MotionComposer: Enhancing Rhythmic Music Generation with Adaptive Retrieval Reference

1st Jinting Wang*
Artificial Intelligence
HKUST (GZ)
Guangzhou, China
jwang644@connect.hkust-gz.edu.cn

2nd Li Liu†
Artificial Intelligence
HKUST (GZ)
Guangzhou, China
avrillliu@hkust-gz.edu.cn

3rd Jun Wang
Tencent AI Lab
Tencent
Shenzhen, China
junjunmin@gmail.com

***Abstract*—With the rise of the AIGC era, rhythmic music generation has extensive applications, particularly with the surge in motion video creation. However, generating music that is rhythmically synchronized and stylistically aligned with motion video presents significant challenges. Although existing methods have made progress, they still face difficulties in producing high-quality long-term music, particularly when addressing complex rhythmic patterns and maintaining style-consistent musical chords. In this work, we present *MotionComposer*, a novel retrieval-augmented, easy-to-hard training approach designed to enhance rhythmic music generation. By leveraging the inherent alignment between motion rhythms and music beats, we first tackle the simpler task of beat prediction with BeatNet, which predicts music beats by analyzing motion patterns. To address the complex musical chord generation, we propose ChordNet, a retrieval-augmented network that integrates external data to enrich chord generation. Additionally, to minimize the impact of irrelevant retrievals, we design RAGate, a retrieval adaptive module that selectively filters out low-relevance retrieval references during the retrieval process. Extensive experiments across three scenarios (*i.e.,* dance, figure skating, and floor exercise) demonstrate that our approach significantly enhances video soundtrack generation, achieving new state-of-the-art performance. Our project is available at https://beria-moon.github.io/Soundtrack-your-Motion/.**

***Index Terms*—Music Generation, Adaptive Retrieval, Diffusion Model, Motion video Soundtrack.**

I. Introduction

Video soundtrack generation focuses on creating music that is precisely tailored to motion videos, ensuring alignment in both rhythm and style. This technology has broad applications across multimedia platforms, including social media and interactive entertainment. Unlike other forms of conditional music generation, such as text-to-music generation [1]–[4], which generate melodies based on global musical attributes, video-conditioned music generation presents greater challenges due to its complex temporal dynamics and the need to maintain stylistic correlations between audio and visual elements.

Music composition centers around two fundamental components: rhythm and chords. While predicting music beats from motion rhythms is feasible [5], [6], generating music that ensures both rhythmic alignment and stylistic consistency remains a significant challenge. MIDI-based methods [6]–[8], which rely on structured symbolic representations, struggle to capture the full complexity and diversity of music. Approaches like CMT [9] and CDCD [10] aim to align video with rhythm but fall short in fine-grained temporal details, leading to sub-optimal rhythmic generation. Advanced models, such as D2M-GAN [11] and the work of Li *et al.* [12], can generate complex music but are constrained by the length of the generated sequences. LORIS [5] attempts to condition chord generation using genre labels but faces challenges in creating stylistic musical chords due to the "one-to-many" problem, where a single genre can correspond to a wide variety of chord progressions. In contrast, Retrieval-Augmented Generation (RAG) [13]–[15] offers a more dynamic alternative by leveraging external data repositories to supplement information. This approach enables models to adapt to diverse stylistic chords with minimal effort, providing a flexible and efficient solution. However, a key challenge lies in seamlessly integrating the retrieved data with the original input without introducing noise or irrelevant information [16]–[18].

In this work, we propose **MotionComposer**, a novel two-stage generation framework enhanced with an adaptive RAG technique, as shown in Fig. 1. Unlike LORIS [5], which directly integrates multiple conditions that may interfere with one another, our approach focuses on rhythmic music generation by separately addressing two key elements: **music beats** and **music chords**. In the first stage, we introduce BeatNet, which determines the music beats by learning the temporal coherence with motion rhythm. This ensures rhythmic alignment between the music beat and motion rhythm, effectively preventing stylistic information from disrupting rhythmic patterns. In the second stage, we design a retrieval-augmented ChordNet to compose appropriate chords based on the visual style and retrieved chord reference, enabling a progressive generation process. To further enhance stylistic chord generation, we develop RAGate, which dynamically retrieves the most relevant examples based on the input visual condition. This deepens the understanding of the music style, resulting in more consistent and fitting chord generation.

