Gaze and Head Joint Representation Learning

Ashwin Murali
M.Sc Computer Science (Student - 3588671)
Universität Stuttgart
Stuttgart, 70569, Germany
ashwin.cse18@gmail.com

Souptik K. Majumdar
M.Sc Computer Science (Student - 3638136)
Universität Stuttgart
Stuttgart, 70569, Germany
souptikmajumdar98@gmail.com

Abstract

Accurate capture and interpretation of gaze data in digital interactions encounter substantial hurdles, primarily attributable to frequent missing data resulting from calibration issues, human errors, and inherent actions such as blinking. Overcoming these challenges is imperative for enhancing the accuracy of digital interactions and subsequent data analysis. This paper addresses these challenges by proposing a novel approach that leverages head movement data to enhance and reconstruct incomplete gaze data. By integrating head movement information, our method aims to mitigate the impact of missing data, ultimately improving the precision and reliability of gaze data in digital interactions. Our Multi-modal architecture exploits the correlations between head and gaze data to effectively reconstructs gaze values, mitigating the impact of missing data by leveraging the inherent relationships between these two crucial elements. This architecture presents a promising solution to the challenges associated with gaze data accuracy in digital interactions, showcasing its potential for enhancing user experience and advancing research in the field.

1 Introduction

Efficient and accurate capture of gaze data stands as a cornerstone in the realm of digital interactions, playing a key role in enhancing user experience, understanding user behavior, and informing the design of responsive interfaces. Despite its critical importance, the reliability of gaze data is frequently compromised by challenges that manifest in the form of missing data. These challenges, rooted in calibration issues, human errors, and inherent ocular phenomena like blinking, pose formidable obstacles to the seamless functioning of digital interaction systems and the precision of subsequent analytical endeavors. The nature of missing gaze data necessitates innovative solutions that go beyond conventional methodologies. In response to this imperative, we present an approach that harnesses head movement data as a complementary source of information to augment and reconstruct incomplete gaze data. This augmentation not only addresses the immediate challenge of missing data but also holds the promise of significantly improving the overall accuracy and robustness of gaze tracking in digital interactions. At the heart of our proposed solution lies the Multi-modal architecture that exploits the inherent correlations between head and gaze data, acknowledging the symbiotic relationship between these two critical components of the user's interaction profile. By leveraging this relationship, the Multimodal seeks to transcend the limitations imposed by missing gaze data, offering a novel perspective on data reconstruction that holds great potential for transformative advancements in the field.

In the subsequent sections of this paper, we embark on the Background, Methodology, Results and Discussion of our work.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Background

In recent years, the integration of machine learning techniques, particularly self-supervised learning, has played a transformative role in extracting generalized features from sensor data. One notable contribution to this domain is the BERT (Bidirectional Encoder Representations from Transformers) model Devlin et al. 2019, originally designed for natural language processing tasks. BERT's ability to capture contextual information through masked token predictions has inspired advancements across various domains, including computer vision and sensor data analysis. The evolution of transformer architectures, notably exemplified by "Attention is All You Need" Vaswani et al. 2023 has played a pivotal role in advancing Multi-modal learning. Multi-modal Transformers Xu, Zhu, and Clifton 2023 have emerged as a powerful paradigm for capturing relationships across diverse data modalities. In the context of Inertial Measurement Unit (IMU) data, the LIMU-BERT model Xu et al. 2021 represents a significant stride forward. LIMU-BERT focuses on harnessing unlabeled IMU data to extract generalized features, utilizing self-supervised training principles. This approach has proven effective in enhancing the understanding of temporal patterns and contextual relationships within IMU sensor measurements, laying the groundwork for improved analysis and interpretation of inertial data. Extending the principles of LIMU-BERT Xu et al. 2021, our work introduces a Multi-modal perspective, merging IMU data with gaze information to address the challenges associated with missing data in gaze tracking. By leveraging the correlations between head movement and gaze data, our extended model aims to enhance the accuracy of gaze data reconstruction in digital interactions. This approach draws inspiration from the successes of Masked Auto Encoder He et al. 2021 techniques, which share commonalities with BERT in their ability to capture contextual information through masked input predictions. To facilitate our research, we leverage a rich dataset Jin et al. 2022 named "Where Are You Looking?: A Large-Scale Dataset of Head and Gaze Behavior for 360-Degree Videos and a Pilot Study," which encompasses diverse scenarios and interactions, enabling a comprehensive evaluation of our extended Multi-modal Transformer architecture. In synthesizing these elements, our work seeks to contribute to the evolving landscape of self-supervised learning and Multi-modal data analysis, with a particular focus on improving the accuracy and reliability of gaze data reconstruction in digital interaction scenarios.

