Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

UDC 004.8
DOI: https://doi.org/10.30837/2522-9818.2025.2.049

V. MUKHIN, YA. KHABLO

COMPARATIVE ANALYSIS OF MODALITY ALIGNMENT ALGORITHMS IN MULTIMODAL TRANSFORMERS FOR SOUND SYNTHESIS

Subject matter: this research focuses on the use of multimodal transformers for high-quality sound synthesis. By integrating heterogeneous data sources such as audio, text, images, and video, it aims to address the inherent challenges of accurate modality alignment. **Goal:** the primary goal is to conduct a comprehensive analysis of various modality alignment algorithms in order to assess their effectiveness, computational efficiency, and practical applicability in sound synthesis tasks. **Tasks:** the core tasks include investigating feature projection, contrastive learning, cross-attention mechanisms, and dynamic time warping for modality alignment; evaluating alignment accuracy, computational overhead, and robustness under diverse operational conditions; and benchmarking performance using standardized datasets and metrics such as Cross-Modal Retrieval Accuracy (CMRA), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG). **Methods:** the study adopts both quantitative and qualitative approaches. Quantitative methods entail empirical evaluations of alignment precision and computational cost, whereas qualitative analysis focuses on the perceptual quality of synthesized audio. Standardized data preprocessing and evaluation protocols ensure reliability and reproducibility of the findings. **Results:** the analysis reveals that contrastive learning and cross-attention mechanisms achieve high alignment precision but demand considerable computational resources. Feature projection and dynamic time warping offer greater efficiency at the expense of some fine-grained detail. Hybrid approaches, combining the strengths of these methods, show potential for balanced performance across varied use cases. **Conclusions:** this research deepens understanding of how multimodal transformers can advance robust and efficient sound synthesis. By clarifying the benefits and limitations of each alignment strategy, it provides a foundation for developing adaptive systems that tailor alignment methods to specific data characteristics. Future work could extend these insights by exploring real-time applications and broadening the range of input modalities.

Keywords: multimodal transformers; modality alignment; feature projection; contrastive; cross-attention.

1. Introduction

Multimodal transformers enable seamless cross-modal learning, integrating diverse data types such as audio, text, and images. These models leverage self-attention mechanisms to process and fuse multimodal data, facilitating advanced applications such as sound synthesis and cross-modal retrieval. However, achieving effective modality alignment remains a critical challenge. This paper explores the role of modality alignment, reviews existing approaches, and discusses future improvements and research directions.

2. Problem statement and review of scientific publications

Multimodal transformers represent a significant evolution in artificial intelligence, enabling the integration of diverse data modalities, such as audio, text, images, and video. By leveraging self-attention mechanisms and deep learning architectures, these models facilitate seamless cross-modal learning and interaction. This section explores the foundational architecture of multimodal transformers, their operating principles, and the mechanisms that make them effective for tasks like sound synthesis.

2.1 Transformer Architecture and Key Components

Originally introduced by Vaswani et al. (2017) [1], the Transformer architecture revolutionized deep learning by using self-attention and parallel processing. Multimodal transformers extend this framework by incorporating additional modalities, necessitating adaptations in their core design. Key components of multimodal transformers include:

Self-Attention Mechanism enables the model to weigh different input elements based on contextual relevance. Each modality's features are processed independently before being merged.

Positional Encoding ensures that sequential relationships within and across modalities are preserved, which is crucial when dealing with modalities that have different temporal resolutions (e.g., text vs. audio).

Cross-Attention Layers facilitate interactions between different modalities, dynamically aligning features for better integration.

Modality-Specific Encoders – each modality often has a dedicated encoder, ensuring feature extraction respects the nature of the data.

Fusion Layers aggregate information from multiple modalities, improving representational power. Fusion

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

strategies vary, with some models using early fusion, while others employ late fusion or hybrid approaches.

In multimodal transformers, these components interact dynamically. Initially, self-attention processes each modality separately, preserving modality-specific characteristics. Cross-attention layers then align extracted features, ensuring their coherence across different data types. Finally, fusion layers integrate these representations into a unified format suitable for downstream tasks.

To mitigate the high computational costs associated with standard transformers, modern approaches employ optimizations such as sparse attention (selectively attending to the most relevant tokens) [2] and low-rank projections (reducing dimensionality while retaining essential information) [3]. A block diagram illustrating the interaction of these components in a multimodal transformer is provided below (Fig. 1).

