MELFUSION: Synthesizing Music from Image and Language Cues using Diffusion Models
Appendix

In this appendix we provide additional information on the following:

A. More Details on TANGO++

Our modified baseline model TANGO++ comprises an early-fusion approach, where we align the visual and the textual modalities through an Image-Text Contrastive (ITC) loss. As the generated music is conditioned on both modalities, bringing them to a common latent space is imperative to the success of the system. The text input is passed through the FLAN-T5 text encoder which we keep as frozen. For image encoding we use ViT [10]. We project the visual and the textual inputs to a common embedding space and align them using ITC loss. The diffusion model is conditioned on this hybrid embedding to produce audio signals. It is then converted into spectrograms using the decoder and then passed through a HiFi GAN vocoder to produce the music signal. The expression for ITC loss ($\mathcal{L}_{\text{ITC}}$) is as follows:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{2\mathcal{N}} \sum_{j=1}^{\mathcal{N}} \underbrace{\log \left[\frac{\exp\left(\langle z_j^I, z_j^T \rangle / \tau\right)}{\sum_{l=1}^{\mathcal{N}} \exp\left(\langle z_j^I, z_l^T \rangle / \tau\right)} \right]}_{\text{Contrasting images with the texts}} - \frac{1}{2\mathcal{N}} \sum_{l=1}^{\mathcal{N}} \underbrace{\log \left[\frac{\exp\left(\langle z_l^I, z_l^T \rangle / \tau\right)}{\sum_{j=1}^{\mathcal{N}} \exp\left(\langle z_j^I, z_l^T \rangle / \tau\right)} \right]}_{\text{Contrasting texts with the images}} \quad (1)$$

where $\langle \cdot, \cdot \rangle$ denotes inner product, and τ is the temperature parameter. z^I and z^T refer to the image and text latent representations respectively.

B. Problem Motivation Revisited

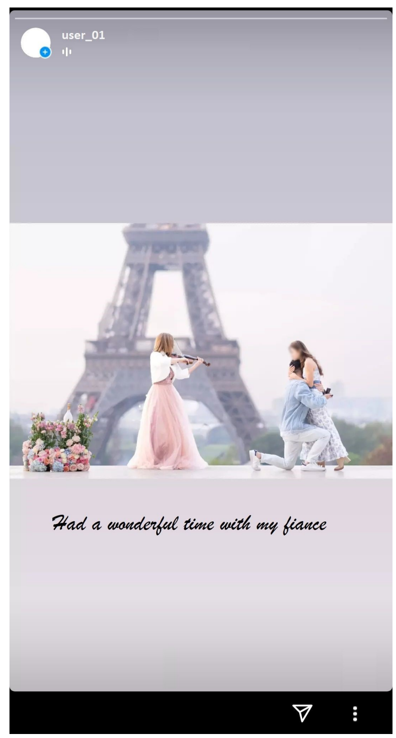

Figure 1. A mock-up of a social media post that contains an image and associated textual content. Our approach MELFUSION, can consume such image-textual pairs as input and synthesize music that can go well with them.

Social media platforms have become ubiquitous and provide a channel for everyone to express their creativity and share their happenings with the world. It is very common for users to upload an image, and write an associated text with it (Fig. 1). Adding music to these social media posts enhances its visibility and appeal. Instead of retrieving music from an existing database, our approach MELFUSION, will be able to generate music tracks that are custom-made, conditioned on the uploaded image and its description. We note that ours is the first approach that operates in this pragmatic setting, to generate music conditioned on both visual and textual modality.

C. Other Baseline Approaches

In addition to our proposed baseline approach, we compare MELFUSION against the following methods. Note that these are text-to-music generation methods unlike our approach and don't support multi-conditioning in input prompts. Hence a direct comparison might not be entirely fair. In most cases these methods don't support introducing an additional modality conditioning as a result we compare our approach against these baselines directly to study the benefits of MELFUSION.

Riffusion [11] base their algorithm on fine-tuning a Stable Diffusion model [42] on mel spectrograms of music pieces from a paired music-text dataset. This is one of the first text-to-music generation methods. Mubert [35] is an API-based service that employs a Transformer backbone. The encoded prompt is used to match the music tags and the one with the highest similarity is used to query the audio generation API. It operates over a relatively smaller set as it produces a combination of audio from a predefined collection. MusicLM [1] generates high-fidelity music from text descriptions by casting the process of conditional music generation as a hierarchical sequence-to-sequence modeling task. They leverage the audio-embedding network of MuLan [17] to extract the representation of the target audio sequence. Moûsai [46] is a cascading two-stage latent diffusion model that is equipped to produce long-duration high-quality stereo music. It achieves this by employing a specially designed U-Net facilitating a high compression rate. Noise2Music [18] introduced a series of diffusion models, a generator, and a cascader model. The former generates an intermediate representation conditioned on text, while the later can produce audio conditioned on the intermediate representation of the text. MeLoDy [26] pursues an LM-guided diffusion model by reducing the forward pass bottleneck and applies a novel dual-path diffusion mode. MusicGen [8] comprises a single-stage transformer LM together with efficient token interleaving patterns. This eliminates the need for hierarchical upsampling.