3 Methodology

In the comparative analysis of gaze prediction models, we evaluated four distinct approaches: Only Gaze Multi-modal Architecture, Head & Gaze Multi-modal Architecture and two baselines Single Modal Architecture and Scipy Interpolation. We utilized the VR Behavior Dataset (Version 2), which comprises approximately 20 million data points of head and gaze tracking information. The following sections provide more clarity on our data preprocessing and model architectures. The Head & Gaze Multi-modal Architecture is only shown as an additional variant and not included in most of our analysis.

3.1 Dataset Preparation

The data includes coordinates, quaternions, and corresponding images captured at a frequency of 120Hz from 100 users across 27 diverse videos. Our preprocessing involved cleaning the data by removing rows with missing values in the Right Gaze Direction and downsampling from 120Hz to 30Hz. This resulted in sequences of 240 data points over 8 seconds, which were then divided into 80% for training, 10% for validation, and 10% for testing, yielding a total of 60926 samples. A compact version about the dataset is provided in Table 1.

Feature	**Description**
Name	VR Behavior (V2)
Modalities	Head and Gaze
Data Representation	Coordinates and quaternions
Sampling Frequency	120Hz among 100 users
Total Samples	60,926

Table 1: Dataset Properties

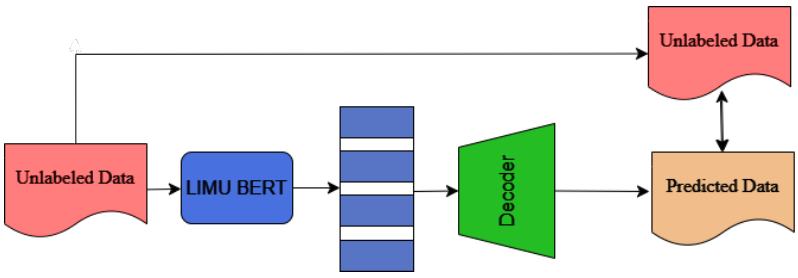


Figure 1: Single Modal LIMU BERT Architecture

3.2 Masking

Our approach introduces a distinct preprocessing methodology tailored for the differential application of masking strategies during the training and testing phases of transformer model pretraining. During training, our method employs a dynamic masking strategy that selects tokens for masking or replacement based on a predetermined probability. This strategy is designed to enhance model robustness by exposing it to a variety of masked inputs, thereby improving its ability to generalize from the training data. Specifically, tokens can be masked with a certain probability or replaced with random tokens, introducing variability in the input data and preventing the model from overfitting to specific patterns. Conversely, in the testing phase, a more uniform masking approach is adopted. Rather than randomly selecting tokens for masking or replacement, a continuous segment of tokens is masked. This methodological shift ensures a consistent and predictable masking pattern during model evaluation, facilitating the assessment of the model's performance in handling masked inputs without the additional variability introduced during training.

This bifurcated approach to masking—employing probabilistic masking and replacement during training and consistent, segment-based masking during testing—allows for a comprehensive evaluation of the transformer model's capabilities in diverse scenarios, highlighting its adaptability and effectiveness across different stages of model development.