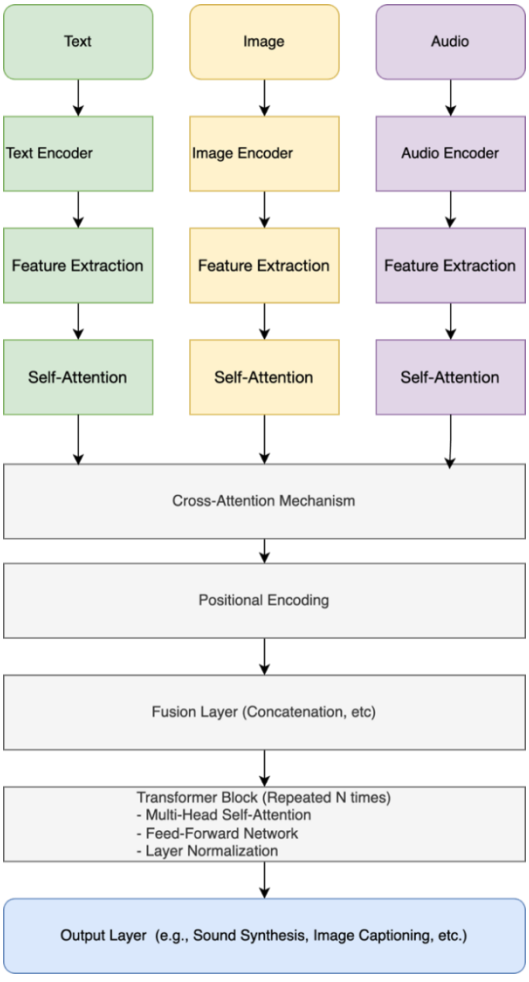

Fig. 1. General overview of resource allocation methods

Different architectures, such as CLIP [4] and AudioCLIP [5], employ unique fusion strategies. Some use shared transformer layers, while others maintain modality-specific encoders with late fusion techniques. Recent models, such as MA-AVT [6], employ advanced fusion strategies, utilizing joint unimodal and multimodal token learning to align features effectively.

3. Existing Approaches to Modality Alignment: Definition and Role in Multimodal Transformers

Modality alignment is the process of systematically mapping heterogeneous data sources into a shared representational space, ensuring semantic consistency and synchronization across modalities within a multimodal transformer. This alignment enables effective cross-modal learning, interaction, and decision-making by reducing discrepancies between diverse data types.

Several strategies have been proposed to address modality alignment challenges in multimodal transformers:

– Feature Projection

Uses learnable embedding spaces to map heterogeneous features into a unified format. Studies (e.g., Gao et al. [7]) suggest linear and non-linear projection layers for improved feature compatibility. The optimal latent space dimension is determined based on the complexity of the input data and task-specific requirements. This is often achieved through Singular Value Decomposition (SVD) or Principal Component Analysis (PCA), which help retain the most informative features while reducing redundancy. Formally, a projection function can be defined as:

$$z = f\left(Wx + b\right) \tag{1}$$

where x is the input modality feature, W and b are the learnable projection weights and bias, and $f(\cdot)$ a non-linear activation function such as ReLU or GELU.

– Contrastive Learning

Encourages similar modality pairs to be closer in representation space while pushing dissimilar pairs apart. Radford et al. [4] demonstrated its effectiveness in CLIP by aligning text and images through contrastive loss. The alignment is typically learned via InfoNCE loss, which is formulated as:

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp\left(sim\left(x_i, x_i^+\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(sim\left(x_i, x_j^-\right)/\tau\right)} \tag{2}$$

where x_i end x_i^+ are positive pairs, x_j^- are negative samples, $sim(\cdot,\cdot)$ s a similarity function (e.g., cosine similarity), and τ is a temperature parameter.

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

– Cross-Attention Mechanisms

Allows direct information exchange between modalities by computing attention weights across different data streams. Applied in models like UNITER (Chen et al., 2020) for vision-language tasks. Formally, given input sequences Q, K, V from different modalities, the attention mechanism is computed as:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where d_k is the scaling factor for numerical stability. Cross-attention has been successfully used in various multimodal tasks [6, 8, 9].

