applied sciences

MDPI

Article

EmoStyle: Emotion-Aware Semantic Image Manipulation with Audio Guidance

Qiwei Shen [1], Junjie Xu [2], Jiahao Mei [2], Xingjiao Wu [3] and Daoguo Dong [2,*]

[1] Software Engineering Institute, East China Normal University, Shanghai 200062, China; 10215101495@stu.ecnu.edu.cn
[2] School of Computer Science and Technology, East China Normal University, Shanghai 200062, China; jjxu_dr@stu.ecnu.edu.cn (J.X.); 10215102440@stu.ecnu.edu.cn (J.M.)
[3] School of Computer Science, Fudan University, Shanghai 200433, China; xjwu_cs@fudan.edu.cn
* Correspondence: dgdong@cs.ecnu.edu.cn

Abstract: With the flourishing development of generative models, image manipulation is receiving increasing attention. Rather than text modality, several elegant designs have delved into leveraging audio to manipulate images. However, existing methodologies mainly focus on image generation conditional on semantic alignment, ignoring the vivid affective information depicted in the audio. We propose an Emotion-aware StyleGAN Manipulator (EmoStyle), a framework where affective information from audio can be explicitly extracted and further utilized during image manipulation. Specifically, we first leverage the multi-modality model ImageBind for initial cross-modal retrieval between images and music, and select the music-related image for further manipulation. Simultaneously, by extracting sentiment polarity from the lyrics of the audio, we generate an emotionally rich auxiliary music branch to accentuate the affective information. We then leverage pre-trained encoders to encode audio and the audio-related image into the same embedding space. With the aligned embeddings, we manipulate the image via a direct latent optimization method. We conduct objective and subjective evaluations on the generated images, and our results show that our framework is capable of generating images with specified human emotions conveyed in the audio.

Keywords: generative model; image manipulation; affective information; audio-based image manipulation

Citation: Shen, Q.; Xu, J.; Mei, J.; Wu, X.; Dong, D. EmoStyle: Emotion-Aware Semantic Image Manipulation with Audio Guidance. *Appl. Sci.* **2024**, *14*, 3193. https://doi.org/10.3390/app14083193

Academic Editor: Andrea Prati

Received: 31 January 2024
Revised: 2 April 2024
Accepted: 8 April 2024
Published: 10 April 2024

1. Introduction

With the skyrocketing popularity of a series of generative models, such as GPT4, LLaMA [1], StyleGAN [2], and Stable Diffusion [3], generative artificial intelligence has garnered significant attention in academia.

Cross-modal generation, one of the most crucial tasks of generative AI, has received great attention for bridging the gap between different modalities. Image manipulation, a sub-task of cross-modal generation, has been extensively studied for its wide applications in digital social media. Image manipulation aims at editing images according to user intention. Recent work has innovated various kinds of methods for manipulating images through human language, such as semantic labels, text, or scene graphs [4]. Meanwhile, some researchers attempt to leverage audio to manipulate the generated images. An emerging line of work has sought to create multi-modal systems by leveraging contrastive learning mechanisms [5]. Among these methods, drawing inspiration from the text-based image manipulation method StyleCLIP [6], which leverages the representational capabilities of the contrastive language-image pretraining (CLIP) [7] model, Lee et al. [8] extended this embedding space to encompass both audio-visual and text modalities.

However, the methods of image manipulation based on audio information, particularly those utilizing music hybrid information, exhibit certain inherent drawbacks. A primary limitation of existing approaches is their focus on leveraging semantic elements derived from music, such as genre or rhythm, while largely neglecting the rich emotional landscape

that music, especially complex compositions, can portray. Music has the capacity to elicit a broad spectrum of emotions that extend beyond its semantic attributes. By overlooking these emotional nuances, the resultant images often do not capture the full depth and nuance of the emotions that the music aims to express. Consequently, the images manipulated in this manner struggle to align effectively with the music, leading to potential inconsistencies and instabilities in the image guidance effects. Furthermore, the guidance process in current methodologies typically relies on semantic information extracted from the music for singular-dimensional guidance, neglecting the application of multi-dimensional constraints. This oversight can result in generation outcomes that lack stability and fidelity to the multi-faceted nature of the music's emotional and semantic content.

To overcome the identified challenges, we introduce a pioneering framework entitled the Emotion-aware StyleGAN Manipulator (EmoStyle). Initially, we deploy the ImageBind [9] model to identify and retrieve images that are semantically aligned with the provided musical content, earmarking these images for further manipulation. Next, we extract the lyrical content from the music and conduct sentiment analysis to classify the emotional tone of the lyrics. Utilizing the insights gained from both the lyrics and their associated sentiment polarity, we craft an emotionally resonant instrumental music segment, intended to complement the primary musical content as auxiliary emotional music. In the final step, we employ pre-trained encoders to transform the musical and textual content alongside the retrieved images into a unified feature space. This convergence facilitates the execution of a hybrid latent vector alignment technique for precise and emotionally attuned image manipulation.

In a nutshell, our contributions can be summarized as follows:

- We introduce a sentiment-auxiliary semantic-guided image manipulation framework, termed the Emotion-aware StyleGAN Manipulator (EmoStyle), which efficiently leverages the affective information from music to manipulate images.
- We devise a simple yet effective retrieval approach that selects images with minimal semantic gaps based on the similarity between music and images, which ensures the stability of subsequent image manipulation.
- We introduce an innovative sentiment-assisted guidance module that explicitly extracts sentiment polarity from the music, generates an affective music branch, and aids in image manipulation with a focus on the sentimental dimension.
- We evaluate the proposed EmoStyle with objective methods, confirming the effectiveness of our approach.