In summary, the main contributions of this work are as follows: **1)** We develop MotionComposer, a novel retrieval-augmented, easy-to-hard generation framework that achieves both rhythmic alignment and stylistic consistency in rhythmic

*Work done during the internship at Tencent AI Lab.
†Corresponding Author

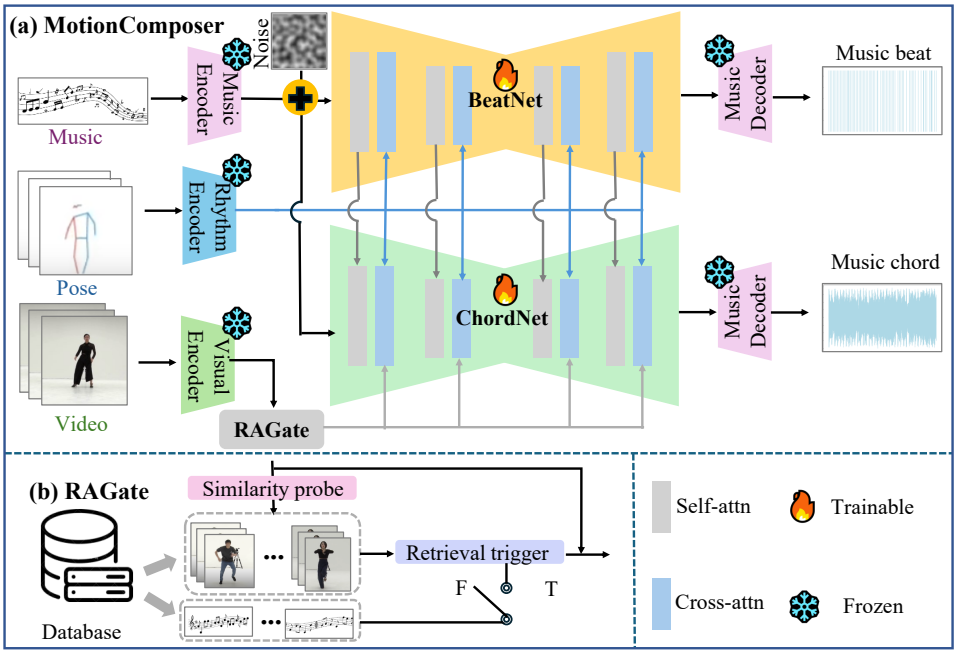

Fig. 1. (a) The architecture of the proposed MotionComposer. To implement an easy-to-hard generation approach, we first develop **BeatNet** to estimate music beats conditioned on dance rhythm. Following this, **ChordNet**, enhanced with **RAGate**, is designed to generate complex music chords. (b) **RAGate** selects adaptive retrieval references to improve the quality and relevance of chord generation.

music generation; **2)** We propose RAGate, an adaptive retrieval module that selectively filters out irrelevant information, thereby improving the quality of the generated music and enhancing overall performance; **3)** Extensive experiments across various scenarios have been conducted. The results show that MotionComposer outperforms current state-of-the-art (SOTA) methods in both rhythmic alignment and stylistic consistency.

II. Related Work

A. Dance-to-Music Generation

Several studies have explored the dance-to-music generation. Dance2Music [19] utilizes a local dance similarity matrix to generate piano music with five notes, but it struggles with rhythmic alignment. RhythmicNet [8] proposes a three-stage method to achieve better rhythmic alignment but overlooks the importance of music chord matching. D2M-GAN [11] employs a pre-trained JukeBox [20] to generate multi-track music from dance videos, but it is limited in terms of music length and quality due to the costly pre-trained music encoder. CDCD [10] maximizes the mutual information of dance and music using contrastive learning and generates music with a diffusion model. LORIS [5] advances long-term music generation using a context-aware latent diffusion model (LDM). Both CDCD and LORIS rely on the dance genre label as a stylistic condition to infer music chords; however, this approach is a coarse-grained constraint for generating long-term suitable music chords. Instead of using genre labels, DanceComposer [6] employs a unified style feature space to learn dance-music style relations and then uses a progressive conditional music generator to produce multi-track MIDI in two stages. However, since each stage is trained separately with independent datasets and training objectives, the overall performance is limited, and the MIDI format music lacks diversity.