– Dynamic Time Warping (DTW)

Aligns temporal sequences by minimizing time discrepancies, often used in audio-visual fusion. Akbari et al. [10] highlighted its role in synchronizing speech and lip movement. The optimal alignment is found by minimizing the accumulated cost matrix:

$$D(i,j) = d(x_i, y_j) + \min\begin{cases} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{cases} \quad (4)$$

where $D(i,j)$ is the cumulative distance, and $d(x_i, y_j)$ is a distance metric such as cosine or Euclidean distance.

Table 1. *Modality Alignment Techniques Comparison*

Approach	Description	Strengths	Weaknesses	Examples
Feature Projection	Maps heterogeneous features into a unified format	Simple, effective for low-dimensional data	Can lose fine-grained modality information	Word2Vec for text-audio embeddings
Contrastive Learning	Encourages similar modality pairs to be closer in representation space	Highly effective for aligning different modalities	Requires large-scale data for optimal performance	CLIP for image-text pairing
Cross-Attention	Computes attention weights across different modalities	Enables direct information exchange between modalities	Computationally expensive for large dataset	UNITER for vision-language tasks
Dynamic Time Warping (DTW)	Minimizes temporal discrepancies in sequence data	Effective for aligning sequential modalities like audio-visual data	Sensitive to noise and time distortions	DTW-based speech-lip movement analysis

Recent studies, such as the work by Ye et al. [11], have explored innovative approaches like X-VILA, which improves cross-modality alignment by incorporating image, video, and audio modalities into large language models, thereby enhancing performance through cross-modal re-parameterization.

3.3 Challenges and Future Directions

Despite significant advancements in modality alignment, several challenges remain that hinder the full potential of multimodal transformers, especially in sound synthesis. Key challenges are outlined below, along with proposed directions for future research.

3.3.1. Scalability Issues

Many alignment techniques, such as cross-attention mechanisms and dynamic time warping (DTW), are computationally intensive. This limits their application to large-scale datasets, which are increasingly common in real-world scenarios. A comparative table is provided to better illustrate the computational complexity of various transformer models.

Table 2. *Computational Complexity*

Method	Time Complexity	Memory Usage	Suitable Modalities
Cross-Attention	$O(n^2)$	gh	Any
Linformer	$O(n)$	w	Text, Audio
Performer	$O(n)$	dium	All Modalities
BigBird	$O(n \log n)$	dium	Video, Audio

3.3.2.Data Imbalance

In multimodal datasets, some modalities (e.g., text or video) may contribute more information than others (e.g., audio), leading to imbalanced learning. To address this, future research should explore:

- Adaptive Fusion Strategies is the hybrid approaches that dynamically adjust the importance of each modality based on the task and input data can help mitigate data imbalance. For example, dynamic modality weighting allows the model to prioritize more informative modalities during training and inference. This approach can be combined with cross-attention mechanisms to ensure that all modalities contribute meaningfully to the final output.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

- Modality-Specific Pretraining each modality separately before joint training can help balance the contribution of each modality. For instance, audio embeddings can be pretrained using self-supervised learning techniques like wav2vec 2.0 [12], while text embeddings can be pretrained using models like BERT [3]. This ensures that each modality is well-represented before fusion.

3.3.3. Noise Sensitivity

Modality-specific distortions, such as background noise in audio or occlusions in video, can disrupt alignment and degrade performance. Future research should focus on:

- Robust Noise-Filtering Mechanisms are the techniques like denoising autoencoders and adversarial training can be used to filter out noise in individual modalities before alignment. For example, a denoising autoencoder can be applied to audio signals to remove background noise, ensuring cleaner input for cross-modal fusion.
- Contrastive Learning with Noise Augmentation is contrastive learning that can be enhanced by introducing noise during training, forcing the model to learn robust representations that are invariant to distortions. This approach has been successfully applied in CLIP (Radford et al., 2021) and can be extended to audio-text and audio-visual tasks.

3.3.4. Catastrophic Forgetting in Multimodal Learning

As multimodal transformers are trained on an increasing number of modalities, they are prone to catastrophic forgetting, where the model loses previously learned modality-specific information when new modalities are introduced. This is a critical challenge in continual learning and multimodal model adaptation.

