A Multimodal Approach for Unified Music-Image Embedding Space based on Elicited Emotions

Edoardo Michele Bufi & Angelantonio Fedele Murolo

Email: e.bufi5@studenti.uniba.it a.murolo7@studenti.uniba.it

***Abstract*—We propose to create a unified embedding space between musical and visual data, enabling cross-modal semantic interconnection based on emotional content. To achieve this, we employed the multimodal models Qwen-VL and Qwen-Audio to generate detailed textual descriptions of fragments from image and music datasets, respectively. The obtained textual descriptions were then processed using the Qwen-3 model, which evaluates them and associates a numerical embedding value, explicitly mapping the elicited emotions to the RGB color space. This innovative projection allows us to represent both modalities (music and images) in a common, emotionally-driven vector space. Pairing between musical and visual instances was subsequently performed by calculating one-to-one cosine similarities, identifying the most significant correspondences between the two domains. This approach demonstrates the feasibility of a coherent multimodal embedding guided by emotional context, opening new perspectives for cross-modal retrieval and research applications.**

I. Introduction

The increasing availability of data in diverse modalities, such as images and music, has spurred growing interest in creating unified embedding spaces. Such spaces are crucial not only for understanding individual modalities but also for capturing complex semantic relationships between them, facilitating novel applications in areas like cross-modal information retrieval, content generation, and context analysis.

Our project addresses the significant challenge of integrating visual and auditory representations into a single vector space, crucial for overcoming the limitations of unimodal systems and enabling a more holistic understanding of multimedia content. Unlike traditional methods, this work focuses on utilizing advanced language and multimodal model architectures, such as Qwen-VL and Qwen-Audio, to extract meaningful information from images and music. Subsequently, we explore how a generative language model (Qwen-3) can be employed to interpret these descriptions and project this information into a comparable format, specifically by mapping emotional content to the RGB color space. This novel approach allows for direct comparison and pairing driven by the nuanced emotional understanding derived from the models.

The remainder of this paper is organized as follows: Section II reviews existing literature on multimodal embeddings and the use of Qwen models. Section III details our proposed methodology, including description extraction and the creation of the unified emotion-driven RGB space. Section IV outlines the experimental setup. Section V presents and discusses the results obtained from cosine similarity analysis. Finally, Section VI concludes the paper and suggests future work.

II. Related Work

We provide an overview of existing research relevant to our multimodal embedding approach, covering foundational embedding models, recent advancements in multimodal large language models (MLLMs), and key datasets used in cross-modal learning. Our work builds upon these advancements to establish an emotion-driven connection between visual and auditory modalities.

A. Multimodal Embedding Models

Early work in connecting different modalities often focused on learning shared representations. CLIP (Contrastive Language-Image Pre-training) [1] revolutionized visual-language understanding by training on a massive dataset of image-text pairs, demonstrating remarkable zero-shot transfer capabilities for image classification and retrieval. Extending this paradigm to audio, CLAP (Contrastive Language-Audio Pre-training) [2] learns joint embeddings for audio and text, proving effective for audio classification and retrieval tasks. These models serve as foundational technologies for extracting robust, modality-specific representations, which we then leverage for subsequent emotional mapping.

B. Multimodal Large Language Models for Captioning

The emergence of Large Language Models (LLMs) has led to significant advancements in multimodal understanding, particularly in generating descriptive captions. LLaVA (Large Language and Vision Assistant) [3] and its successor LLaVA-Next [4] are prominent examples of Visual Language Models (VLMs) that integrate LLMs with visual encoders, enabling capabilities like image description, visual question answering, and multimodal dialogue. For our task, the Qwen series of models provides specialized capabilities essential for generating detailed, context-rich captions: Qwen2.5-VL [5] is designed for general visual-language tasks, used here for generating detailed image captions. Similarly, Qwen2.5-Audio [6] excels in audio-text understanding and is employed for generating descriptive captions of musical fragments. While we primarily leverage these specialized Qwen models for detailed captioning—a critical step before our emotion-to-RGB mapping—the broader trend is towards unified models like Qwen2.5-Omni [7], which integrates text, image, and audio capabilities. As an alternative for audio captioning, LP-MusicCaps [8] offers a lightweight solution that processes audio in chunks. Another notable framework for vision-language models is LAVIS [9], although it was not the primary choice for our captioning tasks.