B. Retrieval-augmented Audio Generation

With direct access to human-written references, retrieval-augmented generation (RAG) has made significant strides across a variety of applications [21]–[24]. Recent efforts in audio generation [25] have also explored RAG techniques to enhance the generation process. For instance, Huang *it al.* [26] introduce a large number of concept compositions by opening up the usage of retrieval audios to alleviate data scarcity. Yuan et al. [25] use retrieved audio-text pairs as supplementary information to enhance the modeling of low-frequency audio events. In the retrieval process, RAG selects samples with the highest scores from the retrieval database as references. However, if the retrieved references are irrelevant, this can lead to misguided responses and hinder the model's ability to utilize its intrinsic knowledge effectively [27]. To achieve more accurate and reliable generation results, we introduce an adaptive retrieval reference approach in the generation process.

III. Method

A. BeatNet for Music Beat Generation

Given the inherent temporal consistency between music beats and visual rhythm, it is intuitive to predict music beats by detecting visual rhythm, which is defined as a sudden deceleration of motion or a dramatic change in direction [28]. For motion estimation, 2D poses $P(t, j, x, y)$ are extracted from motion video frames using OpenPose [29], where t

and j denote the frame number and joint index, respectively, and x and y represent the joint coordinates. To accurately estimate motion amplitude and strength, we follow LORIS [5] by utilizing a directogram to represent motion changes. We then extract the impact envelope from the directogram and apply a peak-picking strategy [30] to simplify the continuous curves into discrete binary codes for conditional generation. This process yields a binary vector $\mathbf{C_r} \in \mathbb{R}^{T\times 1}$, where a value of 1 indicates that the corresponding time step is a rhythm point, and T represents the total number of time steps. After obtaining the rhythm condition, a latent diffusion model (LDM) is employed for conditioned beat generation. The rhythm condition $\mathbf{C_r} \in \mathbb{R}^{T\times 1}$ is fed into the cross-modal attention module in the intermediate layers of BeatNet to interact with the latent embeddings. The training of BeatNet is guided by the following objective function:

$$L_B(\theta) = \mathbb{E}_{Z_m,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(Z_t, t, \mathbf{C_r})\|_2^2\right], \quad (1)$$

where Z_m represents the extracted music embedding by the music encoder, Z_t is a standard Gaussian distribution obtained by injecting Gaussian noise ϵ into Z_m, and t denotes the denoising time steps. To supervise the generation of music beats, we apply binary cross-entropy loss to measure the difference between the generated music beats and the ground truth. The music beats are detected using the onset detection function from the *librosa* toolbox[1].

B. RAGate-augmented ChordNet for Music Chord Generation

To ensure rhythmic alignment, it is crucial that the dance style and music chords harmonize, conveying a shared feeling tone. This stylistic coherence is essential for expressive video soundtrack generation but can be difficult to define explicitly. To address this challenge, we leverage external knowledge to assist in music chord generation. Traditional RAG selects the top-k similar neighbors as retrieved information. However, the reliability of this retrieved information can be problematic, as irrelevant data may lead to inaccurate results. We propose **RAGate**, an adaptive retrieval module that evaluates the relevance of retrieved data thereby improving the accuracy of the generation process.