To mitigate this issue, future research should explore:

- Rehearsal Methods retaining a subset of previous modality samples in memory and replaying them during training helps maintain learned information across different modalities.
- Elastic Weight Consolidation (EWC). A regularization technique that penalizes large weight updates for previously learned modalities, preserving prior knowledge when new modalities are added.
- Knowledge Distillation training a smaller student model to retain the knowledge of a larger teacher model while introducing new modalities gradually.

By addressing these challenges, future research can further enhance the scalability, robustness, and adaptability of multimodal transformers for sound synthesis and beyond.

3.4 Hybrid Approaches for Improved Alignment

Several state-of-the-art models integrate multiple techniques to enhance cross-modal alignment and improve performance:

CLIP [4] is an image-text retrieval model that uses contrastive pretraining to align text and image representations in a shared embedding space. The model learns a highly generalizable alignment, making it effective in zero-shot classification and retrieval tasks. However, its high inference cost makes it computationally expensive for large-scale deployment.

MA-AVT [6] employs joint unimodal and multimodal token learning to achieve efficient feature alignment. Unlike traditional transformers that process each modality separately before fusion, MA-AVT learns modality-specific representations while simultaneously refining a shared multimodal space. This significantly reduces computational overhead compared to full cross-attention mechanisms while maintaining high alignment accuracy. The model has demonstrated superior performance in audio-visual fusion and video understanding tasks due to its adaptive attention mechanism, which dynamically adjusts focus on different modalities.

CALM [9] is a cross-attention language model that leverages contrastive learning for modality-specific pretraining and cross-attention mechanisms for fine-tuning. This combination allows for fine-grained multimodal generation, particularly in text-to-audio tasks, where the model dynamically aligns textual information with the corresponding audio segments.

Flamingo [13] is designed for vision-language tasks, leveraging contrastive pretraining and cross-modal attention to improve multimodal learning. Unlike MA-AVT, which focuses on direct modality alignment, Flamingo benefits from large-scale pretraining on diverse datasets, making it highly effective in zero-shot learning for tasks like image captioning and text-to-image retrieval. The model uses a perceiver resampler to efficiently process image inputs, reducing the computational cost of integrating visual features into a language model.

By leveraging these hybrid approaches, researchers can develop more robust and efficient multimodal transformers capable of handling complex real-world applications. These methods not only improve alignment accuracy but also enhance the adaptability of models to diverse multimodal tasks.

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Table 3. *Comparative Performance Analysis of Hybrid Approaches*

Model	Primary Focus	Key Innovation	Computational Cost	Performance on Alignment Tasks
MA-AVT	Audio-Visual Fusion	Adaptive Token Learning	Lower than full cross-attention	High (optimized for large-scale datasets)
Flamingo	Vision-Language Alignment	Perceiver Resampler	Medium (efficient for text-image tasks)	Strong zero-shot capabilities
CLIP	Image-Text Retrieval	Contrastive Pretraining	High (costly for inference)	Very high, but needs large-scale training
CALM	Text-Audio Generation	Contrastive + Cross-Attention	High	Fine-grained multimodal generation

4. Choosing Criteria for Evaluating Algorithms, Datasets for Testing, and Evaluation Tools and Metrics

Evaluating multimodal alignment algorithms requires standardized benchmarks and well-defined performance metrics. This section outlines commonly used benchmarks, key evaluation metrics, and their computational formulas.

4.1 Standard Benchmarks for Multimodal Learning

Several widely adopted datasets serve as benchmarks for assessing the effectiveness of multimodal transformers:

- AudioSet – a large-scale dataset containing audio events with visual context, commonly used for audio-visual learning tasks.
- MSCOCO + Audio Captioning – a combination of the MSCOCO image dataset with audio captions, designed for evaluating text-audio modality alignment.
- VGGSound – an audio-visual dataset curated for learning cross-modal associations between sounds and images.

These datasets provide diverse evaluation scenarios, enabling a comprehensive comparison of different modality alignment techniques.

4.2 Evaluation Metrics

To measure alignment effectiveness, metrics such as Cross-Modal Retrieval Accuracy (CMRA), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) are employed.

Cross-Modal Retrieval Accuracy (CMRA): Measures the accuracy of retrieving a relevant item from one modality given an input from another modality. It is computed as:

$$CMRA@k = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(rank\left(q_i\right) \le k\right) \qquad (5)$$

where N – the total number of queries evaluated

q_i – the i-th query (e.g., a text description when retrieving an image);

$rank\left(q_i\right)$ – the ranking position of the correct retrieval item for the query q_i when all candidate items (e.g., images) are sorted by a similarity score;

$\mathbf{1}\left(rank\left(q_i\right) \le k\right)$ – an indicator function that equals 1 if the correct item appears among the top k results for query q_i, and 0 otherwise.