D. Implementation Details

Our text-to-music LDM contains 3 encoder blocks and 3 decoder blocks, similar to Ghosal et al. [13]. Empirically we find that finetuning from its pre-trained checkpoint helps convergence. FLAN-T5 [7] is used as the text encoder. MELFUSION is trained for 30 epochs using AdamW optimizer [33]. We attach our visual synapse only on the decoder layers of the LDM. Similar to earlier works [13, 27], we find that using classifier-free guidance improves the result. Our training takes 42 hours on 4 NVIDIA A100 GPUs.

E. More Experimental Analysis

E.1. Choice of Text-to-Image Diffusion Model

Model	MusicCaps			MeLBench		
	FD ↓	KL ↓	FAD ↓	FD ↓	KL ↓	FAD ↓
Stable Diffusion V1.2	1.84	1.52	22.88	1.49	1.14	21.44
Stable Diffusion V1.3	1.62	1.29	22.72	1.34	1.03	21.02
Stable Diffusion V1.4	1.31	1.13	22.67	1.20	0.91	20.53
Stable DiffusionV1.5	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 1. MELFUSION with different versions of Stable Diffusion.

We study the effect of employing different variants of the text-to-image Stable Diffusion model (V1.2 through V1.5) in Tab. 1. We note that the best results are obtained with the latest variant. This brings to light that our proposed visual synapse is able to cascade the usage of better text-to-image models into improving the quality of music generation. The Stable Diffusion V1.4 and V1.5 checkpoints were initialized with the weights of the Stable Diffusion V1.2 checkpoint and subsequently fine-tuned on 225k steps at resolution 512×512 on the LAION dataset and 10% dropping of the text-conditioning to improve classifier-free guidance sampling.

E.2. Performance with Different Text Encoders

Model	MusicCaps			MeLBench		
	FAD ↓	KL ↓	FD ↓	FAD ↓	KL ↓	FD ↓
BERT [9]	2.82	2.23	24.73	2.91	1.94	22.13
RoBERTa [31]	2.35	2.02	24.09	2.17	1.87	21.95
T5-Small [40]	1.98	1.79	23.68	1.89	1.66	21.23
T0 [45]	1.42	1.25	22.96	1.32	1.19	20.76
CLIPText	1.24	0.94	22.78	1.16	0.91	20.58
FLAN-T5 [7]	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 2. Performance of MELFUSION with different text encoders

In Tab. 2 we compare the performance of MELFUSION under different text encoders. We note that the best results are achieved when an instruction-tuned text encoder is employed (FLAN-T5 [7]) over other non-instruction-based models, which correlates with the findings in Ghosal et al. [13]. This is very closely followed by the ClipText [39] encoder.

E.3. Variation Across Genres

Genre name	Objective metrics				Subjective metrics	
	FD ↓	KL ↓	FAD_{VGG} ↓	IMSM ↑	OVL ↑	REL ↑
Pop	22.47	0.78	1.21	0.95	86.31	90.10
Rock	21.11	0.95	0.85	0.81	88.41	84.92
Hip-Hop/Rap	19.73	0.65	1.24	0.69	83.05	88.78
Electronic Dance Music	20.03	1.06	0.93	0.72	85.39	86.18
Country	19.56	0.89	0.88	0.98	89.94	87.22

Table 3. A study on the diversity analysis of MELFUSION. We evaluate the performance of our model on generating musical tracks of five different genres on MeLBench.

Tab 3 reports the performance of MELFUSION across the 5 most popular genres (chosen through a study undertaken by [16]) on the genre-wise test set collected from MeLBench. We find a steady performance of our approach across different genres substantiating the ability of the model to capture the musical nuances like the composition of the instruments, track progression, sequence of instruments introduced, rhythm, tonality, tempo, and beats. Due to the highly subjective nature of the problem, we also perform a human evaluation by subject matter experts. To this end, we employ 7 individuals formally trained in music to independently listen and report OVL and REL scores considering the aforementioned aspects to assess the quality of genre-wise samples. We report the mean OVL and REL values from all the evaluators on a subset of the corresponding genre-wise test splits. We find that the overall performance of our method is highly encouraging as reported in Tab 3.

E.4. Ablating choice of layers

When we fuse subset of Decoder Blocks, we see drop in performance in Tab. 4, as coupling becomes weak. We also ablate encoder and decoder layer separately (refer to Tab. 2 of main paper). Learned α values for each blocks (0.37, 0.59 and 0.63 respectively) improves over α=0.5 on all metrics, thus avoiding an extra hyper-parameter to tune. With a few layers to account for dimension mismatch, visual synapse can scale to different architectures and avoid layer-to-layer correspondence. We will explore this in a future work.

