1808.03195. Available online: https://doi.org/10.48550/arXiv.1808.03195. (accessed on 26 Sep 2025)
77. Hamghalam, M.; Frangi, A. F.; Lei, B.; Simpson, A. L. Modality completion via Gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*. Springer, Cham; 2021. pp. 442-52. DOI
78. Zhou, T.; Canu, S.; Vera, P.; Ruan, S. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE. Trans. Image. Process.* **2021**, *30*, 4263-74. DOI
79. Sun, J.; Zhang, X.; Han, S.; Ruan, Y. P.; Li, T. RedCore: relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. *Proc. AAAI. Conf. Artif. Intell.* **2024**, *38*, 15173-82. DOI
80. Park, K. R.; Lee, H. J.; Kim, J. U. Learning trimodal relation for audio-visual question answering with missing modality. *arXiv* **2024**, arXiv:2407.16171. Available online: https://doi.org/10.48550/arXiv.2407.16171. (accessed on 26 Sep 2025)
81. Kim, D.; Kim, T. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. *arXiv* **2024**, arXiv:2407.12616. Available online: https://doi.org/10.48550/arXiv.2407.12616. (accessed on 26 Sep 2025)
82. Guo, Z.; Jin, T.; Zhao, Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. *arXiv* **2024**, arXiv:2407.05374. Available online: https://doi.org/10.48550/arXiv.2407.05374. (accessed on 26 Sep 2025)
83. Lin, X.; Wang, S.; Cai, R.; et al. Suppress and rebalance: towards generalized multi-modal face anti-spoofing. *arXiv* **2024**, arXiv:2402.19298. Available online: https://doi.org/10.48550/arXiv.2402.19298. (accessed on 26 Sep 2025)
84. Lu, Z.; Ewing, R.; Blasch, E.; Li, J. Explainable diffusion model via Schrödinger Bridge in multimodal image translation. In *Dynamic data driven applications systems*. Springer, Cham; 2026. pp. 391-402. DOI
85. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: https://doi.org/10.48550/arXiv.1409.1556. (accessed on 26 Sep 2025)
86. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA. June 27-30, 2016. IEEE; 2016. pp. 2818-26. DOI
87. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, Bellevue, USA, July 02, 2012. PMLR; 2012. pp. 37-49. https://proceedings.mlr.press/v27/baldi12a.html. (accessed on 26 Sep 2025)
88. Hafner, S.; Ban, Y. Multi-modal deep learning for multi-temporal urban mapping with a partly missing optical modality. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, Pasadena, USA. July 16-21, 2023. IEEE; 2023. pp. 6843-6. DOI