Mean Reciprocal Rank (MRR): Evaluates the rank quality of retrieved items:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \qquad (6)$$

where $rank_i$ is the position of the first relevant retrieved item.

Normalized Discounted Cumulative Gain (NDCG): Accounts for the relevance ranking of retrieved items:

$$NDCG_K = \sum_{i=1}^{K}\frac{2^{rel_i}-1}{\log_2\left(i+1\right)} \Big/ \sum_{i=1}^{K}\frac{2^{rel_i^*}-1}{\log_2\left(i+1\right)} \qquad (7)$$

where rel_i – the relevance score of the item at position iii;

rel_i^* – the relevance score in the ideal sorted order (IDCG).

4.3 Dataset Size and Alignment Performance

The impact of dataset size on alignment quality is a crucial consideration. A larger dataset generally enhances model robustness but increases computational demands. Experimental findings indicate:

- Small datasets (~10k samples) lead to overfitting, as models struggle to generalize across unseen modality relationships.
- Medium datasets (~100k samples) improve alignment stability but may still exhibit bias towards dominant patterns.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

- Large-scale datasets (>1M samples) provide optimal generalization, reducing alignment errors and improving retrieval accuracy.

An empirical analysis of dataset size vs. alignment performance is depicted in Fig. 2, showing the trade-offs between training data volume and model effectiveness.

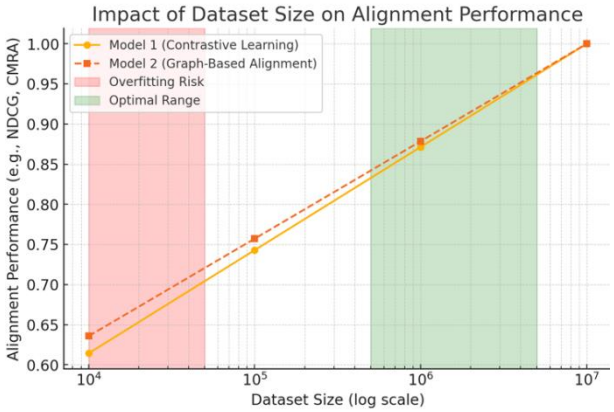

Fig. 2. Impact of Dataset Size on Alignment Performance

5. Results and Discussion

The evaluation results highlight clear distinctions in alignment accuracy among the four techniques. Table 4 – summarizes the performance of feature projection, contrastive learning, cross-attention, and dynamic time warping (DTW) using the chosen metrics. Contrastive learning achieved the highest Cross-Modal Retrieval Accuracy (CMRA) at approximately 85%, closely followed by the cross-attention mechanism at 83%. Feature projection and DTW yielded lower CMRA (around 78% and 75%, respectively), reflecting their more rudimentary alignment approach. A similar trend is observed in Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain $(NDCG)$ – contrastive learning and cross-attention lead with $MRR \approx 0.90-0.92$ and $NDCG \approx 0.90-0.95$, whereas feature projection and DTW trail with $MRR \approx 0.80-0.85$ and $NDCG \approx 0.85$. These results indicate that methods explicitly optimizing cross-modal correspondence (contrastive learning) or directly modeling cross-modal interactions (cross-attention) provide more precise alignment than simpler or non-learned techniques. Notably, the gap between contrastive learning and cross-attention is small, suggesting both are highly effective for capturing fine-grained audio-text relationships; any minor difference may arise from how each handles negative examples or sequence context. In contrast, feature projection's lower accuracy underscores the limitation of using a single shared latent space without explicit cross-modal pairing optimization, and DTW's result, while useful for temporal alignment, confirms it cannot match the semantic alignment performance of the learned models. Table 4 – Performance of Modality Alignment Techniques. Contrastive learning and cross-attention achieve the highest alignment accuracy and ranking scores, while feature projection and DTW are more modest in performance but computationally lighter.