3.3 Single Modal Architecture

The Single Modal architecture as shown in Figure 1 is defined as follows: Given an input tensor $X \in \mathbb{R}^{B\times S\times 3}$, where B denotes the batch size, S the sequence length, and 3 the dimensionality of the unit vector representing gaze direction, the embedding layer transforms X into an embedded representation E. The transformer encoder, $\mathcal{T}$, then processes E to obtain an encoded output O, such that $O = \mathcal{T}(E)$. If a set of masked positions $M \subset \{1, \ldots, S\}$ is provided, the encoder's output is filtered to yield $O' = O[:, M, :]$. The activation function GELU (Lee 2023) and a normalization layer $\mathcal{N}$ are subsequently applied to O', resulting in a transformed tensor $T = \mathcal{N}(\text{GELU}(O'))$. Finally, the decoder $\mathcal{D}$ projects T onto the predicted gaze directions, yielding the output tensor $Y \in \mathbb{R}^{B\times|M|\times 3}$, which is subsequently normalized to unit length to obtain the final predictions $\hat{Y} = \frac{Y}{\|Y\|_2}$, ensuring the output is a unit vector for each predicted gaze direction.

3.4 Multi-modal Architecture

The Multi-modal architecture as depicted in Figure 2 incorporates both gaze and head movement data, and is defined as follows: Let $X_g \in \mathbb{R}^{B\times S\times 3}$ and $X_h \in \mathbb{R}^{B\times S\times 3}$ denote the input tensors for gaze and head sequences respectively, where B is the batch size, S is the sequence length, and 3 is the unit vector dimension. The embedding layers transform these inputs into embedded representations E_g and E_h. These representations are concatenated along the sequence dimension to form a combined tensor $E_{gh} = \text{concat}(E_g, E_h)$, which is then passed through the transformer encoder $\mathcal{T}$, yielding the encoded output $O_{gh} = \mathcal{T}(E_{gh})$. If masked positions M are specified, a filtered output $O'_{gh} = O_{gh}[:, M, :]$ is obtained. The activations are applied using the GELU function followed by normalization, $T_{gh} = \mathcal{N}(\text{GELU}(O'_{gh}))$. Decoding is performed by separate decoders for gaze $\mathcal{D}_g$ and head $\mathcal{D}_h$, resulting in predictions $Y_g = \mathcal{D}_g(T_{gh})$ and $Y_h = \mathcal{D}_h(T_{gh})$. These predictions are normalized to unit vectors, $\hat{Y}_g = \frac{Y_g}{\|Y_g\|_2}$ and $\hat{Y}_h = \frac{Y_h}{\|Y_h\|_2}$. Given the predicted output tensor

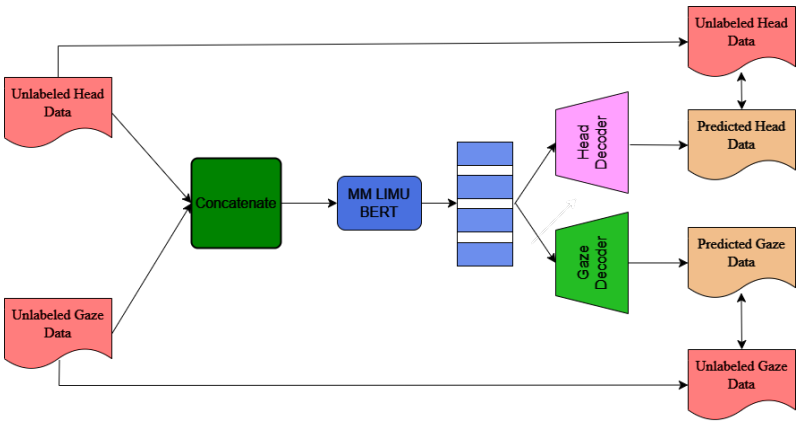


Figure 2: Multi-modal Architecture

$\hat{Y} \in \mathbb{R}^{B\times|M|\times 3}$ and the corresponding ground truth tensor $Y \in \mathbb{R}^{B\times|M|\times 3}$ for the actual masked positions, the MSE loss $\mathcal{L}_{MSE}$ is computed as:

$$\mathcal{L}_{MSE} = \frac{1}{B \cdot |M|} \sum_{i=1}^{B} \sum_{j=1}^{|M|} \|\hat{Y}_{ij} - Y_{ij}\|_2^2 \tag{1}$$

where $\| \cdot \|_2^2$ denotes the squared Euclidean norm. This loss function quantifies the discrepancy between the model's predictions and the actual data, guiding the optimization process. We designed two variations of this architecture depending on whether head sequences are also reconstructed. These variants are named the Only Gaze Multi-Modal Architecture and the Head & Gaze Multi-Modal Architecture. The latter includes an additional decoder specifically for reconstructing head sequences.