As shown in Fig. 1(b), ChordNet processes two parallel style conditions: a visual input $\mathbf{C_v}$, representing low-level style information, and a retrieval reference $\mathbf{C_{ar}}$, representing high-level style information. The visual embedding $\mathbf{C_v}$ is obtained using the I3D model [31], while the retrieval reference $\mathbf{C_{ar}}$ is selected by the RAGate module. Specifically, a **similarity probe** is first performed on the visual embedding to retrieve the most similar candidate:

$$S_{\text{highest}} = \max\langle \mathbf{C_v}, \mathbf{C_v^i} \rangle,$$

where $\langle\cdot,\cdot\rangle$ represents the cosine similarity between two feature vectors, and $\mathbf{C_v^i}$ is the visual feature of the i_{th} sample in the retrieval database. To eliminate irrelevant information, a **retrieval trigger** is employed, setting a relevance threshold to filter the low-relevant references. When the similarity score between the retrieved candidate and the visual query exceeds the threshold, the paired music features $\mathbf{Z_{rm}}$ are extracted and used as the reference $\mathbf{C_{ar}}$ in the cross-attention module. If the score falls below the threshold, a zero vector is used instead, allowing the model to rely on its internal knowledge to generate the music chords. This process can be formalized as:

$$\mathbf{C_{ar}} = \begin{cases} \mathbf{Z_{rm}} & \text{if } S_{\text{highest}} \geq \text{threshold}, \\ \mathbf{0} & \text{if } S_{\text{highest}} < \text{threshold}. \end{cases}$$

After obtaining the retrieval reference, the retrieval-augmented ChordNet is trained using the following loss function:

$$L_C(\theta) = \mathbb{E}_{Z_m,\epsilon,t}\left[\|D(Z_m, t, C_{ar}, C_v, \hat{Z}_{mb}) - Z_m\|_2^2\right], \quad (2)$$

where $\hat{Z}_{mb}$ denotes the predicted music beat from BeatNet, and D represents the music decoder.

IV. Experiments

A. Datasets

Following the previous work, we use the dance dataset AIST++ [32], the Figure skating dataset, and the floor exercise dataset collected in LORIS [5] to evaluate the effectiveness of our proposed method.

B. Experimental Settings

The music sampling rate is set to 22,050 Hz for a duration of 25 seconds. Both BeatNet and ChordNet use audio-diffusion [33] as their backbone. We use AdamW as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.96$, and a weight decay of 4.5×10^{-2}. The entire framework is optimized jointly. We train our model for 200 epochs on the dancing dataset, 200 epochs on the floor exercise dataset, and 250 epochs on the figure skating dataset, using one NVIDIA A6000 GPU. For music sampling, we employ classifier-free guidance [34] to perform conditional generation with a guidance scale $w = 20$. During inference, we use 50 diffusion steps to balance music quality and inference speed. The **threshold** of RAGate is set based on the lowest similarity score of improved samples between Two-stage and Two-stage+RAG in ablation studies.

C. Evaluation Metrics

To quantitatively evaluate the generated music, we utilize several objective metrics, assessing from two aspects: rhythmic consistency and stylistic consistency. **Rhythmic Alignment**. Following previous work [5], we use the refined Beats Coverage Scores (BCS) and Beats Hit Scores (BHS) to evaluate the rhythmic alignment between the dance and the generated music. We integrate these assessments using the F1 scores of BCS and BHS and report their standard deviations (referred to as CSD and HSD, respectively) to evaluate generative stability. **Stylistic Consistency**. In line with DanceComposer [6], we assess Genre Accuracy (GAC) to evaluate the consistency of stylistic chords with respect to the dance genre.

[1]https://librosa.org/doc/main/generated/librosa.onset.onset_detect.html

D. Comparison with SOTAs

We compare our framework with six baselines: Foley [7], Dance2Music (D2M) [19], CMT [9], D2M-GAN [11], CDCD [10], and LORIS [5]. Comparisons with RhythmicNet [8], Li *et al.* [12], and DanceComposer [6] are not conducted due to the unavailability of their source code. Tables I, II, and III present the comparison results for the dance, floor exercise, and figure skating datasets, respectively. The results show that our proposed method consistently outperforms competitors across all evaluations. In terms of rhythmic consistency, our method excels in all metrics, demonstrating that the easy-to-hard generation pipeline effectively enhances rhythmic alignment. For stylistic consistency, the incorporation of retrieval references through RAGate significantly improves performance in GAC, underscoring the effectiveness of using external information as supplementary style signal.

TABLE I
QUANTITATIVE EVALUATION ON AIST++ DATASET.