Decoder Block	Extended MusicCaps			MeLBench		
	FAD ↓	KL ↓	FD ↓	FAD ↓	KL ↓	FD ↓
1	1.79	1.12	22.97	1.71	1.02	21.20
1,2	1.53	1.05	22.76	1.27	0.86	20.93
1,2,3	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 4. Ablation of different decoder blocks

E.5. On conditioning image

MELFUSION generates music from complementary information from text and image modalities. While selecting images randomly, we have lower FAD/KL/FD scores of 6.38/1.73/26.45 and 8.33/1.57/28.64 on the extended MusicCaps and MeLBench datasets respectively, as it gets conditioned on random image semantics. We see similar trend in the baselines too, and MELFUSION still outperforms them. Retrieving or generating image from conditioning text, will also have similar effect due to semantic similarity in both conditioning domains.

OVL Range	Reasons
0-25	Discordant sound, unpleasant, poor quality, mismatched genre, not cohesive, repetitive melody, distractive background noise, unpleasant timbre, lack of contrast.
26-50	Unappealing instrumentation, lack of emotional resonance, unusual degree of dissonance, complex narrative, unrelatable theme, abrupt transition, unbalanced sound levels.
51-75	Inconsistent mood, uninteresting chord progression, uneven transition between sections, has a nostalgic appeal, cinematic quality, spirituality.
76-100	Exudes calmness, cohesive, pleasing sequence of notes, well balanced combinations, engaging rhythmic pattern, evoke a sense of groove, nice arrangement of instruments, strong sense of expression, authentic, vibrant texture, catchy, intuitive and natural flow.

Table 6. Subjective analysis on generated samples

E.6. Alternate visual conditioning

We compare alternate conditioning from ViT features and ControlNet here. The semantics contained in these representations are inferior to those from text-to-image models (similar to findings in [54]). Further, our visual synapse effectively adapts them by learning to modulate the representations, specific to music synthesis. Moreover compared to the generalist model (that consumes multiple modalities) in AudioLDM2 [28], our specialist synaptic model generates better music. Also, their feature concatenation strategy is inferior to our visual synapse, as evident from Tab. 5.

Model	Extended MusicCaps			MeLBench		
	FAD ↓	KL ↓	FD ↓	FAD ↓	KL ↓	FD ↓
CLIP ViT Feats [39]	1.83	1.15	23.03	1.77	1.04	21.48
Control Net [56]	1.65	1.09	22.94	1.25	0.85	20.91
AudioLDM2 [28]	1.77	1.13	22.96	1.74	1.02	21.42
Ours	**1.12**	**0.89**	**22.65**	**1.05**	**0.72**	**20.49**

Table 5. Comparison against different visual conditioning

E.7. Subjective analysis

We complement our OVL scores with subjective descriptions, where we ask the evaluators to justify the score, stratify them based on OVL scores, and report the most frequent reasons in Tab. 6.

E.8. Learnable versus Fixed α Parameters

Fusion parameter α	Extended MusicCaps			MeLBench		
	FAD ↓	KL ↓	IMSM ↑	FAD ↓	KL ↓	IMSM ↑
$\alpha = 0$	3.07	1.21	-	3.11	1.19	-
$\alpha = 0.10$	2.98	1.17	0.51	3.03	1.07	0.56
$\alpha = 0.50$	1.17	0.93	0.71	1.12	0.79	0.77
$\alpha = 0.90$	4.96	1.38	0.85	4.11	1.29	0.89
$\alpha = 1.0$	5.62	1.54	-	4.16	1.37	-
Learnable α	**1.12**	**0.89**	**0.76**	**1.05**	**0.72**	**0.83**

Table 7. Analyzing the effect of having fixed versus learnable α.

We study the impact when α is kept frozen as compared to being learnable here. The first five entries in Tab. 7 denote the cases where the value of α is unaltered during training and kept constant at 0, 0.10, 0.50, 0.90, and 1.0 respectively. Experimental results demonstrate that a learnable value of α produces significantly better results as compared to the fixed counterpart, as the model has the flexibility to learn them to effectively balance between both the conditioning modalities.

F. Dataset Details

F.1. MeLBench Statistics

Type of image	# Pieces	Percentage (%) in Dataset
Natural image	3206	28
Animation	2404	21
Poster	2748	24
Painting / Sketch	3092	27

Table 9. Image categories in MeLBench.

Tab. 9 presents the distribution of the image samples in MeLBench. To maintain a fair balance across different distributions we collect samples from 4 different categories: natural images, animations, posters, paintings/sketches. This ensures that MELFUSION is trained with ample examples from each of these classes and is equipped to tackle images from any of these very frequent and popular classes better. MeLBench comprises 11,250 samples which is $\sim$ 2x larger than the next largest dataset MusicCaps [1].

Fig. 2 presents the frequency of the top 90 words in MeLBench. The annotators were asked to write free-form text descriptions of the musical pieces with an emphasis on the musicality of the samples. We observe that the annotation contains important cues about the nature of the audio track (e.g., 'live performance', 'chaotic', 'forceful vocals', etc). These can supplement a model with useful pieces of information regarding the aesthetics of the composition.