4 Results

In this section, we compare the results of different models trained including Gaze Multi-modal and compare the results against the two baselines. We present results visually comparing the actual and predicted sequences of spherical coordinates. We also compare metrics like Euclidean distance, Dynamic Time Warping (DTW) and overall test loss for different models. The DTW metric reflects the model's ability to align predicted sequences with the actual sequences over time, accounting for temporal shifts. The Euclidean Distance metric measures the average spatial distance between predicted and actual gaze points.

4.1 Hyperparameter Tuning

To ascertain the optimal configuration for our gaze prediction models, we undertook a hyperparameter optimization process as visualised in 3. This involved experimenting with a myriad of permutations and combinations of batch sizes, masking ratios, and sequence lengths. Our objective was to identify a set of hyperparameters that would minimize the DTW score, Euclidean Distance and Test Loss. The optimization was executed systematically, ensuring that each variant of the hyperparameters was evaluated under consistent conditions to produce reliable and reproducible results. Through iterative trials and performance evaluations, we determined the hyperparameter set that yielded the best performance metrics, which informed the final model configuration.

4.2 Linear Interpolation using Scipy

The outcomes of applying a linear interpolation method to gaze prediction across various sequences are depicted in Figure 4. The linear interpolation method provides a smooth transition between known data points, evident in the continuity of the predicted lines. This method is inherently limited to the information available from the immediate neighboring points and cannot account for more complex patterns or abrupt changes in data.

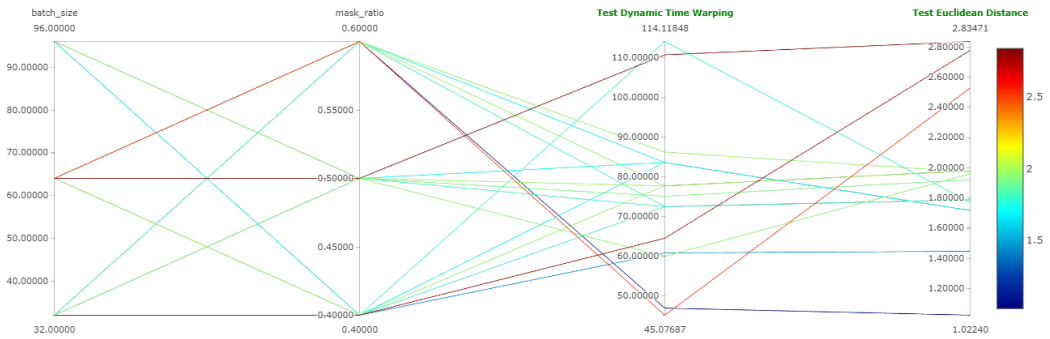


Figure 3: Parallel Coordinates Plot Showing Different Metrics during Hyperparameter Tuning.