C. Datasets for Cross-Modal Learning

The development of effective multimodal models heavily relies on rich and diverse datasets. For the visual component of our project, we utilize the Artgraph dataset [10], specifically its image collection. This dataset, known for its artistic and expressive content, is particularly suitable for exploring the emotional nuances required for our RGB mapping. For the audio modality, the FMA (Free Music Archive) dataset [11], particularly its smaller version, serves as our primary source of musical fragments due to its diverse genre and emotional range. These datasets provide the necessary raw data from which our Qwen models extract their modality-specific textual descriptions, forming the crucial textual basis for our cross-modal emotional embedding approach.

III. Methodology

Our proposed approach integrates multimodal perception, large language model-based emotional mapping, and similarity computation to establish cross-modal connections between visual and auditory data. The overall process is illustrated in Figure 1.

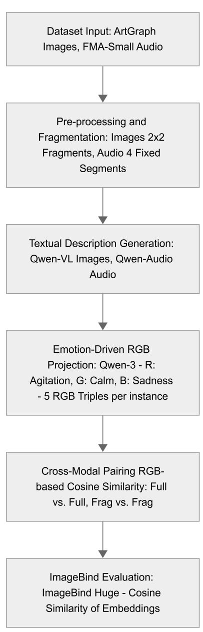

Fig. 1: Overall Methodology Pipeline for Emotion-Driven Cross-Modal Embedding and Evaluation.

A. Description Generation with Qwen-VL and Qwen-Audio

The initial step involves translating raw image and audio content into rich textual descriptions. For audio, we utilized the Qwen2-Audio-7B-Instruct model from Hugging Face Transformers. Each full audio track from the FMA dataset was processed in two distinct ways: first, the complete track was analyzed, and concurrently, it was programmatically split into four fixed-duration segments. Both these individual segments and the complete track were then fed to the Qwen2-Audio model for detailed textual analysis. Audio data was consistently resampled to a target sample rate of 16kHz to ensure compatibility with the model's input requirements.

A meticulously designed system prompt and user instruction guided the model's output to ensure descriptive richness and focus on relevant attributes. The system prompt established the model's persona as a "highly specialized and descriptive music analyst," instructing it to provide comprehensive descriptions covering musical elements such as instruments, tempo/rhythm, timbre, and, crucially, the emotions evoked, along with an imagined contextual setting. The user prompt explicitly reinforced these requirements and specified whether the analysis pertained to a particular fragment or the entire song. The max_new_tokens parameter for generated responses was capped at 512 tokens to control output verbosity and ensure efficient processing.

Similar principles of fragmentation and detailed prompting were applied for image description generation using the Qwen2.5-VL model. Each image from the Artgraph dataset was similarly processed as a whole and divided into four distinct fragments. For each (full image + four fragments), Qwen2.5-VL generated a textual description following a specific prompt designed to elicit information about visual elements, colors, composition, and their emotional impact. This dual-modal captioning process ensures a rich textual foundation for the subsequent emotional mapping.

B. Emotion-Driven RGB Space Projection with Qwen-3

Following the generation of detailed textual descriptions for both music and image (full and fragmented) instances, the next critical step involves projecting these descriptions into a unified, emotionally-driven embedding space. For this purpose, we employed the Qwen3-4B large language model [12]. Qwen-3 was specifically tasked with evaluating the emotional content of each textual description and mapping it to a 3-dimensional RGB color value.