F.2. Dataset Hierarchy and Samples

Tab. 8 contains the genre and sub-genre-wise division of the samples collected in MeLBench. We categorise the collected musical samples into 15 broad categories with each of them having 22 sub-genres to facilitate fine-grained control over the composition through the image (theme) and text-instructions (details on musicality). The samples are divided across different genres roughly equally to maintain a good balance.

Fig. 3 presents one sample from each of the remaining 13 categories (Electronic and Folk Acoustic present in the main paper). As can be seen from the examples, the captions are of varied lengths and the images are from different distributions (natural images, animation, paintings, etc.).

F.3. Extended MusicCaps Data Collection

MusicCaps [1] is a music caption dataset comprising music clips from AudioSet [12] paired with corresponding text descriptions in English. The collection consists of a total of 5,521 examples, out of which 2,858 are from the AudioSet eval and 2,663 are from the AudioSet train split. The authors further tag 1,000 samples as a balanced subset of the dataset - equally divided across genres. All examples in the balanced subset are from the AudioSet eval split. As our setup is not restricted to text and requires joint conditioning in the form of images as well, we supplement this dataset by collecting 2 carefully chosen image frames for each of the 10-second samples from the corresponding YouTube video or web. As some of the samples are not live anymore, we were able to collect a total of 7,684 samples which we divided into a 60%/20%/20% split for train/validation/test respectively.

G. User Study Details

Fig. 4 presents the user study interface. To obtain the OVL and REL scores, we provide the participants with an image-text pair and the audio sample generated by MELFUSION. For the overall audio quality score (OVL) the participants are instructed to add their score between [1,10] while for the relevance score (REL), they are required to rate the sample based on its similarity with the input image-text pairs.

In Fig. 5 we compare our method against prior text-to-music methods and report the OVL and REL scores in the main paper (Tab. 1). In this case, the participants were presented with only the text-music pairs.

Fig. 6 shows the user study interface for the IMSM score. For this, the participants were presented with image-music pairs and asked to provide their rating between [1,10], with 1 being the lowest. The higher the score, the more perceptually similar the participant has found the image-music pair to be.

H. Inspiration from Conditional Image Generation

Powered by architectural improvements and the availability of large-scale, high-quality paired training data, conditional image generation methods have made considerable progress in the generative AI space. Promising results from transformer-based auto-regressive approaches [41, 55] were boosted by diffusion model-based methods [36, 42, 44]. These approaches have been naturally extended to generate videos from text prompts too [15, 48, 53]. Latent diffusion models [42] do the diffusion process in the latent space of a pre-trained VQ-VAE [51]. This significantly reduced