2. Related Work

2.1. Cross-Modal Generation

In the realm of cross-modal generation, the task of translating information from one modality to another presents a fascinating and open field of research. Various tasks across diverse domains have been explored, such as text-to-image generation [10], where textual descriptions are transformed into visual images. Similarly, text-to-video generation [11] delves into creating video sequences from textual input. In the domain of image processing, the image-to-caption task, as demonstrated in works like Flamingo [12], focuses on generating textual descriptions from visual inputs. Likewise, the conversion of audio to textual captions, as explored in ClipCap [13], showcases the potential of translating auditory information into text. A significant trend in these cross-modal endeavors is the leveraging of existing pre-trained models. Notably, many studies have extended the pre-trained CLIP [7] embedding space, which is anchored in the text-visual modality, to bridge heterogeneous modalities in cross-modal generations. In line with this trend, our work focuses on the task of generating manipulated images from audio by leveraging text-audio signals extracted from the audio, highlighting the multifaceted nature of cross-modal generative tasks.

2.2. Text-Driven Image Manipulation

With the popularity of text-driven image generation [14–18], text-driven image manipulation [19–25] has achieved substantial developments. This emerging field centers on modifying images with textual descriptions that specify desired visual characteristics, such as altering the age, emotion, or atmospheric conditions in an image. Early explorations, including those by Dong et al. [19], Li et al. [21], and Nam et al. [23], leveraged the encoder–decoder architecture inherent in generative adversarial networks (GANs). These studies successfully demonstrated how images could be manipulated while preserving their fundamental attributes, setting a foundational precedent for subsequent research. Further progression in this field saw the integration of these methods with sophisticated architectures like StyleGAN [2]. Notably, Xia et al. [24] introduced a pioneering approach that incorporated a visual-linguistic similarity module within the StyleGAN framework. This innovation significantly improved the alignment between text and image modalities, enhancing the accuracy and relevance of the resultant image manipulations.

A remarkable evolution in this area is the incorporation of pre-trained models' latent spaces, such as StyleGAN, with advanced language-image understanding models like CLIP [7]. The work of StyleCLIP [6] exemplifies this approach, wherein the intuitive understanding of the text by CLIP is utilized to guide and refine the manipulation direction within StyleGAN's latent space. This synergy represents a paradigm shift, offering a more nuanced and precise method for manipulating images based on textual descriptions.

2.3. Audio-Visual Learning

Audio-visual learning is a sophisticated computational endeavor aimed at aligning audio and visual modalities within a unified latent embedding space. The primary goal of this task is to minimize the distance between corresponding audio and visual pairs while simultaneously maximizing the divergence between non-corresponding pairs. Many previous studies [26–32] have been conducted to delve into the realm of audio-visual alignment. This cross-modal alignment holds significant potential for other related audio-based tasks, such as audio/speech separation [29,33–35], audio-visual event parsing [36–39], and audio spatialization [31,40,41]. In this work, our main focus is to learn compact audio-visual representations on audio-guided image manipulation. By leveraging the strong multi-modal representation capacity of CLIP, Lee et al. [8] trained an audio encoder to embed the given audio into the embedding space of CLIP.

Audio-visual learning, a complex and burgeoning field in computational research, revolves around the challenge of synchronizing audio and visual modalities within a unified latent embedding space. The core objective in this discipline is to intricately align corresponding audio and visual pairs, thereby minimizing the distance between them. Concurrently, it is crucial to ensure a substantial separation between non-corresponding pairs, maintaining a clear distinction between unrelated audio and visual elements.

The pursuit of audio-visual alignment has seen substantial research [26–32], with significant contributions from a variety of studies. This line of research has greatly influenced subsequent studies in related audio-visual tasks, such as audio/speech separation [29,33–35], audio-visual event parsing [36–39], and audio spatialization [31,40,41].

Our present work seeks to advance this field by focusing on the development of compact and effective audio-visual representations specifically tailored for audio-guided image manipulation. We draw inspiration from and build upon the multi-modal representation capacity of CLIP. Following the methodology of Lee et al. [8], our approach involves training an audio encoder designed to integrate audio inputs seamlessly into the CLIP embedding space. This novel approach not only aligns with the fundamental principles of audio-visual learning but also pushes the boundaries of how audio can influence and guide visual content creation.

2.4. Audio-Driven Image Manipulation

In contemporary cross-modal generation research, the focus has increasingly shifted towards more nuanced methodologies, such as manipulating images using combined audio-text input rather than direct audio-to-image generation. A notable example of this approach is the work of Lee et al. [8], who adeptly utilized text-audio synergy. Their strategy involved integrating audio semantic information into a shared latent space aligned with the text modality. This integration was further enhanced by modifying the latent code of an image using StyleGAN2 [42], thereby creating a sophisticated multi-modal alignment model.

Our research methodology, however, diverges in two fundamental ways from Lee et al.s' approach. First, we place a heightened emphasis on extracting and utilizing affective information from the audio. This involves not just capturing the semantic content but also interpreting and translating the emotional undertones and nuances inherent in the audio into visual elements. By doing so, our method aims to create images that resonate more deeply with the emotional context of the audio source. Second, we employ a retrieval-based method to mitigate the semantic disparity between the audio input and the visual output. This method is crucial in ensuring the stability of the generated images. By narrowing the semantic gap, our approach facilitates a more accurate and coherent translation of the audio content into the visual domain. This retrieval-based strategy is instrumental in maintaining the integrity and relevance of the generated images to the original audio input.