To achieve this, the model was instructed to quantify the perceived presence of three core emotional axes, mapping them to the primary color channels: 'excitement/movement' was assigned to the Red channel, 'calmness/stillness' to the Green channel, and 'sadness/nostalgia' to the Blue channel. Values for each channel were constrained to the standard 0 to 255 range. The prompt engineering for Qwen-3 was meticulously designed to ensure the output adhered to a strict, machine-readable format: specifically, five complete RGB triples (corresponding to the four audio/image fragments and one total description), all on a single line, prefixed by 'RGB:'. This strict format facilitated automated parsing of the model's output. The extracted RGB values were then programmatically clamped to the valid 0-255 range to prevent overflow and ensure consistency. This RGB representation serves as a compact and visually intuitive encoding of the elicited emotional state, allowing for direct comparison between modalities based on a shared emotional interpretation.

$$\text{Embedding}_{\text{RGB}} = \text{Qwen-3}_{\text{EmotionMapper}}(\text{Textual Description}) \tag{1}$$

Where $\text{Qwen-3}_{\text{EmotionMapper}}$ acts as a mapping function from semantically and emotionally rich textual descriptions to a 3-dimensional RGB vector.

C. Cross-Modal Pairing using Cosine Similarity

Once the emotion-driven RGB embedding vectors were generated for all full and fragmented images and audio tracks, the next stage involved pairing. For each full audio track, we aimed to find the most emotionally aligned image from our dataset. This pairing was performed by calculating the cosine similarity between the RGB embedding of the full audio track and the RGB embeddings of all full images in the dataset. The image yielding the highest cosine similarity with a given audio track was selected as its most suitable pair, indicating the strongest emotional alignment in the RGB space.

We performed a similar pairing process for the fragmented data. For each audio fragment, its RGB embedding was compared against the RGB embeddings of all corresponding image fragments (i.e., fragment 1 of audio vs. fragment 1 of all images, and so on). This allowed for a more granular, localized emotional matching.

The cosine similarity metric was chosen due to its effectiveness in measuring the angular distance between two non-zero vectors in a multi-dimensional space, making it ideal for quantifying the degree of emotional alignment between our RGB representations. The formula for cosine similarity is given by:

$$\text{Similarity}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \tag{2}$$

Where $\mathbf{v}_1$ and $\mathbf{v}_2$ are the emotion-driven RGB embedding vectors for a music and an image instance (or fragment), respectively. The output of this stage is a set of 1:1 pairings, forming the input for the final validation step using ImageBind [13].

IV. Experiments

We outline the datasets used, the distinct metrics applied for evaluation, and the overall setup of our cross-modal analysis pipeline.

A. Datasets

Our evaluation utilizes balanced subsets from two multi-modal sources:

- **ArtGraph Visual Subset**: 100 curated artistic images from Zenodo's collection (average original resolution 1200×800). Before processing with Qwen-VL, images underwent center-cropping and resizing to 448×448. For analysis, each image was programmatically divided into a 2x2 grid, yielding four distinct fragments, in addition to the analysis of the entire image. We also leveraged the associated ‘artgraph_metadata.parquet‘ for data handling and organization.
- **FMA-Small Audio Subset**: 100 diverse music tracks (full-length) from Kaggle. Each full audio track was loaded and then programmatically split into four fixed-duration segments. Both these individual segments and the complete track were then processed. Audio data was consistently resampled to 16000 Hz before being fed to the model.

For both subsets, Qwen's multimodal capabilities were used to generate rich textual descriptions. These descriptions inherently include aspects such as ”emotions evoked”. No dataset-specific preprocessing was applied beyond format standardization and the aforementioned fragmentation/resampling.

B. Evaluation Metrics

To assess the effectiveness of the cross-modal pairing, two primary evaluation approaches were employed, focusing on both semantic and emotional consistency:

- **ImageBind Similarity Score:** This metric quantifies the semantic alignment between paired image and audio content. It is calculated as the cosine similarity between the multimodal embeddings generated by the ImageBind model for each image-audio pair. The distribution of these similarity scores is analyzed to understand the overall alignment.
- **Emotional Coherence:** Metrics assessing how well the paired RGB values (derived from Qwen's emotion-driven textual descriptions) align between modalities. This evaluates the consistency of the implied emotional content. This approach also involves analyzing the distribution of these RGB-based similarity measures.
- **Qualitative Assessment:** Human evaluation of selected best and worst pairings, specifically judging their emotional congruence and overall multimodal coherence. This provides crucial insights beyond quantitative measures.