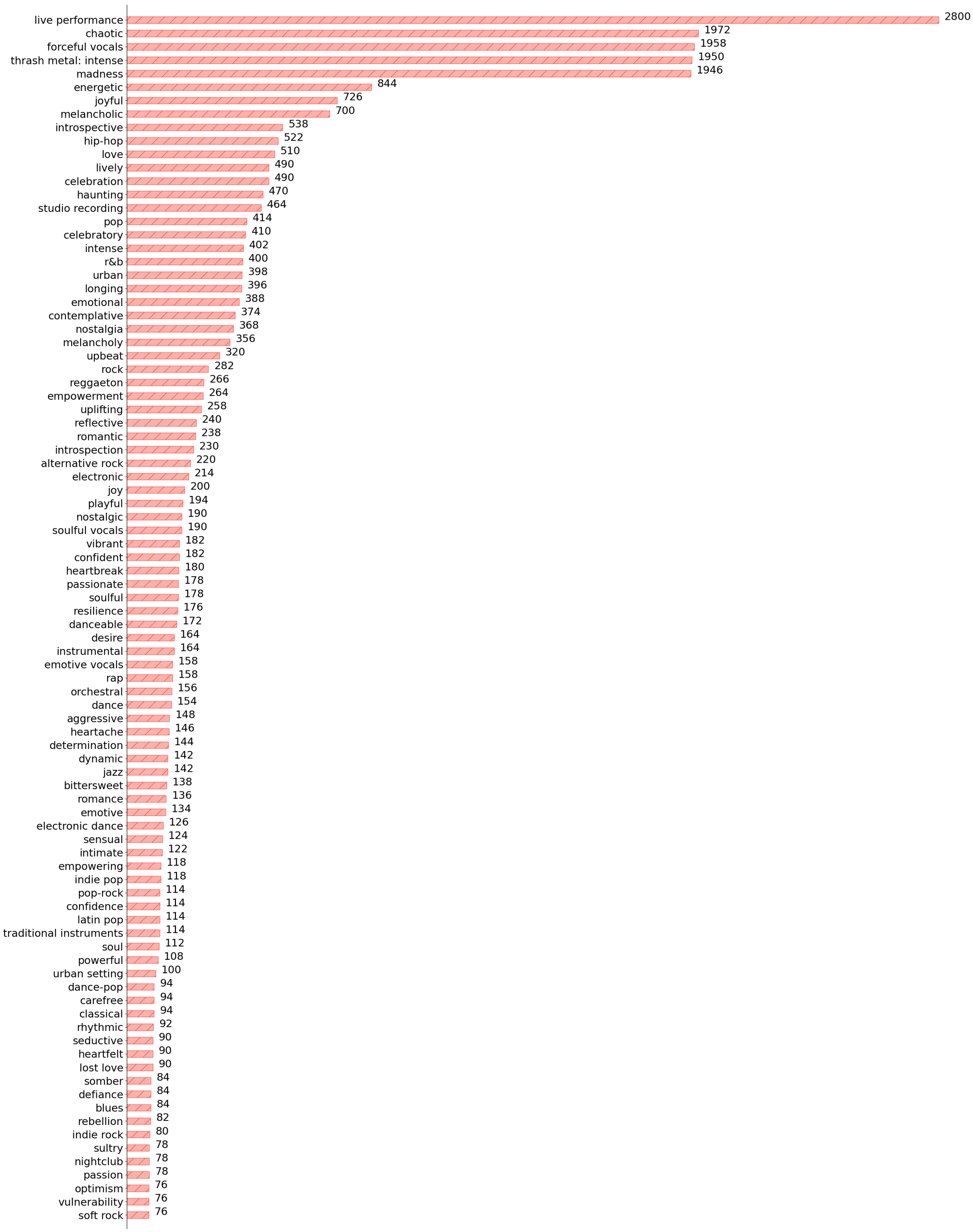

Figure 2. Frequency of top 90 words from MeLBench

New-Age

The track falls within the alternative hip-hop and pop genres. It incorporates electronic beats, synthesizers, and rap-style vocals, creating a contemplative yet catchy composition.

Classical

The baroque composition is executed by a string orchestra, the instruments involved include violins, violas, cellos, and a double bass.

Latin

The track is nestled within the Latin Trap genre. Electronic beats and synthesizers form the backbone of the composition, initiating a rhythmic and engaging melody.

R & B

The track embodies the essence of Indie R&B/Soul, weaving soulful vocals and gentle guitar melodies. The composition contains a background chorus enhancing the emotional depth of the song.

Easy-Listening

The composition features acoustic guitar chords as the foundational instrument, accompanied by soft, melodic strings or subtle orchestration. The composition begins with a gentle guitar introduction and gradually builds as additional instruments are layered in. The track belongs to the chanson genre.

Metal

The track fits into the post-grunge and alternative metal genres. Its composition follows a sequence of instruments, typically kicking off with heavy electric guitars and aggressive, anguished vocals, creating a powerful and emotionally charged atmosphere. As the song progresses, this intensity remains, with drums and bass joining in to maintain the heavy and relentless sound. The vocals are delivered with a raspy and aggressive tone.

Rock

The music track falls under the rock and roll genre and prominently features instruments like electric guitar, bass, drums, and keyboards, with a driving rhythm that propels the composition forward. The vocals are delivered with a confident and energetic tone, fitting the overall spirited nature of the song.

Jazz

The track falls under the traditional jazz genre. It features a harmonious blend of traditional jazz instruments like trumpet, saxophone, piano, and double bass, creating a melodious and timeless musical composition.

World-Traditional

The music track belongs to the world-traditional genre. It features acoustic instruments such as the acoustic guitar, accordion, and a string instrument being introduced in a distinct sequence, creating a haunting and memorable melody.

Hip-Hop

The instrumentation features a mix of electronic beats, piano chords, and smooth vocal delivery, creating a laid-back and introspective composition with subtle chattering noises in the background.

Pop

The song can be classified as a country-rock or country-pop-rock song. It prominently features electric guitars, drums, and horns in a sequence that creates an upbeat and energetic composition.

Blues

The track features powerful electric guitars, drums, bass, keyboards, and possibly horns, creating an emotive and intense musical composition. The vocals are emotive and soulful, conveying determination and strength, resonating with the song's themes of struggle and resilience. The track belongs to blues rock genre.

Country

The track is a country ballad composed of acoustic guitar, pedal steel guitar, and a rhythm section, arranged to form a mournful and wistful composition.

Figure 3. Samples from MeLBench.