Overall, our approach represents a refined and innovative direction in audio-influenced image manipulation within the field of cross-modal generation. By focusing on affective information and employing a retrieval-based method, we aim to enhance the fidelity and emotional congruence of the images generated from audio inputs.

3. Methodology

In this section of the paper, we present a detailed exposition of our proposed framework, EmoStyle, as illustrated in Figure 1. Initially, we discuss the semantic similarity-based image retrieval module, designed to retrieve images aligned with semantic congruity. Following this, the affective music generation module is introduced, responsible for the composition of music enriched with emotional depth. Concluding our framework's description, we elaborate on the sentiment-assisted guidance module, an innovative component leveraging sentiment analysis to enhance the output's relevance and emotional resonance. Collectively, these modules form the cornerstone of the EmoStyle framework, aimed at augmenting the quality and emotional pertinence of content generation through an integrated approach.

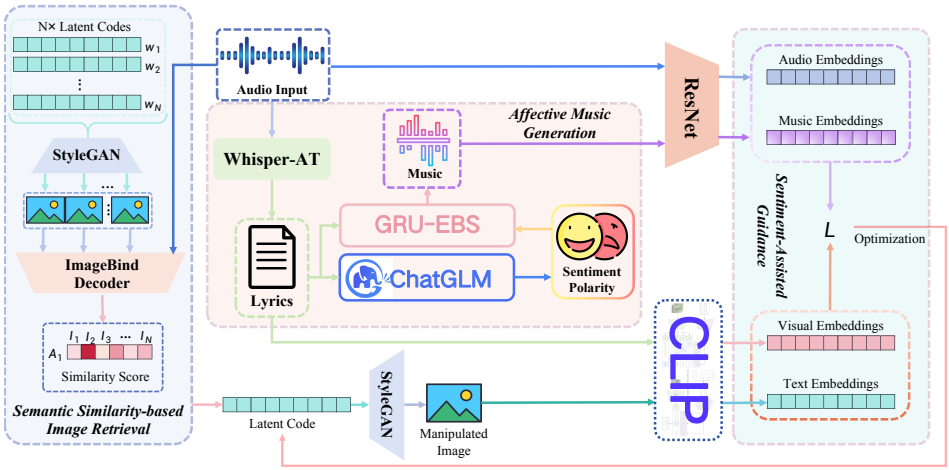

Figure 1. The overall architecture of the proposed EmoStyle. EmoStyle consists of three modules: semantic similarity-based image retrieval, affective music generation, and sentiment-assisted guidance.

3.1. Semantic Similarity-Based Image Retrieval

For effective image manipulation stability, it is imperative to diminish the semantic discrepancy between the input audio and the resultant image. Cross-modality models play a pivotal role in bridging the gap between auditory and visual modalities, thanks to their capability to encode content from disparate modalities into a unified latent space. Initially, we generate N random latent codes. For each latent code, w, an image, $G(w)$, is synthesized using StyleGAN2, denoted by G. With the input audio, A, and the N randomly generated images, $G(w)$, at our disposal, we first utilize ImageBind [9] for the dual encoding of both the audio and images. This process ensures a more cohesive integration of auditory and visual information, facilitating a seamless cross-modal translation that narrows the semantic gap between the given audio input and the generated visual output:

$$\begin{aligned} E_a &= AudioEncoder_{ImageBind}(A) \\ E_{img} &= ImageEncoder_{ImageBind}(G(w)) \end{aligned} \tag{1}$$

where E_a and E_{img} stand for the embeddings of the audio and image. Then, we calculate the similarity between the two modalities' embeddings:

$$s_i = E_{img} \cdot E_a^T \tag{2}$$

where s_i denotes the similarity of the ith image with the audio. After applying a $softmax$ layer, we select the latent code, w, which generates the image with the highest similarity score as the manipulated latent code.

3.2. Affective Music Generation

Lyrics and melodies serve as essential vehicles for conveying emotional content and semantic information within audio. While melody predominantly represents semantic information, lyrics offer a direct avenue for extracting sentiment data from audio.

Initially, the extraction of lyrics from the audio is accomplished through the application of audio-captioning techniques. In particular, we deploy the pre-trained model Whisper-AT [43] to accurately transcribe the lyrics from the audio. Subsequently, to ascertain the sentiment polarity, P (designated as "positive" or "negative") of these lyrics, we employ ChatGLM [44], a state-of-the-art language model. With the obtained lyrics, L, and determined sentiment polarity, P, in hand, we utilize GRU-EBS, a sophisticated model tailored for generating emotionally resonant music segments, thereby enriching the auditory experience with a specified emotional depth:

$$M = \text{GRU-EBS}\,(L, P) \tag{3}$$

3.3. Sentiment-Assisted Guidance Module

Capitalizing on the outputs from the prior modules—controlled images, emotionally enriched audio, and lyrics—we are positioned to engage this multi-faceted information for the purpose of image manipulation.