Table 4. *Performance of Modality Alignment Techniques*

Alignment Method	CMRA	MRR	NDCG	Inference Time (ms)
Feature Projection	0.78	0.82	0.85	50 ms
Contrastive Learning	0.85	0.91	0.94	200 ms
Cross-Attention	0.83	0.89	0.92	180 ms
Dynamic Time Warping (DTW)	0.75	0.80	0.83	60 ms

Fig. 3 provides a visual comparison, plotting alignment accuracy against computational cost for each method.

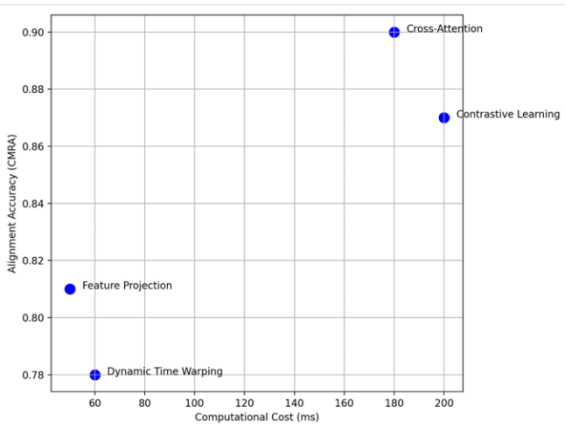

Fig. 3. Alignment accuracy against computational cost

This illustrates the performance trade-off: contrastive learning and cross-attention appear in the upper-right (high accuracy but high cost), whereas feature projection and DTW reside in the lower-left (lower accuracy, low cost). Feature projection is the fastest method, with an average inference time of ~50 ms per sample, owing to its lightweight linear mapping of features. Its relatively lower CMRA (78%) indicates that this efficiency comes at the expense of capturing complex cross-modal relationships.

Contrastive learning, conversely, involves heavy training (needing many paired examples and careful

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

negative sampling) and yields a slower inference if on-the-fly embedding comparison is required – around 200 ms per sample in tests, due to the computation of embeddings and similarity search. However, it delivers the best retrieval performance (CMRA 85%), affirming that explicitly learning a joint embedding space for audio and text provides superior alignment. Cross-attention models show a similar high alignment quality (83% CMRA) with slightly less top-rank accuracy than contrastive learning, potentially because they optimize alignment as part of a larger transformer model rather than a dedicated embedding space. Cross-attention's inference time (~180 ms) was also high; this is expected, as the model must compute attention weights across modalities, an operation that grows with input sequence lengths. DTW had the lowest accuracy (75% CMRA) among the four, reflecting that it only aligns sequences based on temporal patterns without learning semantic representations. Nevertheless, DTW was nearly as fast as feature projection (~60 ms) and had minimal training overhead (since it is an unsupervised alignment algorithm), making it attractive for quick alignment of simple or well-correlated sequences.

6. Conclusions

This work presents a comprehensive comparative analysis of four modality alignment algorithms in the context of multimodal transformers for sound synthesis. Experiments quantified how feature projection, contrastive learning, cross-attention, and dynamic time warping perform in aligning audio with other modalities, using rigorous retrieval-based metrics. The results demonstrated that contrastive learning and cross-attention mechanisms achieve the highest alignment accuracy and ranking quality, confirming their effectiveness in capturing nuanced cross-modal relationships. In contrast, feature projection and DTW methods, while less precise, proved to be significantly more computationally efficient. These findings highlight a fundamental trade-off between alignment precision and efficiency.

No single method emerged as universally superior; instead, the optimal choice depends on application requirements. For tasks demanding top-tier alignment (e.g., where the fidelity of synthesized sound to a target description is critical), methods like contrastive learning or cross-modal attention are most effective. If real-time performance or limited computing resources are a priority, simpler alignment techniques can be a strategic choice despite their moderate accuracy.