Genre	Subgenre
Hip-Hop	Alternative Hip Hop, Rap, Pop Rap, Trap, Melodic Rap, Gangster Rap, Southern Hip Hop, Urban Contemporary, Crunk, German Hip Hop, Rap Conscient, Italian Hip Hop, East Coast Hip Hop, Hardcore Hip Hop, Atl Hip Hop, Dirty South Rap, Russian Hip Hop, Polish Trap, Underground Hip Hop, Funk Carioca, West Coast Rap, Cloud Rap
Pop	Dance Pop, Pov- Indie, Singer-Songwriter Pop, Mexican Pop, J-Pop, Latin Arena Pop, Indie Pop, Modern Country Pop, Art Pop, Alt Z, Indietronica, New Wave Pop, Spanish Pop, Italian Adult Pop, Electropop, Turkish Pop, Reggae Fusion, Post-Teen Pop, Hip Pop, Ccm, Indonesian Pop, Pop Nacional
Latin	Latin Pop, Trap Latino, Urbano Latino, Reggaeton, Musica Mexicana, Rock En Espanol, Norteno, Sierreno,R&B Francais, Reggaeton Colombiano, Sad Sierreno, Mpb, Sertanejo, Tropical, Latin Alternative, Banda, Corrido, Grupera, Ranchera, Trap Brasileiro, Rap Conciencia, Urbano Espanol
Electronic	Edm, Pop Dance, Uk Dance, Electronica, Electro House, House, German Dance, Tropical House, Downtempo, Brostep, Stutter House, Progressive House, Slap House, Big Room, Chill House, New French Touch, Dancefloor Dnb, Chillhop, Pop Edm, Lo-Fi Beats, Trance, Metropopolis
R&B	Soul, Indie Soul, Quiet Storm, Neo Soul, Funk, Alternative R&B, Disco, Pop Soul, Afrobeats, Bedroom R&B, Dark R&B, Reggae, British Soul, Contemporary R&B, Hi-Nrg, Classic Soul, Uk Contemporary R&B, Motown, New Jack Swing, Gospel, Roots Reggae, Philly Soul
Easy listening	Adult Standards, Chanson, Soundtrack, Show Tunes, Hollywood, Movie Tunes, Cartoon, Japanese Soundtrack, Broadway, Deutsch Disney, Swing, British Soundtrack, Lounge, Preschool Children's Music, Scorecore, Romantico, Classic Girl Group, Children's Music, Electro Swing, French Soundtrack, French Movie Tunes, Classic Soundtrack
World / traditional	Folkmusik, Modern Bollywood, Filmi, Pop Urbaine, World, Afroswing, Dancehall, World Worship, Entehno, Sufi, Naija Worship, Classic Bollywood, Nouvelle Chanson Francaise, Modern Reggae, Laiko, Classic Opm, Uk Dancehall, South African Pop Dance, Chutney, Celtic, Manila Sound, Azontobeats
Jazz	Vocal Jazz, Bossa Nova, Dinner Jazz, Contemporary Post-Bop, Jazz Fusion, Nu Jazz, Background Jazz, Smooth Jazz, Jazz Funk, Contemporary Vocal Jazz, Jazz Piano, Jazztronica, Hard Bop, Smooth Saxophone, Cool Jazz, Nz Reggae, Soul Jazz, Torch Song, Folclore Salteno, Indie Jazz, Contemporary Jazz, Brazilian Jazz
Rock	Permanent Wave, Modern Rock, Classic Rock, Mellow Gold, Album Rock, Soft Rock, Pop Rock, Alternative Rock, Hard Rock, Folk Rock, New Wave, New Romantic, Indie Rock, Heartland Rock, Latin Rock, Art Rock, Blues Rock, Dance Rock, Country Rock, Alternative Dance, Pop Punk, Punk
Classical	Orchestral Soundtrack, Compositional Ambient, Classical Performance, Javanese Dangdut, Italian orchestra, Orchestral Performance, Neo-Classical, Orchestra, Classical Piano, British Orchestra, Choral, Opera, Indian Classical, Hungarian Classical, Epicore, Impressionism, Chamber Orchestra, Historically Informed Performance, Violin, Baroque Ensemble, Symfonicky Orchestra, Japanese Guitar
Blues	Electric Blues, Jazz Blues, British Blues, Modern Blues, Malian Blues, Rebel Blues, Acoustic Blues, Rhythm And Blues, Doo-Wop, Traditional Blues, Soul Blues, Louisiana Blues, Garage Rock Revival, Indie Quebecois, New Orleans Blues, Texas Blues, Country Blues, Australian Garage Punk, Chicago Blues, Delta Blues, Memphis Blues, Slack-Key Guitar
Metal	Alternative Metal, Post-Grunge, Nu Metal, Rap Metal, Groove Metal, Power Metal, Melodic Metalcore, Metalcore, Skate Punk Glam Metal, Thrash Metal, Speed Metal, Death Metal, Funk Metal, Screamo, Nerdcore Brasileiro, Industrial Metal, Comic Metal, Symphonic Metal, Deathcore, Gothic Metal, Progressive Metal,
Country	Contemporary Country, Agronejo, Arrocha, Country Road, Sertanejo Universitario, Outlaw Country, Nashville Sound, Pop Rap Brasileiro, Pagode Novo, Arrochadeira, Forro, Forro De Favela, Funk 150 Bpm, Progressive Bluegrass, Black Americana, Axe, Bandinhas, Funk Ostentacao, Alternative Country, Piseiro, Jam Band, Classic Texas Country
Folk/ acoustic	Singer-Songwriter, Neo Mellow, Indie Folk, New Americana, Stomp And Holler, British Singer-Songwriter, Melancholia, Lilith, Turbo Folk, Countrygaze, Neo-Psychedelic, Pop Folk, Turkish Folk, Ambient Folk, Modern Indie Folk, Rune Folk, Indian Folk, Fantasy, Alternative Americana, Ska Punk, Vbs, German Indie
New age	Rain, Color Noise, Sleep, Sound, Healing Hz, Solfeggio Product, Indie Game Soundtrack, Ocean, Environmental, Water, Piano Cover, Acoustic Guitar Cover, Lullaby, High Vibe, Instrumental Worship, Atmosphere, Background Music, Ambient Worship, Binaural, Brain Waves, Background Piano, Fourth World