Employing the diverse data derived from lyrics, audio, and emotionally charged music, we advocate for the utilization of multi-dimensional direct latent code optimization. Drawing inspiration from StyleCLIP [6], our approach seeks to harness direct latent code optimization as a means for conducting sophisticated image manipulation. This method allows for the integration of complex, multi-layered information into the image editing process, enabling a more nuanced and emotionally congruent visual representation:

$$\mathcal{L} \;=\; \underset{w_a \in \mathcal{W}+}{\operatorname{argmin}}\, \mathcal{L}_{\cos} + \lambda_{\mathrm{ID}} \mathcal{L}_{\mathrm{ID}}(w_a) + \lambda_{\mathrm{sim}}\, \mathcal{L}_{\mathrm{reg}} \tag{4}$$

where w_a stands for the manipulated latent code, $\mathcal{W}+$ denotes the latent space of StyleGAN, and λ_{ID} and λ_{sim} are hyperparameters.

3.3.1. Multi-Dimension Guidance Loss

Initially, we encode the image, audio, music, and lyrics into a cohesive latent space to facilitate their integration. This is achieved by utilizing the image encoder and text encoder provided by CLIP for images and lyrics, respectively, while embedding the given audio and generated music through the use of a pre-trained ResNet model, as outlined by Lee et al. [8].

With the resultant visual embeddings, E_v, text embeddings, E_t, audio embeddings, E_a, and music embeddings, E_m, in hand, we introduce a multi-dimensional guidance loss. This loss function is designed to minimize the cosine distance between the visual embedding vector and the embedding vectors from the other modalities. This approach aims to ensure a harmonious integration of information across different dimensions, facilitating a more cohesive and unified image manipulation outcome:

$$\mathcal{L}_{\cos} = \lambda_{\text{Aud}} d_{\cos}(G(w_a), a) + \lambda_{\text{Mus}} d_{\cos}(G(w_a), m) + \lambda_{\text{Tex}} d_{\cos}(G(w_a), t) \tag{5}$$

where w_a stands for the latent code manipulated during the audio-guided process, and $d_{\cos}()$ denotes the cosine distance. The hyperparameters λ_{Aud}, λ_{Mus}, and λ_{Tex} are instrumental in modulating the influence levels of audio, music, and text, respectively. With this loss function, we guide the latent code, w_a, and manipulate the image, $G(w_a)$, with triple dimensions (audio, music, and text).

3.3.2. Identity Loss

To ensure the manipulated image retains similarity to the original input image, we employ the identity loss function, $\mathcal{L}_{\text{ID}}$, formulated as follows:

$$\mathcal{L}_{\text{ID}}(w_a) = 1 - \langle R(G(w_s), R(G(w_a)))\rangle. \tag{6}$$

Here, R denotes the pre-trained ArcFace model [45] utilized for face recognition, aiming to minimize the cosine distance $\langle R(G(w_s)), R(G(w_a))\rangle$ between the generated image and the target in ArcFace's latent space. This loss function is crucial for altering facial expressions while maintaining the subject's identity intact. For manipulations that do not involve facial identities, we nullify this loss by setting $\lambda_{\text{ID}} = 0$, thus allowing for broader image modifications beyond facial adjustments.

3.3.3. Adaptive Layer Masking

Drawing inspiration from the methodology proposed in [46], we incorporate L_2 regularization to ensure the generated image remains visually coherent with the original. Within the StyleGAN latent space, image similarity is preserved by minimizing the L_2-distance between the adjusted latent code, w_a, and the original latent code, w, as delineated by the following equation:

$$\mathcal{L}_{\text{reg}} = \|w_a - w\|_2 \tag{7}$$

To facilitate nuanced control over various attributes at different layers within StyleGAN, we employ an adaptive layer masking strategy. This approach enables the retention of compact content information within the style latent codes, as captured by:

$$\mathcal{L}_{\text{reg}} = \frac{1}{L} \sum_{i=1}^{L} g_i \cdot \|(w_i^a - w_i)\|_2 \tag{8}$$

where L signifies the total number of StyleGAN layers, and g represents a trainable vector that selectively masks specific style layers. This vector, g, is subject to iterative optimization, thereby facilitating adaptive adjustments to the latent code in accordance with the desired image manipulations.

4. Experiments

4.1. Datasets

LHQ (Landscapes High-Quality): presented by Skorokhodov et al. [47], the LHQ dataset comprises 90,000 high-resolution images of natural landscapes. These images were sourced from Unsplash and Flickr, and underwent preprocessing using Mask R-CNN and Inception V3 to ensure quality and consistency. This dataset serves as a significant resource for studies requiring detailed and diverse landscape imagery.

FFHQ (Flickr-Faces-High-Quality): developed by Karras et al. [2], the FFHQ dataset contains 70,000 high-quality images of human faces, sourced from Flickr. Each image in the dataset has been meticulously processed to maintain consistency in terms of quality and resolution, making it an invaluable asset for research in facial recognition, generative modeling, and other areas requiring detailed facial imagery. The dataset is renowned for its diversity in age, ethnicity, and background settings, providing a comprehensive foundation for robust machine learning applications.

4.2. Implementation Details

We apply StyleGAN2 [42], pre-trained, respectively, on LHQ and FFHQ, to generate images. We employ the ImageBind [9] model to compute the similarity between different audio and image modalities and select the image most similar to the given audio. We acquire the lyrics from the provided audio using Whisper-AT [43], and subsequently input them into ChatGLM-6B [44] with the prompt "Judge the sentiment polarity (positive or negative):" to determine the sentiment polarity. Using the lyrics as seeds and sentiment polarity, we employ GRU-EBS to generate emotionally rich music.