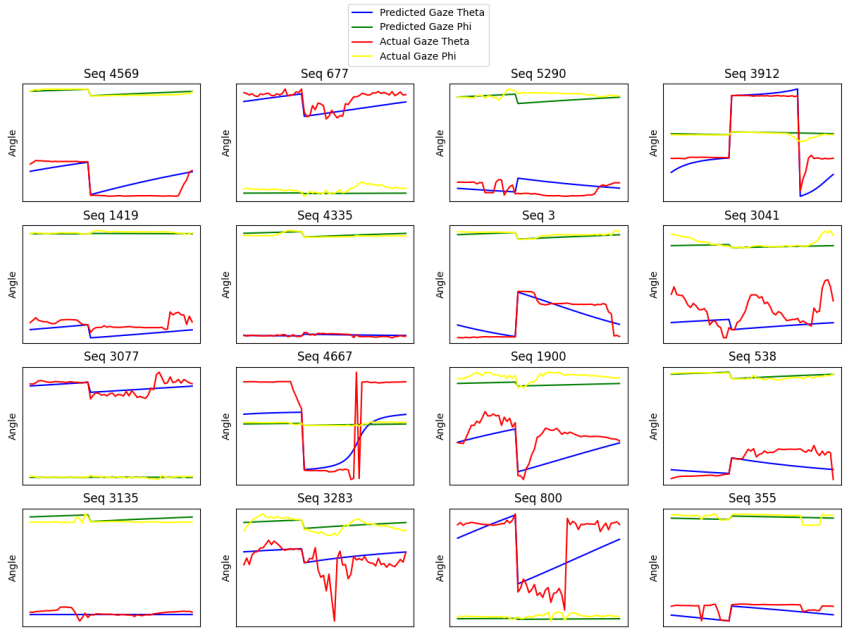


Figure 4: Comparison of Reconstructed and Actual Sequences: Linear Interpolation

4.3 Single Modal Architecture

The output of a Single Modal gaze prediction model across various sequences is showcased in Figure 5. Each subplot corresponds to a unique sequence and plots the predicted versus actual theta and phi angles of gaze over time. The predicted gaze theta and phi are represented by the blue and yellow lines, respectively, while the actual gaze theta and phi are depicted by the red and green lines. In sequences where the Single Modal Architecture closely aligns with the actual data (e.g., Seq 4569, Seq 677), the model appears to perform well. However, discrepancies between prediction and actual data are evident in other sequences (e.g. Seq 538). The Single Modal gaze prediction model provides a valuable baseline for gaze estimation.

4.4 Multi-modal Architecture

An analysis of the Multi-modal architecture's performance in reconstructing gaze angles across various sequences is presented in Figure 6. Each subplot represents a distinct sequence, labeled with a unique identifier (e.g., Seq 4569, Seq 677, etc.), and displays the temporal evolution of two

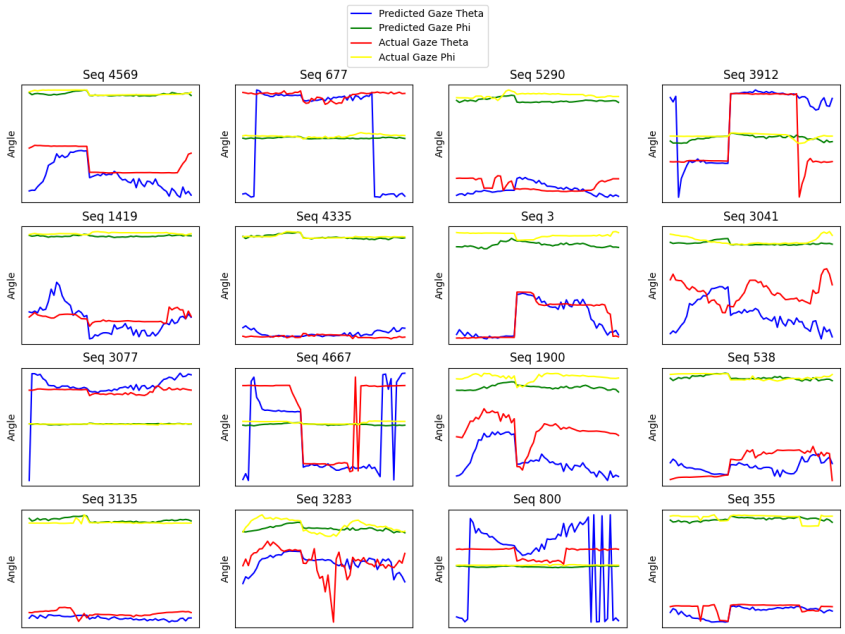


Figure 5: Comparison of Reconstructed and Actual Sequences: Single Modal Architecture