Table 8. Genre and sub-genre-wise division of the collected samples. Our dataset encompasses samples from 15 different genres each further divided into 22 sub-genres

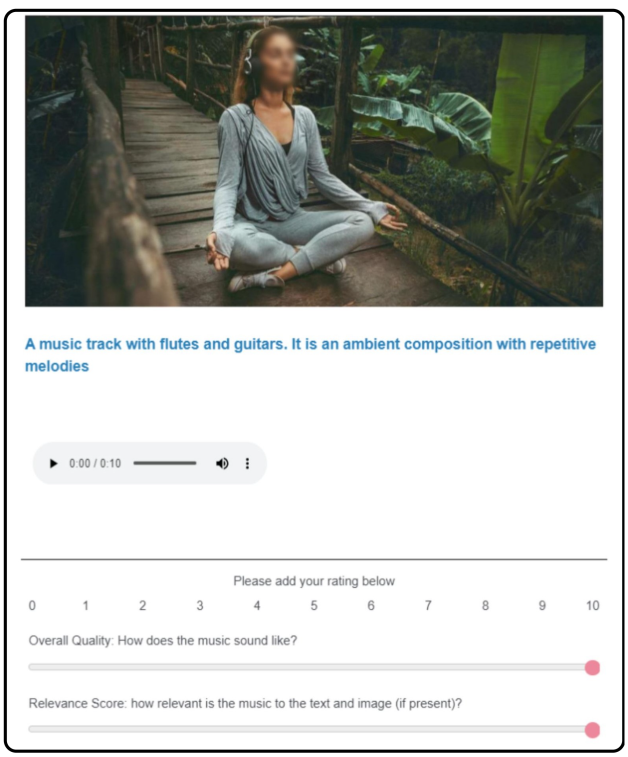

Figure 4. User study interface to collect OVL and REL scores.

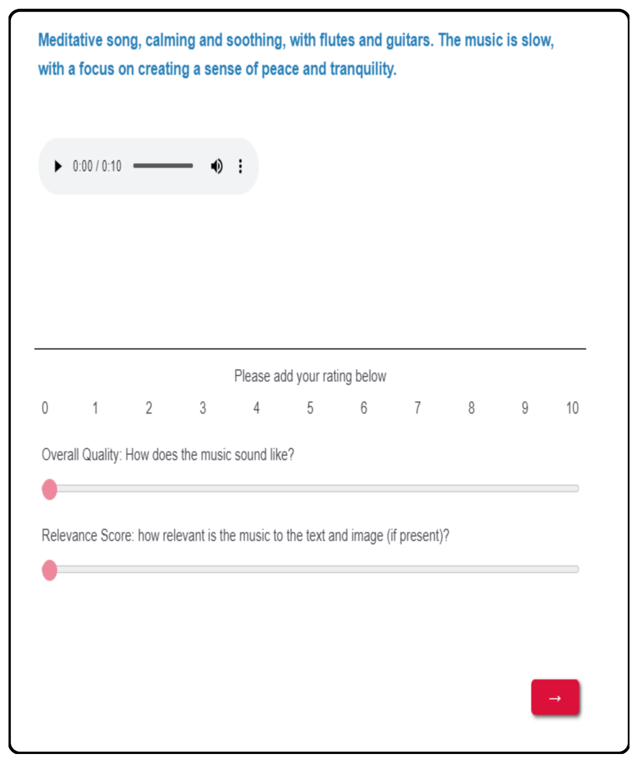

Figure 5. User study interface for comparison against prior text-to-music methods

the compute requirements when compared with image diffusion methods. Ho and Salimans [14] proposed classifier-free guidance to enhance image quality. Text-to-music and text-to-audio methods are heavily inspired by the success of text-to-image generative methods, and so are we.