Method	BCS↑	CSD↓	BHS↑	HSD↓	F1↑	GAC↑
Foley [7]	88.2	18.2	44.2	17.2	57.5	8.2
D2M [19]	90.4	16.1	41.6	19.2	64.7	9.7
CMT [9]	93.1	14.2	48.1	17.2	58.4	9.9
D2M-GAN [11]	94.6	9.2	86.3	12.3	92.6	20.3
CDCD [10]	94.2	10.8	84.1	13.9	91.4	14.6
LORIS [5]	96.1	9.3	90.4	10.8	93.6	10.2
MotionComposer	**98.8**	**5.2**	**99.4**	**2.6**	**98.7**	**47.8**

TABLE II
QUANTITATIVE EVALUATION ON FLOOR EXERCISE DATASET.

Method	BCS↑	CSD↓	BHS↑	HSD↓	F1↑
Foley [7]	38.7	27.4	34.3	25.8	37.2
D2M [19]	42.5	24.4	52.1	27.4	46.3
CMT [9]	46.4	30.1	57.4	29.8	51.3
D2M-GAN [11]	47.6	26.5	56.4	28.2	50.3
CDCD [10]	48.2	24.1	57.5	26.2	52.4
LORIS [5]	61.7	26.3	63.6	24.3	59.2
MotionComposer	**67.5**	**21.7**	**73.6**	**15.4**	**67.6**

TABLE III
QUANTITATIVE EVALUATION ON FIGURE SKATING DATASET.

Method	BCS↑	CSD↓	BHS↑	HSD↓	F1↑
Foley [7]	33.6	27.1	23.6	17.4	27.9
D2M [19]	38.3	22.6	29.2	19.9	31.4
CMT [9]	39.2	22.8	46.4	27.5	50.7
D2M-GAN [11]	43.1	24.6	49.7	27.4	43.2
CDCD [10]	44.3	25.6	42.7	19.4	46.8
LORIS [5]	54.6	17.8	58.4	20.3	56.6
MotionComposer	**59.5**	**10.7**	**64.2**	**13.1**	**63.7**

E. Ablation Studies

We further evaluate the necessity of each model component, with results presented in Table IV. To investigate the effectiveness of the two-stage generation approach, we compare the framework's performance with one-stage and two-stage models. In this context, "one-stage" refers to using multi-conditions directly in the LDM to generate rhythmic music. The results indicate that the two-stage model performs better in both rhythmic and stylistic alignment, demonstrating that the easy-to-hard generation strategy is crucial for mitigating the interference of style information. We also analyze the impact of the genre condition, finding that it weakens the model's performance due to the one-to-many problem. To explore the effectiveness of the RAGate, we compare traditional RAG with our proposed RAGate. We observe that filtering irrelevant retrieval information with RAGate helps the model balance between intrinsic and external knowledge, leading to improved generation.

TABLE IV
ABLATION STUDIES ON AIST++ DATASET.

Method	BCS↑	CSD↓	BHS↑	HSD↓	F1↑	GAC↑
One-stage	96.2	9.6	86.7	12.4	91.1	22.6
Two-stage	96.7	9.2	96.4	11.7	95.3	31.7
Two-stage+Genre	96.4	8.8	85.3	10.9	86.6	26.4
Two-stage+RAG	97.4	8.6	94.2	7.9	94.9	32.7
MotionComposer	**98.8**	**5.2**	**99.4**	**2.6**	**98.7**	**47.8**

F. Case Study

In Fig. 2, we visualize the rhythm of the dance and the corresponding music, alongside their mel-spectrograms, for a qualitative analysis of rhythmic alignment. The generated music beats produced by our method show a clear alignment with both the ground truth and the visual rhythm. The comparison between the generated and ground truth music mel-spectrograms reveals that the generated music exhibits a similar crest distribution to the ground truth.