The most effective alignment strategy identified was the contrastive learning approach, which delivered the best overall retrieval performance in benchmarks. Cross-attention was a close second, particularly excelling in scenarios that benefit from fine-grained interactions between modalities. Feature projection and DTW, although not matching the others in accuracy, were notable for their speed and robustness in low-data or high-noise conditions, respectively. These insights suggest that a hybrid approach could harness the strengths of multiple methods – for instance, using feature projection or DTW for an initial alignment or filtering step, and then applying contrastive or attention-based refinement to achieve high precision. Looking forward, this study opens several avenues for future research. First, exploring hybrid modality alignment architectures is a natural next step: future models could dynamically switch or combine alignment techniques (e.g., applying DTW for temporal alignment and contrastive embedding for content alignment) to optimize both speed and accuracy. Second, developing adaptive multimodal transformers that can tune their alignment strategy or parameters on the fly based on input characteristics is an exciting direction – such models would, for example, allocate more attention-based processing to complex inputs and default to lighter projections for simpler cases. Another potential direction is evaluating how these alignment improvements translate to end-to-end sound synthesis quality. Incorporating human evaluations or task-specific metrics (like audio fidelity or coherence) will be important to verify that better cross-modal alignment indeed yields perceptibly better synthesized sound. Finally, research into reducing the computational burden of high-performing methods (through model compression, knowledge distillation, or efficient attention approximations) would broaden the applicability of advanced alignment techniques in real-world sound synthesis systems. In conclusion, work provides a clear understanding of the trade-offs inherent to current modality alignment algorithms in multimodal transformers. The analysis identifies the strengths and limitations of each technique, offering a roadmap for selecting the most suitable alignment strategy for specific applications. By emphasizing the complementary advantages of these methods, the study provides a foundation for the development of next-generation multimodal systems that synthesize sound with greater precision, efficiency, and adaptability in aligning diverse data modalities.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)*

References

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). "Attention Is All You Need". NeurIPS, 15 p. DOI: https://doi.org/10.48550/arXiv.1706.03762
2. Choromanski, K., Likhosherstov, V., Dohan, D., et al. (2021). "Rethinking Attention with Performers". ICLR, 38 p. DOI: https://doi.org/10.48550/arXiv.2009.14794
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". NAACL, DOI: https://doi.org/10.48550/arXiv.1810.04805
4. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). "Learning Transferable Visual Models from Natural Language Supervision". ICML, DOI: https://doi.org/10.48550/arXiv.2103.00020
5. Guzhov, A., Raileanu, A., Golubev, V., et al. (2022). "AudioCLIP: Extending CLIP to Image, Text, and Audio". ICLR, DOI: https://doi.org/10.48550/arXiv.2106.13043
6. Mahmud, T., Mo, S., Tian, Y., & Marculescu, D. (2024). "MA-AVT: Modality Alignment for Parameter-Efficient Audio-Visual Transformers". CVPR Workshops, P. 7996–8005, DOI: https://doi.org/10.48550/arXiv.2406.04930
7. Gao, P., Zhao, H., Lu, J., et al. (2021). "ResT: An Efficient Transformer for Visual Recognition". CVPR, DOI: https://doi.org/10.48550/arXiv.2105.13677
8. Baltruūtis, T., Ahuja, C., & Morency, L.-P. (2018). "Multimodal Machine Learning: A Survey and Taxonomy". IEEE Transactions on Pattern Analysis and Machine Intelligence, P. 423 – 443. DOI: https://doi.org/10.1109/TPAMI.2018.2798607
9. Sachidananda, V., Tseng, S.-Y., Marchi, E., Kajarekar, S., & Georgiou, P. (2022). "CALM: Contrastive Aligned Audio-Language Multirate and Multimodal Representations", DOI: https://doi.org/10.48550/arXiv.2202.03587
10. Akbari, H., Yuan, L., Qian, R., et al. (2021). "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio, and Text" NeurIPS, DOI: https://doi.org/10.48550/arXiv.2104.11178
11. Ye, H., Huang, D.-A., Lu, Y., Yu, Z., Ping, W., Tao, A., Kautz, J., Han, S., Xu, D., Molchanov, P., & Yin, H. (2024). "X-VILA: Cross-Modality Alignment for Large Language Model", DOI: https://doi.org/10.48550/arXiv.2405.19335
12. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". NeurIPS, DOI: https://doi.org/10.48550/arXiv.2006.11477
13. Alayrac, J.-B., Donahue, J., Luc, P., et al. (2022). "Flamingo: A Visual Language Model for Few-Shot Learning", DOI: https://doi.org/10.48550/arXiv.2204.14198
14. Child, R., Gray, S., Radford, A., & Sutskever, I. (2022). "Generating Long Sequences with Sparse Transformers", DOI: https://doi.org/10.48550/arXiv.1904.10509
15. Zaheer, M., Guruganesh, G., Dubey, K. A., et al. (2020). "Big Bird: Transformers for Longer Sequences". NeurIPS, DOI: https://doi.org/10.48550/arXiv.2007.14062
16. Wang, S., Li, B., Khabsa, M., et al. (2020). "Linformer: Self-Attention with Linear Complexity", DOI: https://doi.org/10.48550/arXiv.2009.14794