We employ the pre-trained CLIP [7] model as the image and text encoder. For the audio encoding, we utilize the pre-trained ResNet50 [8] model with an output dimension set to 512, aligning it with the image and text encoders. To start, we convert audio and music inputs into Mel-spectrogram acoustic features. Subsequently, these features are fed into our audio encoder, resulting in a 512-dimensional latent vector.

In the sentiment-assisted guidance process, we set λ_{ID} and λ_{sim} in Equation (4) to 0.008 and 0.004 for the FFHQ dataset, and 0.002 and 0 for the LHQ dataset. Additionally, we configure the direct code optimization process to run for 300 epochs. We take the text-guided image manipulation method StyleCLIP [6], and the audio-guided image manipulation method proposed by Lee et al. [8] as baselines.

All experiments are conducted on a computing platform equipped with an NVIDIA RTX 3090 graphics card. This high-performance GPU provides the necessary computational power to support complex image generation and processing tasks, ensuring efficient and stable execution of the experimental procedures.

4.3. Qualitative Analysis

4.3.1. Comparison in Lyric-Based Audio Settings

In our study, we specifically collect music tracks that featured lyrics rich in emotional content, aiming to analyze how this emotional richness influences image manipulation when used as input. This approach allows us to compare the effectiveness of different methods in translating the emotional nuances of songs into visual representations.

For the StyleCLIP method, we extract the lyrical content of the songs and use it as textual input. This provides a basis for assessing how well text-based manipulation captures the emotional essence of the lyrics. As shown in Figure 2, we find that image manipulation driven by audio input generally produces better quality results when compared with methods that rely solely on textual input, like StyleCLIP, or methods that use audio in a different way, such as the approach proposed by Lee et al. [8].

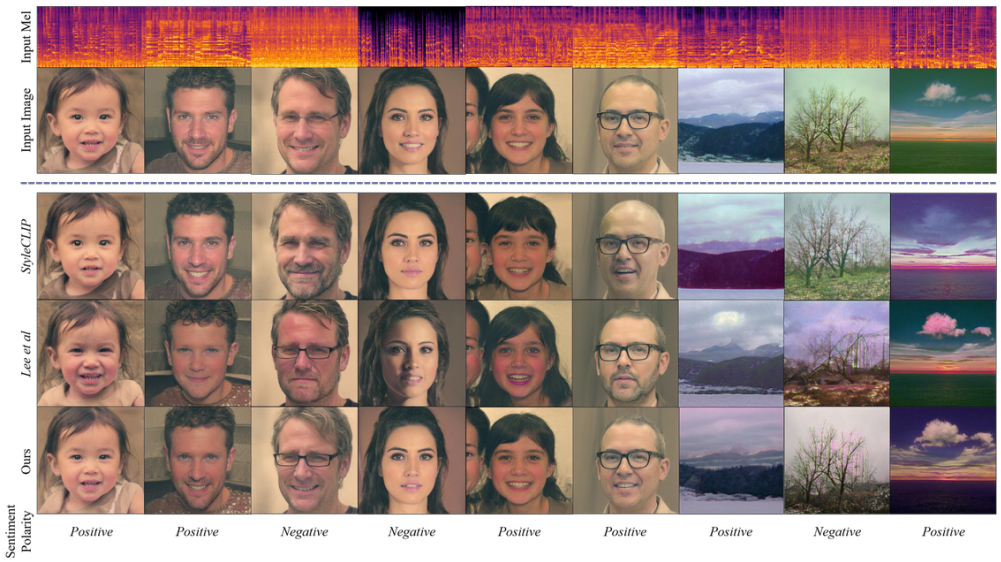

Figure 2. Comparison with different baselines under lyric-based audio settings.

In the context of facial generation within the FFHQ setting, it was observed that StyleCLIP, guided by lyrics text, often struggled with accurately interpreting and representing the emotional dimensions of the songs. This misalignment indicates a limitation in the method's ability to fully grasp and translate the emotional subtleties present in the lyrics into corresponding visual cues.

Similarly, the Lee et al. [8] method, which utilizes audio input, showed some instability in maintaining image fidelity. The images generated by this method often lacked naturalness, suggesting difficulty in achieving a harmonious balance between the audio input and the visual output.

The landscape images within the LHQ setting further highlighted these discrepancies. When examining the natural scenery results (shown in the rightmost three columns of Figure 2), both StyleCLIP and the Lee et al. methods seemed to falter in accurately representing the emotional dimensions conveyed by the audio. This inadequacy was evident in their failure to capture the mood and tone suggested by the music's emotional content.

In stark contrast, our proposed method, EmoStyle, demonstrated a superior ability to effectively handle audio inputs with lyrics. It is noteworthy that, for the third image in Figure 2, our method generated an expression depicting a bitter smile, representing a negative emotional change. This contrasts with images produced by other methods, as our approach excels in capturing and expressing emotional nuances while maintaining facial texture features. Similarly, for the fifth image in Figure 2, which originally had a positive expression, our method was tested for its ability to accurately identify and edit the image to enhance the appropriate emotional tone, ensuring the facial features and visual recognizability were preserved. The slight rosiness added to the cheeks in our generated image subtly amplified the positive emotion, illustrating the effectiveness and sensitivity of our approach in both scenarios. Thus, it is evident that EmoStyle not only accurately represented the emotional content of the songs but also maintained high image fidelity after incorporating this emotional information. This suggests that EmoStyle is more adept at interpreting and visually rendering the complex interplay of emotions conveyed in music, offering a more nuanced and faithful translation of audio-lyrical content into visual imagery.