C. Experimental Setup

The experimental workflow involved three main stages: multimodal description generation, emotion-driven RGB projection, and cross-modal pairing/evaluation.

Initially, for multimodal description generation, the Qwen2-Audio-7B-Instruct model was used for audio analysis and Qwen2.5-VL for image analysis. These models generated detailed textual descriptions for both the four fragments and the entire instance of each image and audio file. The descriptions captured relevant content features and evoked emotions, serving as the raw input for the subsequent stage.

Next, for emotion-driven RGB projection, the Qwen3-4B large language model was employed. This model processed the textual descriptions from the previous stage, mapping their emotional content to a 3-dimensional RGB color value. Specifically, 'excitement/movement' was mapped to the Red channel, 'quietness/stillness' to Green, and 'sadness/nostalgia' to Blue, with values constrained from 0 to 255. Prompt engineering ensured that Qwen-3 outputted precisely five complete RGB triples (corresponding to the four fragments and the total description) per instance, facilitating automated parsing. The generated RGB values were then compiled into separate CSV files for audio and images.

Finally, for cross-modal pairing and evaluation, the ImageBind ‘imagebind_huge‘ model was utilized. This model generated multimodal embeddings for the 100 original images and 100 audio tracks. Cosine similarity was then calculated between the ImageBind embeddings of predefined 1:1 image-audio pairs, identified from a classification CSV. The resulting ImageBind scores, representing semantic alignment, were saved for further analysis and used to identify best and worst matches.

V. Results and Discussion

We presents the numerical and qualitative findings of our multimodal embedding and pairing experiments, analyzes the results, and discusses their implications, particularly focusing on the role of emotion-driven embeddings.

A. Quantitative Results

The quantitative outcomes, including comprehensive comparative statistics on the generated textual descriptions, the intra-modal coherence assessed via cosine similarities, the cross-modal RGB pairing effectiveness, and ImageBind scores.

1) Description Length Statistics Comparison: To characterize the output of the Qwen models, we analyzed the length of the generated textual descriptions for both full instances and aggregated fragments. Table I compares description length statistics for total audios versus aggregated audio fragments generated by Qwen-Audio. Similarly, Table II provides a comparison for total images versus aggregated image fragments generated by Qwen-VL.

TABLE I: Comparison of Qwen-Audio Description Lengths (Words).

Metric	Total Audios	Agg. Audio Fragments
Max Description Length	265	436
Min Description Length	38	68
Mean Description Length	83.56	105.90
Descriptions > Mean Length	44	36
Descriptions < Mean Length	56	64
Descriptions = Mean Length	0	0
Total Descriptions Count	100	100
Complete Descriptions	0	0
Incomplete Descriptions	100	100

2) Intra-Modal Cosine Similarities Comparison: We evaluated the coherence between the original modality and its Qwen-generated textual description using established embedding models (CLAP for audio, CLIP for images). Table III compares CLAP cosine similarities for total audios versus aggregated audio fragments. Table IV provides a similar comparison for CLIP cosine similarities, focusing on total images versus aggregated image fragments.

TABLE II: Comparison of Qwen-VL Image Description Lengths (Words).

Metric	Total Images	Agg. Image Fragments
Max Description Length	430	816
Min Description Length	273	369
Mean Description Length	388.63	492.87
Descriptions > Mean Length	66	36
Descriptions < Mean Length	34	64
Descriptions = Mean Length	0	0
Total Descriptions Count	100	100
Complete Descriptions	46	100
Incomplete Descriptions	54	0

TABLE III: Comparison of CLAP Cosine Similarities (Audio: Modality vs. Description).

Metric	Total Audios	Aggregated Audio Fragments
Number of Pairs	100	100
Mean	0.3051	0.2789
Median	0.2970	0.2870
Standard Deviation	0.0875	0.1134
Min	0.0597	−0.0448
Max	0.5060	0.4834

3) Cross-Modal Emotion-Driven RGB Pairing: The core quantitative result of our methodology is the evaluation of cross-modal pairings based on the emotion-driven RGB embedding space. Table V presents the distribution of cosine similarities between the RGB embeddings of the paired music and image instances, which enabled the pairing process.