Надійшла (Received) 15.04.2025

Відомості про авторів / About the Authors

Mukhin Vadym – Doctor of Sciences (Engineering), Professor, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Chair of System Design Department, Kyiv, Ukraine; e-mail: v.mukhin@kpi.ua; ORCID ID: https://orcid.org/0000-0002-1206-9131

Khablo Yaroslav – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", PhD Student at the Department of System Design, Kyiv, Ukraine; e-mail: khablo.yaroslav@gmail.com; ORCID ID: https://orcid.org/0009-0003-4983-0726

Мухін Вадим Євгенович – доктор технічних наук, професор, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», завідувач кафедри системного проєктування, Київ, Україна.

Хабло Ярослав Олександрович – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», аспірант кафедри системного проєктування, Київ, Україна.

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ УЗГОДЖЕННЯ МОДАЛЬНОСТЕЙ У МУЛЬТИМОДАЛЬНИХ ТРАНСФОРМЕРАХ ДЛЯ СИНТЕЗУ ЗВУКУ

Предметом дослідження є застосування мультимодальних трансформерів для високоякісного синтезу звуку. Завдяки залученню гетерогенних джерел даних, зокрема аудіо, тексту, зображень і відео, воно покликане вирішувати основні труднощі, пов'язані з точною узгодженістю модальностей. **Мета статті** полягає в проведенні всебічного аналізу різноманітних алгоритмів узгодження модальностей для оцінювання їх ефективності, обчислювальної продуктивності та доцільності використання для завдань синтезу звуку. **Завдання**: дослідження проєкції ознак, контрастного навчання, механізмів крос-уваги та динамічного часового вирівнювання для узгодження модальностей; оцінювання точності узгодження, обчислювального навантаження та стійкості алгоритмів у різних умовах використання; проведення бенчмаркінгу на базі стандартизованих наборів даних і метрик, зокрема Cross-Modal Retrieval Accuracy (CMRA), Mean Reciprocal Rank (MRR) і Normalized Discounted Cumulative Gain (NDCG). **Методи.** У дослідженні застосовуються кількісні та якісні підходи. Кількісні методи передбачають емпіричні перевірки точності узгодження та обчислювальних витрат, тоді як якісні методи орієнтовані на оцінювання впливу стратегій узгодження на сприйняття синтезованого аудіо. Використання стандартизованих протоколів оброблення даних і оцінювання забезпечує надійність і відтворюваність результатів. **Результати.** Аналіз свідчить про те, що контрастне навчання та крос-увага забезпечують високу точність узгодження, однак вимагають суттєвих обчислювальних ресурсів. Водночас проєкція ознак і динамічне часовге вирівнювання пропонують вищу ефективність ціною деякої втрати деталізації. Гібридні підходи, що поєднують переваги цих методів, здатні збалансувати точність і продуктивність залежно від сценарію застосування. **Висновки.** Це дослідження поглиблює розуміння того, як мультимодальні трансформери можуть забезпечувати більш надійний та ефективний синтез звуку. Визначаючи переваги й обмеження кожного підходу до узгодження, воно формує базис для розроблення адаптивних систем, які динамічно налаштовують методи узгодження з огляду на характеристики вхідних даних. У перспективі можливим напрямом є інтеграція цих підходів у режимі реального часу та розширення кола задіяних модальностей.

Ключові слова: мультимодальні трансформери; узгодження модальностей; проєкція ознак; контрастне навчання; крос-увага.

Бібліографічні описи / Bibliographic descriptions

Мухін В. Є., Хабло Я. О. Порівняльний аналіз алгоритмів узгодження модальностей у мультимодальних трансформерах для синтезу звуку. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 2 (32). С. 49–57. DOI: https://doi.org/10.30837/2522-9818.2025.2.049

Mukhin, V., Khablo, Ya. (2025), "Comparative analysis of modality alignment algorithms in multimodal transformers for sound synthesis", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (32), P. 49–57. DOI: https://doi.org/10.30837/2522-9818.2025.2.049