4.3.2. Comparison in Semantically Rich Audio Settings

In order to evaluate the efficacy of different methods in guiding image semantics based on audio input, we conducted a comparative analysis using eight semantically rich audio samples. Given the absence of lyrics in these samples, we employed textual labels

derived from the audio as a substitute for lyrical content in our EmoStyle method. This approach was aimed at enriching the semantic guidance of our model. As depicted in Figure 3, we observed that all three methods under consideration—StyleCLIP, Lee et al. [8], and our EmoStyle—were capable of extracting semantic information from the audio to influence the generation of images. However, notable differences emerged upon closer examination of their performance. The StyleCLIP method, while effective in semantic extraction, exhibited shortcomings in terms of image fidelity. This was particularly evident in instances where certain elements were missing in the generated images, such as the absence of glasses in an image of a man. This indicated a limitation in the method's ability to fully capture and translate all semantic details from the audio. On the other hand, the method proposed by Lee et al. [8] showed some instability in its semantic control generation. This instability often led to unnatural image guidance effects, highlighting a potential area for improvement in ensuring consistency and naturalness in image manipulation. In contrast, our proposed EmoStyle method demonstrated a more balanced performance. It not only showed natural and stable manipulation of image semantics but also maintained high fidelity in the generated images. This suggests that EmoStyle effectively bridges the gap between semantic extraction from audio and its translation into visually coherent and complete images, outperforming the other methods in creating more accurate and holistic visual representations based on audio semantics.

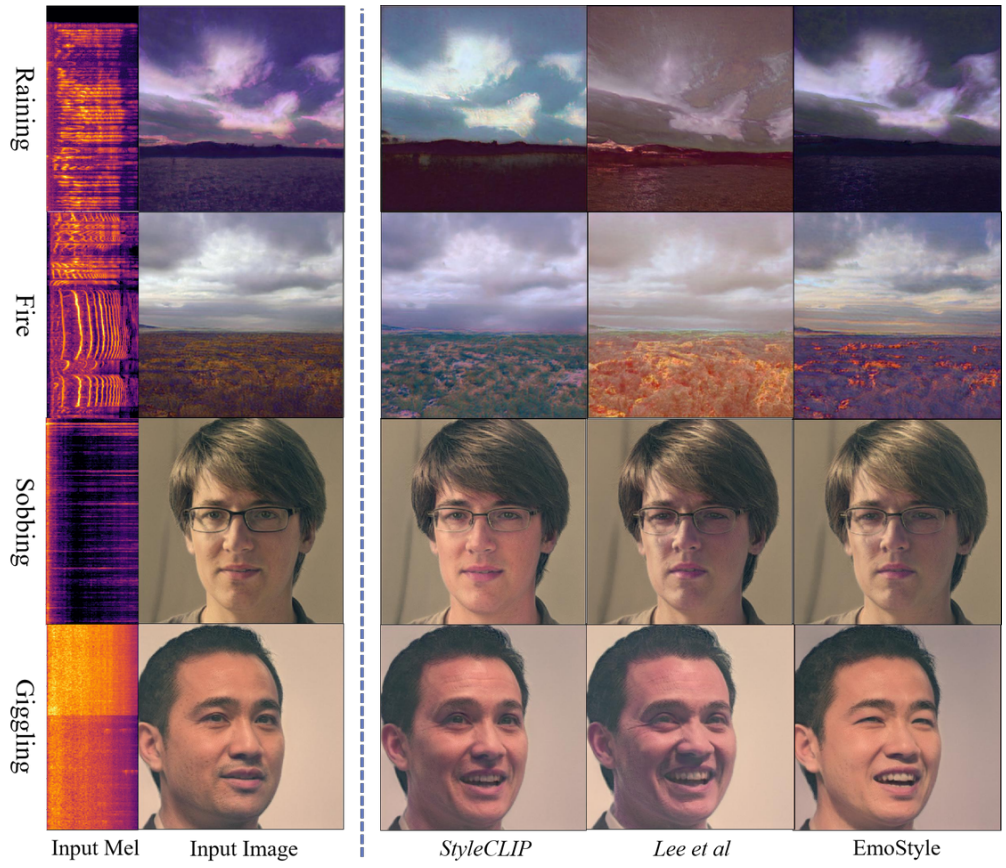

Figure 3. Comparison with selected baselines under semantically rich audio settings.

4.3.3. Ablation Study

Our proposed EmoStyle integrates multi-dimensional information for image manipulation, encompassing text, audio, and music components. To discern the individual effectiveness of each component, we conducted a series of ablation studies with three additional branches: "EmoStyle wo Audio", "EmoStyle wo Music", and "EmoStyle wo Text". In the "EmoStyle wo Audio" branch, where audio input was excluded, a reliance solely on generated music resulted in a noticeable decline in image fidelity, particularly in facial

images, as depicted in Figure 4. This highlighted the critical role of audio in enhancing image realism. Conversely, the "EmoStyle wo Music" branch, devoid of music guidance, revealed images with subdued emotional expressions and lower fidelity, especially in facial representations. This suggested that music components significantly contribute to the emotional depth of the images. Lastly, the "EmoStyle wo Text" branch, which operated without text guidance, demonstrated a propensity for errors in emotional judgment, leading to less natural image generation. This underlined the importance of text in providing contextual and emotional clarity. Comparing these branches with the complete EmoStyle model underscored the collective importance of text, audio, and music components in achieving more nuanced, emotionally resonant, and realistic image manipulation. This multi-faceted integration ensures a richer and more faithful representation, validating the effectiveness of our comprehensive approach in audio-visual image manipulation.