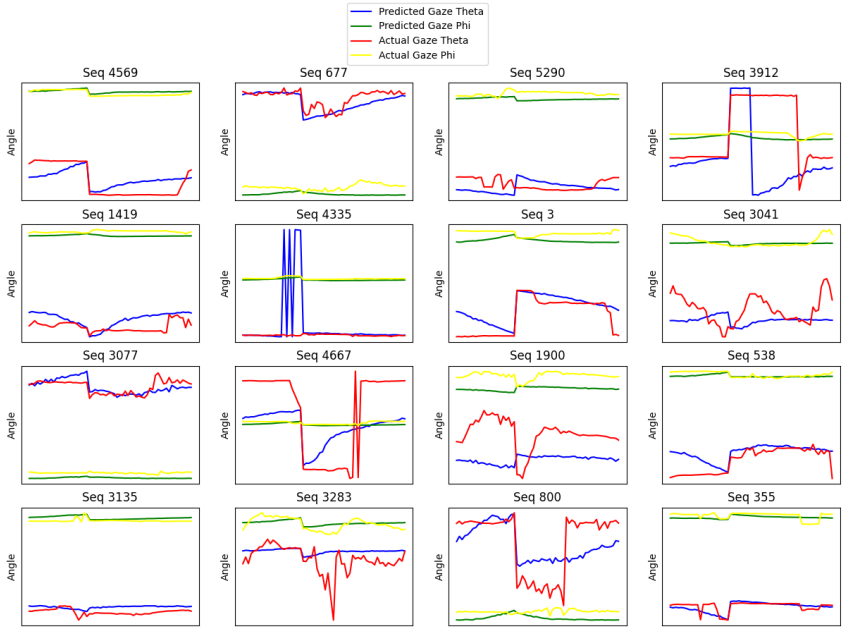
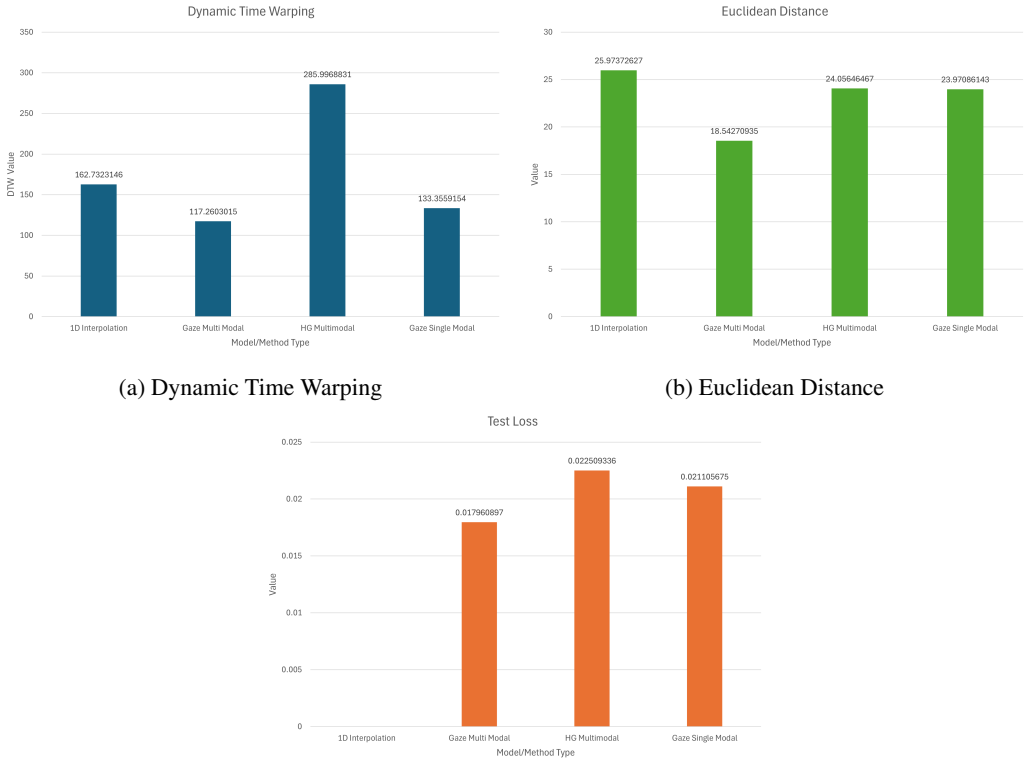