4) ImageBind Score Statistics: For additional cross-modal evaluation, ImageBind scores were computed between the image and audio embeddings. Table VI summarizes the statistics of these scores, providing another perspective on the relationships between modalities.

TABLE IV: Comparison of CLIP Cosine Similarities (Image: Modality vs. Description).

Metric	Total Images	Aggregated Image Fragments
Number of Pairs	100	100
Mean	0.3101	0.3039
Median	0.3149	0.3017
Standard Deviation	0.0356	0.0379
Min	0.2113	0.1879
Max	0.3998	0.3911

TABLE V: Distribution of Cosine Similarities for Emotion-Driven Paired Music-Image Embeddings.

Metric	Value
Total Songs	100
Total Images	100
Pairings Made	100
Unpaired Songs	0
Unpaired Images	0
Average Similarity	0.9214
Maximum Similarity	0.9987
Minimum Similarity	0.4016

TABLE VI: ImageBind Score Statistics for Paired Music-Image Embeddings.

Metric	Value
Number of Pairs Processed	100
Skipped Pairs	0
Mean ImageBind Score	0.0366
Median ImageBind Score	0.0289
Standard Deviation ImageBind Score	0.0597
Min ImageBind Score	−0.0859
Max ImageBind Score	0.2045

B. Qualitative Analysis

We present visual examples of pairings, highlighting the effectiveness of the emotion-driven RGB mapping. We specifically showcase the top-performing pairings based on cosine similarity, linking them to their emotional content and RGB mappings.

Fig. 2: Paired Example 1: Top-performing pairing with Similarity 0.9987. Image: *Lucian Freud naked man with his friend.jpg*. Audio: *000211.mp3* (Link to Audio 1).

Fig. 3: Paired Example 2: Top-performing pairing with Similarity 0.9987. Image: *Juan-Gris the open book 1925.jpg*. Audio: *005157.mp3* (Link to Audio 2).

C. Discussion of Findings

We analyze the quantitative and qualitative results, focusing on the efficacy of emotion-driven RGB mapping in establishing a unified embedding space for cross-modal pairing.

The core finding of our analysis is the remarkable effectiveness of the emotion-driven RGB mapping in creating a coherent and unified embedding space for cross-modal pairing between music and images. The quantitative results, specifically the high average cosine similarity of 0.9214 between paired music and image instances in the RGB space (Table V), provide strong evidence for this. With a near-perfect maximum similarity of 0.9987 and the successful pairing of all 100 songs and 100 images without any mismatches, this approach underscores its robustness. This suggests that the RGB representation effectively captures and aligns the underlying emotional content across disparate modalities, facilitating highly effective and emotionally congruent pairings.

Regarding the Qwen models' performance in generating textual descriptions, variations were observed across modalities. Qwen-Audio, when processing entire audio tracks, consistently produced descriptions marked as "Incomplete" (100 out of 100), despite achieving a mean length of 83.56 words (Table I). This indicates a potential limitation in comprehensively summarizing extensive temporal data into a single, complete textual representation. In contrast, Qwen-VL performed more robustly for images, yielding "Complete" descriptions for 46% of total images and all aggregated fragments (Table II). However, despite these differences in description completeness, the subsequent emotion-driven RGB mapping proved highly effective.

A crucial insight emerges from the comparison between intra-modal coherence and cross-modal RGB similarities. The CLAP cosine similarities for audio-description pairs (mean 0.3051) and CLIP cosine similarities for image-description pairs (mean 0.3101) were relatively low (Tables III and IV). This suggests that while Qwen models generate descriptive text, the direct semantic alignment of these descriptions with raw modality embeddings (as measured by CLAP and CLIP) might not always be strong. However, the significant disparity between these lower intra-modal similarities and the remarkably high cross-modal RGB similarities is key. It strongly implies that the emotion-driven RGB layer acts as a powerful bridge, effectively prioritizing and extracting the emotional essence from the Qwen-generated descriptions to facilitate accurate cross-modal pairing, proving more effective for this specific task than relying solely on general semantic alignment.