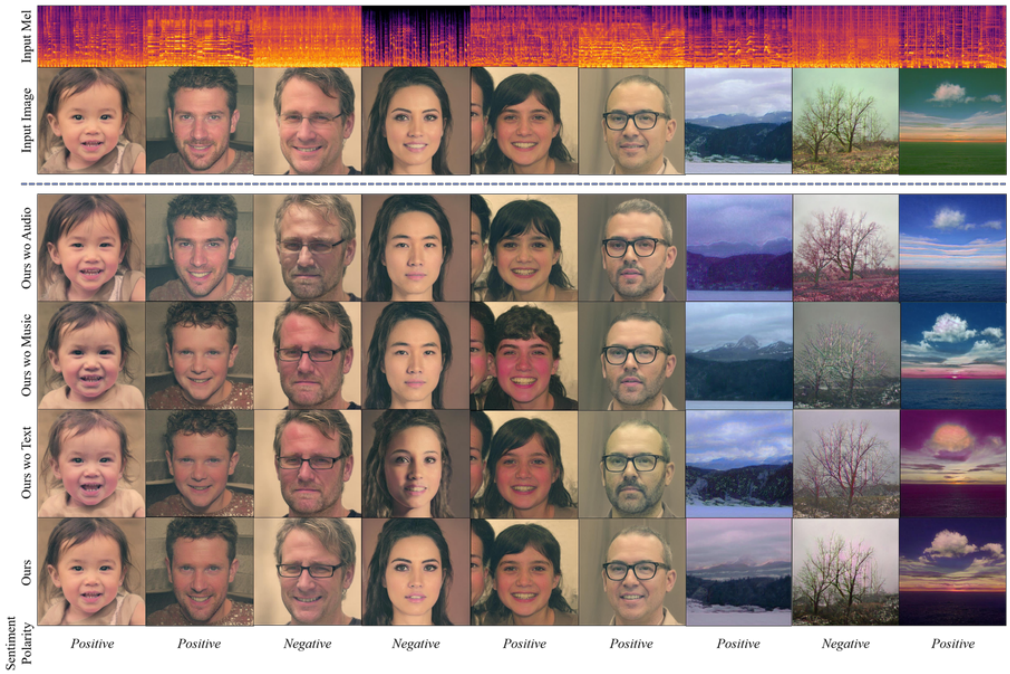

Figure 4. Ablation studies of the EmoStyle.

4.4. Evaluation of Visual Attractiveness

Within the domain of image synthesis, the aesthetic quality of the generated visuals stands as a pivotal metric for assessment. This dimension accentuates not merely the need for algorithmic accuracy but also the significance of the visual allure and impact conveyed by the images, mirroring their capacity to captivate and harmonize with viewer perceptions. Nevertheless, the reliance on sampling methodologies for human evaluation invariably instills bias, imbuing the experimental framework with considerable subjectivity and undermining the dependability of the findings. Against this backdrop, our methodology pivots to a quantitative analytical paradigm, harnessing the capabilities of the contemporary, comprehensive aesthetic evaluation framework, VILA [48]. This approach facilitates rigorous experimentation across dual datasets, employing both methods under scrutiny to derive the mean aesthetic evaluation. Empirical evidence, as delineated in Table 1, establishes EmoStyle's ascendancy in performance over competing models across the evaluated datasets, substantiating its prowess in engendering images of paramount aesthetic merit. This ascendant aesthetic achievement infers that EmoStyle adeptly amalgamates emotional and stylistic nuances within its image synthesis protocol, culminating in creations that not only exhibit enhanced visual enticement but also align more closely with the nuanced contours of human aesthetic preferences.

Table 1. Comparison of VILA aesthetic scores across three models on datasets LHQ and FFHQ.

Method	LHQ	FFHQ
StyleCLIP [6]	0.315	0.303
Lee et al. [8]	0.363	0.331
EmoStyle	**0.523**	**0.479**

5. Conclusions

In this paper, we introduce EmoStyle, a pioneering framework that integrates emotional information and leverages multi-dimensional audio data for image manipulation. We employ a semantic similarity-based image retrieval module to select the appropriate image for manipulation. The affective music generation module is introduced for explicit emotion information extraction, enhanced by an emotional music branch to aid in manipulating emotional dimensions. Utilizing vectorized embeddings of multi-modal data, we propose the sentiment-assisted guidance module, which employs direct latent code optimization for image manipulation. Through objective experiments, we have demonstrated the effectiveness of EmoStyle in manipulating images along emotional dimensions.

In conclusion, EmoStyle not only showcases the successful integration of emotional information with image manipulation techniques but also holds significant potential for application in real-world environments. It could serve as a personalized and emotionally rich visual content creation tool in sectors such as advertising, social media, and entertainment. For future work, the framework could be optimized and expanded by delving deeper into the interaction mechanisms between emotion analysis and image processing, enhancing the precision and naturalness of emotional dimension manipulation. Exploring the application of this technology in broader domains such as virtual and augmented reality also presents a promising direction.