Figure 6: Comparison of Reconstructed and Actual Sequences: Multi-modal Architecture



(a) Dynamic Time Warping

(b) Euclidean Distance

(c) Test Loss

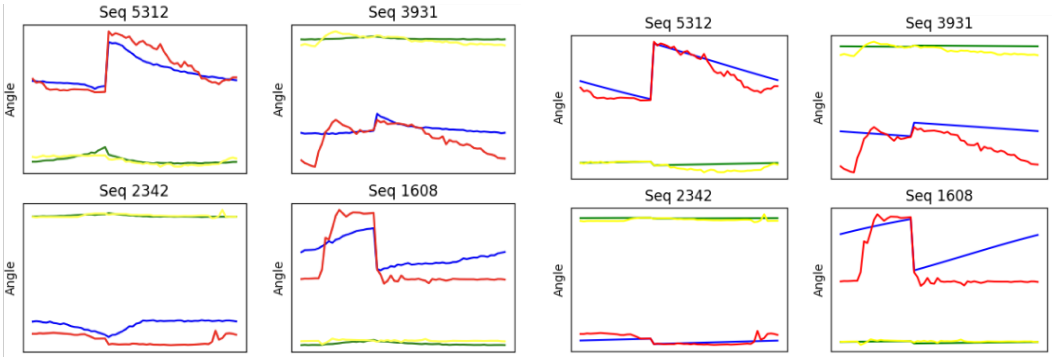
Figure 7: Comparison of metrics between different architectures

angular components: theta and phi. The blue and green lines depict the predicted gaze theta and phi angles, respectively, while the red and yellow lines represent the actual gaze theta and phi angles. The congruence between predicted and actual lines serves as a visual quantifier of the model's accuracy. In several sequences (e.g., Seq 4569, Seq 677), there is a close alignment between the predicted and actual angles, indicating a high degree of predictive accuracy. This suggests that the Multi-modal architecture has successfully captured the underlying patterns in the gaze data, enabling it to predict the gaze orientation with a high level of precision. However, in certain sequences (e.g., Seq 1900, Seq 3912), there are noticeable discrepancies between the predicted and actual lines, particularly in the gaze theta component. These deviations could be attributed to complex temporal dynamics within the data that the model has not fully captured or the methodology used previously mentioned to mask the test sequences. Furthermore, the ability of the architecture to reconstruct missing data can be inferred from the continuity of the predicted lines in the presence of gaps or abrupt changes in the actual data lines. For instance, in Seq 4667, despite sharp transitions in the actual data, the model generates smooth and continuous predictions, suggesting robustness in handling incomplete or noisy input data.

Overall, the visual representation of the model's performance across a diverse set of sequences demonstrates the Multi-modal architecture's capacity to predict gaze orientation with varying degrees of success. The evaluation of the model's performance, particularly in sequences with high prediction accuracy, underscores its potential utility in applications requiring gaze estimation. Conversely, the instances of prediction error provide a critical avenue for further refinement of the model, potentially through incorporating additional modalities or optimizing model parameters.

5 Discussion

The different metrics we used for quantitative analysis of the trained models is shown in Figure 7. The evaluation metrics included Test Loss, DTW, and Euclidean Distance, each offering insights into different aspects of model performance. All models displayed almost equivalent test loss



(a) Multi-modal Architecture (b) Linear Interpolation

Figure 8: Comparison of reconstructed sequences between Multi-modal Architecture and Linear Interpolation

values, indicative of a similar average magnitude of error per sample in the test set. The Gaze Multi-modal approach demonstrated the lowest DTW values in comparison to the Single Modal and Scipy Interpolation methods, indicating a superior temporal alignment with the actual data. The Gaze Multi-modal Architecture again achieved the lowest Euclidean distance, followed by the Single Modal Architecture. The convergence of test loss values across models does not translate to an equivalence in temporal and spatial prediction fidelity, as evidenced by the DTW and Euclidean distance metrics.