Furthermore, a comparative perspective is offered by the ImageBind scores (Table VI), which directly assess cross-modal coherence between raw image and audio embeddings. With a mean ImageBind score of 0.0366, these scores suggest a weaker inherent alignment in a general-purpose multimodal space compared to the high similarities achieved through our emotion-driven RGB approach. This further emphasizes the unique and effective contribution of our methodology in creating a highly aligned space specifically through emotional congruence. The qualitative examples, such as the pairings with cosine similarity of 0.9987 (Figure 2 and Figure 3), visually corroborate the strong emotional alignment achieved by our method. These high-similarity pairings exemplify cases where the emotion-driven RGB embeddings successfully matched music and images sharing a clear and strong common emotional valence, aligning with the growing interest in affective computing within multimodal AI. Even for less optimal pairings, the minimum RGB similarity of 0.4016 indicates that a relevant emotional connection was consistently maintained, preventing entirely incongruous matches and highlighting the robustness of the emotional layer.

VI. Conclusion

We successfully addressed the challenge of cross-modal music-image pairing by developing an emotion-driven methodology. By utilizing Qwen models for textual descriptions and mapping these into an RGB emotion-driven color space, we created a unified, emotionally-aware embedding space.

The key finding is the remarkably high average cosine similarity of 0.9214 in the RGB space, with all 100 music and image instances successfully paired (Table V). This robust alignment, achieved despite relatively lower intra-modal similarities (Tables III and IV), underscores the RGB layer's effectiveness as a crucial emotional bridge for cross-modal content. Our work thus significantly contributes to creating a highly coherent and emotionally congruent space between music and images.

Future work could involve exploring alternative emotion representation models beyond RGB, integrating a wider range of emotional nuances, or developing more sophisticated evaluation metrics for cross-modal pairing quality, particularly for emotional congruence, in the absence of explicit emotional ground truth.

VII. Acknowledgments

This work was proposed by PhD student Ivan Rinaldi and Prof. Giovanna Castellano at the University of Bari, as part of the Master's Degree program in Computer Science: Artificial Intelligence.

References

[1] A. Radford *et al.*, "Clip: Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020.

[2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," *arXiv preprint arXiv:2206.04769*, 2022. [Online]. Available: https://arxiv.org/abs/2206.04769.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Fu, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08485.

[4] H. Liu *et al.*, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024. [Online]. Available: https://arxiv.org/abs/2407.07895.

[5] Q. Team, "Qwen2.5-vl: The new vision language model from qwen team," *arXiv preprint arXiv:2502.13923*, 2025. [Online]. Available: https://arxiv.org/abs/2502.13923.

[6] Y. Chen *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024. [Online]. Available: https://arxiv.org/abs/2407.10759.

[7] Q. Team, "Qwen2.5-omni: Towards a general-purpose multimodal large language model," *arXiv preprint arXiv:2503.20215*, 2025. [Online]. Available: https://arxiv.org/abs/2503.20215.

[8] S. Doh, G.-s. Jung, M. Lee, H.-W. Kim, and S.-G. Lee, "Lp-musiccaps: Boosting music captioning with a large pool of pseudo-captions," *arXiv preprint arXiv:2402.09117*, 2024. [Online]. Available: https://arxiv.org/abs/2402.09117.

[9] J. Li, D. Li, C. Savarese, and S. C. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *arXiv preprint arXiv:2201.1072*, 2022. [Online]. Available: https://arxiv.org/abs/2201.1072.

[10] G. Castellano, V. Digeno, G. Sansaro, R. Scaringi, and G. Vessio, *Artgraph*, 2023. DOI: 10.5281/zenodo.8172374. [Online]. Available: https://zenodo.org/records/8172374.

[11] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. [Online]. Available: https://arxiv.org/abs/1612.05971.

[12] Q. Team, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025. [Online]. Available: https://arxiv.org/abs/2505.09388.

[13] R. Girdhar *et al.*, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20268–20278. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html.