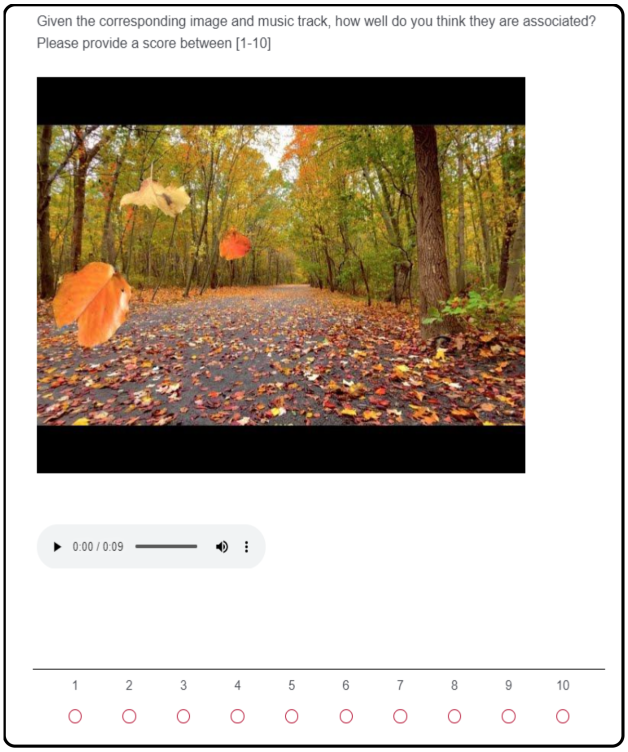

Figure 6. User study interface to obtain IMSM scores

I. Related Audio Concepts

The Multimodal Variational Auto-encoders (MVAEs) are latent variable generative models to learn more generalizable representations from diverse modalities through joint distribution estimation. Arik et al. [2] pioneered a neural audio synthesis model based on VAEs. Their approach demonstrated promising results in generating realistic audio samples by learning a latent representation of the audio data. Inspired by this VAEs have been widely used in the audio processing domain for speech synthesis [30, 50, 57], audio generation [4, 13, 22], and audio denoising [3, 43].

Vocoders are used for a variety of purposes across different domains due to their ability to manipulate and synthesize audio signals efficiently. Among other prominent applications of vocoder, neural voice cloning [2, 21], voice conversion [29], and speech-to-speech synthesis [20] are very popular. GAN-based vocoders [25] have been employed to generate high-fidelity raw audio conditioned on mel spectrogram. More recently, WaveRNN [24] has been applied for universal vocoding task [23, 32, 38].

Spectrograms are a powerful tool for analyzing time-varying signals such as audio and speech. They provide a visual representation of the frequency content of a signal over time, making them widely used in speech processing [6, 34, 47], music analysis [26, 46], and audio synthesis [5, 13, 19, 27, 52] in general. Audio spectrograms are also massively deployed in different audio visual applications [5, 37, 49].

Acknowledgements: We would like to sincerely thank the data annotators and the volunteers who took part in the user study. We would also like to extend our gratitude to the anonymous reviewers for their constructive and thoughtful feedbacks.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 2, 4

[2] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. *Advances in neural information processing systems*, 31, 2018. 9

[3] Yoshiaki Bando, Kouhei Sekiguchi, and Kazuyoshi Yoshii. Adaptive neural speech enhancement with a denoising variational autoencoder. In *Interspeech*, pages 2437–2441, 2020. 9

[4] Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021. 9

[5] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7884–7896, 2023. 9

[6] Shang-Yi Chuang, Hsin-Min Wang, and Yu Tsao. Improved lite audio-visual speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1345–1359, 2022. 9

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[8] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[11] Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation, 2022. *URL https://riffusion.com/about*, 6. 2

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 4

[13] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023. 2, 9

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 8

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4

[16] https://primesound.org/popular-music genres/. https://primesound.org/popular-music-genres/. https://primesound.org/popular-music-genres/, 2023. 3

[17] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022. 2

[18] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 2

[19] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023. 9

[20] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model (2019). *arXiv preprint arXiv:1904.06037*, 1904. 9

[21] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018. 9

[22] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 516–520. IEEE, 2020. 9

[23] Yunlong Jiao, Adam Gabryś, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa, and Viacheslav Klimkov. Universal neural vocoding with parallel wavenet. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6044–6048. IEEE, 2021. 9

[24] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018. 9

[25] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020. 9

[26] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. Efficient neural music generation. *arXiv preprint arXiv:2305.15719*, 2023. 2, 9

[27] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 2, 9

[28] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023. 3

[29] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai. Wavenet vocoder with limited training data for voice conversion. In *Interspeech*, pages 1983–1987, 2018. 9

[30] Peng Liu, Yuewen Cao, Songxiang Liu, Na Hu, Guangzhi Li, Chao Weng, and Dan Su. Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention. *arXiv preprint arXiv:2102.06431*, 2021. 9

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[32] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal. Towards achieving robust universal neural vocoding. *arXiv preprint arXiv:1811.06292*, 2018. 9

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[34] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021. 9

[35] 2023 Mubert Inc. Mubert. URL https://mubert.com/. Mubert inc. mubert. url https://mubert.com/, 2023,. *Mubert Inc. Mubert. URL https://mubert.com/, 2023,*, 2023. 2

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4

[37] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, pages 5711–5720, 2024. 9

[38] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. Speaker conditional wavernn: Towards universal neural vocoder for unseen speaker and recording conditions. *arXiv preprint arXiv:2008.05289*, 2020. 9

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 4

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[43] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1788–1800, 2020. 9

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 4

[45] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 2

[46] Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023. 2, 9

[47] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Avformer: Injecting vision into frozen speech models for zero-shot av-asr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22922–22931, 2023. 9

[48] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 4

[49] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023. 9

[50] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 9

[51] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4

[52] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 9

[53] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 4

[54] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *ICCV*, 2023. 3

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 4

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[57] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019. 9