When assessing the linear interpolation baseline against the more sophisticated Multi-modal Architecture 8, following comparative conclusions can be drawn. In the presence of gaps or missing data, linear interpolation generates predictions that linearly bridge the gap between known data points. This can be observed in sequences where the interpolated lines remain unbroken despite discontinuities in the actual data. However, this simplistic approach may not accurately reflect the true trajectory of gaze movements, which could be nonlinear or influenced by external factors not captured by the baseline. While linear interpolation is computationally straightforward and robust to noise in small gaps, it lacks the complexity to model dynamic systems accurately. In contrast, Single Modal and Multi-modal Architectures, which employ learning algorithms, can potentially accommodate a wider range of dynamics inherent in gaze movements. Therefore, the baseline results underscore the necessity of more complex models, such as the Single Modal and Multi-modal systems, to capture the nuanced behaviors of gaze patterns. The comparison elucidates the improvements in prediction accuracy and the ability to handle complex temporal dynamics offered by these advanced models over the linear interpolation method.

When comparing this Single Modal Architecture to the Multi-modal Architecture, several observations can be made. The single model may be less robust to variations and noise within the data, as indicated by abrupt changes in the predicted values (e.g., Seq 4667). The Multi-modal Architecture, benefiting from the fusion of diverse inputs, may demonstrate greater resilience to such data inconsistencies, yielding smoother predictions. Both models show the capacity for reconstructing missing data, as suggested by the continuous predicted lines in sequences with gaps or abrupt changes in the actual lines. However, the Multi-modal Architecture may offer a more nuanced reconstruction due to the combined insights from multiple data modalities. The Single Modal Architecture may also be more prone to predicting extreme values or demonstrating erratic behavior in the presence of anomalies within the data (e.g., the spikes in Seq 800). The Multi-modal Architecture's use of additional context may mitigate these effects, leading to more stable predictions. The comparison of these models underscores the potential advantages of a Multi-modal approach in complex predictive tasks.

6 Conclusion and Future Work

In conclusion, the empirical evaluation of our Multi-modal presents a nuanced picture of the efficacy of incorporating head movement data for the reconstruction of occluded gaze data. While the model demonstrates a marked improvement in performance metrics over baseline models in scenarios with

artificially masked gaze data, the real-world applicability and advantages of such an architecture remain to be fully ascertained. The complexity of real-world gaze patterns and the potential variability in missing data may present challenges that were not encapsulated within the controlled conditions of our experimental setup. Future work will be directed towards refining the Multi-modal architecture to enhance its capability for reconstructing gaze data in real-world scenarios. This will include Deploying the reconstructed gaze data in real-world applications, such as user experience research, psychological studies, or human-computer interaction, to assess the tangible benefits of the Multi-modal approach. We also want to explore the synergistic effects of additional modalities, such as environmental context or user interaction data, to potentially enhance the reconstruction accuracy and ensure the model is not only robust to various types of missing data but also generalizes well across different populations and settings.

References

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv: 1810.04805 [cs.CL].

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. *Masked Autoencoders Are Scalable Vision Learners.* arXiv: 2111.06377 [cs.CV].

Jin, Yili, Junhua Liu, Fangxin Wang, and Shuguang Cui. 2022. "Where Are You Looking?: A Large-Scale Dataset of Head and Gaze Behavior for 360-Degree Videos and a Pilot Study." In *MM '22: Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022.* ACM. https://doi.org/10.1145/3503161.3548200.

Lee, Minhyeok. 2023. *GELU Activation Function in Deep Learning: A Comprehensive Mathematical Analysis and Performance.* arXiv: 2305.12073 [cs.LG].

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. *Attention Is All You Need.* arXiv: 1706.03762 [cs.CL].

Xu, Huatao, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. "Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications." In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems,* 220–233.

Xu, Peng, Xiatian Zhu, and David A. Clifton. 2023. *Multimodal Learning with Transformers: A Survey.* arXiv: 2206.06488 [cs.CV].