Author Contributions: Conceptualization, Q.S. and J.X.; methodology, J.X.; validation, Q.S. and J.M.; data curation, J.M.; writing—original draft preparation, Q.S., J.X. and X.W.; writing—review and editing, J.M., X.W. and D.D.; visualization, Q.S.; supervision, D.D.; project administration, D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
2. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
3. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
4. Johnson, J.; Gupta, A.; Li, F.-F. Image generation from scene graphs. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1219–1228.
5. Chen, L.; Srivastava, S.; Duan, Z.; Xu, C. Deep cross-modal audio-visual generation. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Silicon Valley, CA, USA, 23–27 October 2017; pp. 349–357.
6. Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 2085–2094.
7. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8748–8763.
8. Lee, S.H.; Roh, W.; Byeon, W.; Yoon, S.H.; Kim, C.; Kim, J.; Kim, S. Sound-guided semantic image manipulation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3377–3386.

9. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. Imagebind: One embedding space to bind them all. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 15180–15190.
10. Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. Cogview: Mastering text-to-image generation via transformers. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021; Volume 34, pp. 19822–19835.
11. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-a-video: Text-to-video generation without text-video data. *arXiv* **2022**, arXiv:2209.14792.
12. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 23716–23736.
13. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734.
14. Sun, J.; Li, Q.; Wang, W.; Zhao, J.; Sun, Z. Multi-caption text-to-face synthesis: Dataset and algorithm. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Chengdu, China, 20–24 October 2021; pp. 2290–2298.
15. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1316–1324.
16. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5907–5915.
17. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1947–1962. [CrossRef] [PubMed]
18. Cheng, J.; Wu, F.; Tian, Y.; Wang, L.; Tao, D. RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis From Prior Knowledge. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
19. Dong, H.; Yu, S.; Wu, C.; Guo, Y. Semantic image synthesis via adversarial learning. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5706–5714.
20. Kim, G.; Kwon, T.; Ye, J.C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2426–2435.
21. Li, B.; Qi, X.; Lukasiewicz, T.; Torr, P.H. Manigan: Text-guided image manipulation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7880–7889.
22. Liu, Y.; De Nadai, M.; Cai, D.; Li, H.; Alameda-Pineda, X.; Sebe, N.; Lepri, B. Describe what to change: A text-guided unsupervised image-to-image translation approach. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Seattle, WA, USA, 16 October 2020; pp. 1357–1365.
23. Nam, S.; Kim, Y.; Kim, S.J. Text-adaptive generative adversarial networks: Manipulating images with natural language. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018 .
24. Xia, W.; Yang, Y.; Xue, J.H.; Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2256–2265.
25. Xu, Z.; Lin, T.; Tang, H.; Li, F.; He, D.; Sebe, N.; Timofte, R.; Van Gool, L.; Ding, E. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18229–18238.
26. Aytar, Y.; Vondrick, C.; Torralba, A. Soundnet: Learning sound representations from unlabeled video. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
27. Korbar, B.; Tran, D.; Torresani, L. Cooperative learning of audio and video models from self-supervised synchronization. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018; Volume 31.
28. Owens, A.; Wu, J.; McDermott, J.H.; Freeman, W.T.; Torralba, A. Ambient sound provides supervision for visual learning. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 801–816.
29. Gan, C.; Huang, D.; Zhao, H.; Tenenbaum, J.B.; Torralba, A. Music gesture for visual sound separation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10478–10487.
30. Su, K.; Liu, X.; Shlizerman, E. Audeo: Audio generation for a silent performance video. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 3325–3337.
31. Morgado, P.; Li, Y.; Nvasconcelos, N. Learning representations from audio-visual spatial alignment. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 4733–4744.
32. Mo, S.; Morgado, P. A Unified Audio-Visual Learning Framework for Localization, Separation, and Recognition. *arXiv* **2023**, arXiv:2305.19458.

33. Gao, R.; Feris, R.; Grauman, K. Learning to separate object sounds by watching unlabeled video. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–53.
34. Tian, Y.; Hu, D.; Xu, C. Cyclic co-learning of sounding object visual grounding and sound separation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2745–2754.
35. Gao, R.; Grauman, K. Visualvoice: Audio-visual speech separation with cross-modal consistency. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscateville, NJ, USA, 2021; pp. 15490–15500.
36. Tian, Y.; Shi, J.; Li, B.; Duan, Z.; Xu, C. Audio-visual event localization in unconstrained videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 247–263.
37. Tian, Y.; Li, D.; Xu, C. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 436–454.
38. Lin, Y.B.; Tseng, H.Y.; Lee, H.Y.; Lin, Y.Y.; Yang, M.H. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021; Volume 34, pp. 11449–11461.
39. Mo, S.; Tian, Y. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 34722–34733.
40. Morgado, P.; Nvasconcelos, N.; Langlois, T.; Wang, O. Self-supervised generation of spatial audio for 360 video. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018; Volume 31.
41. Gao, R.; Grauman, K. 2.5D visual sound. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 324–333.
42. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
43. Gong, Y.; Khurana, S.; Karlinsky, L.; Glass, J. Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers. *arXiv* **2023**, arXiv:2307.03183.
44. Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. GLM-130B: An Open Bilingual Pre-trained Model. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
45. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
46. Lee, S.H.; Oh, G.; Byeon, W.; Yoon, S.H.; Kim, J.; Kim, S. Robust sound-guided image manipulation. *arXiv* **2022**, arXiv:2208.14114.
47. Skorokhodov, I.; Sotnikov, G.; Elhoseiny, M. Aligning latent and image spaces to connect the unconnectable. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14144–14153.
48. Ke, J.; Ye, K.; Yu, J.; Wu, Y.; Milanfar, P.; Yang, F. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 10041–10051.