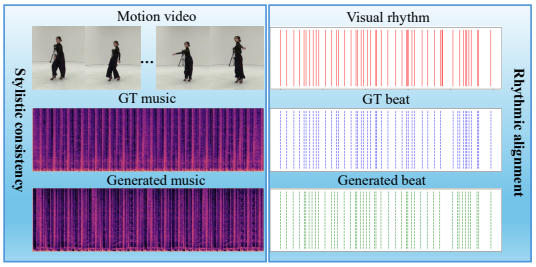

Fig. 2. Example visualization of rhythms and musical mel-spectrograms, demonstrating that our model successfully generates music with well-aligned rhythms and coherent chords.

V. CONCLUSION

In this work, we present a novel retrieval-augmented two-stage generation pipeline to achieve long-term, high-quality video soundtrack generation with both rhythmic alignment and stylistic consistency. Comparisons with SOTA methods across various performance metrics demonstrate that our method significantly improves the quality of video soundtrack generation. In future work, we plan to extend this approach by integrating large-scale music generation models and exploring the potential of our framework in zero-shot scenarios.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62471420 and 62101351), the Tencent AI Lab (RBFR2023014), the Guangzhou-HKUST (GZ) Joint Funding Program (Grant No. 2023A03J0008), and the Education Bureau of Guangzhou Municipality.

REFERENCES

[1] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1206–1210.

[2] 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," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 286–290.

[4] Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis, "Investigating personalization methods in text to music generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1081–1085.

[5] Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Yu Qiao, "Long-term rhythmic video soundtracker," in *International Conference on Machine Learning (ICML)*, 2023.

[6] Xiao Liang, Wensheng Li, Lifeng Huang, and Chengying Gao, "Dance-composer: Dance-to-music generation using a progressive conditional music generator," *IEEE Transactions on Multimedia*, pp. 1–14, 2024.

[7] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba, "Foley music: Learning to generate music from videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 758–775.

[8] Kun Su, Xiulong Liu, and Eli Shlizerman, "How does it sound?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29258–29273, 2021.

[9] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan, "Video background music generation with controllable music transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2037–2045.

[10] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan, "Discrete contrastive diffusion for cross-modal music and image generation," *arXiv preprint arXiv:2206.07771*, 2022.

[11] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov, "Quantized gan for complex music generation from dance videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 182–199.

[12] Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu, "Dance-to-music generation with encoder-based textual inversion of diffusion models," *arXiv preprint arXiv:2401.17800*, 2024.

[13] Uday Kamath, Kevin Keenan, Garrett Somers, and Sarah Sorenson, "Retrieval-augmented generation," in *Large Language Models: A Deep Dive: Bridging Theory and Practice*, pp. 275–313. Springer, 2024.

[14] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 17754–17762.

[15] Alireza Salemi and Hamed Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2395–2400.

[16] Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li, "Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation," *arXiv preprint arXiv:2406.19215*, 2024.

[17] Huanshuo Liu, Hao Zhang, Zhijiang Guo, Kuicai Dong, Xiangyang Li, Yi Quan Lee, Cong Zhang, and Yong Liu, "Ctrla: Adaptive retrieval-augmented generation via probe-guided control," *arXiv preprint arXiv:2405.18727*, 2024.

[18] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng, "Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models," *arXiv preprint arXiv:2402.10612*, 2024.

[19] Gunjan Aggarwal and Devi Parikh, "Dance2music: Automatic dance-driven music generation," *arXiv preprint arXiv:2107.06252*, 2021.

[20] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[21] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," *arXiv preprint arXiv:2209.14491*, 2022.

[22] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan, "Lift yourself up: Retrieval-augmented text generation with self-memory," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani, "Fid-light: Efficient and effective retrieval-augmented text generation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 1437–1447.

[24] Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin, "Retrieval-generation synergy augmented large language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11661–11665.

[25] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang, "Retrieval-augmented text-to-audio generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 581–585.

[26] 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*. PMLR, 2023, pp. 13916–13932.

[27] Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang, "Improving retrieval augmented language model with self-reasoning," *arXiv preprint arXiv:2407.19813*, 2024.

[28] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz, "Dancing to music," *Advances in neural information processing systems*, vol. 32, 2019.

[29] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[30] Sebastian Böck, Florian Krebs, and Markus Schedl, "Evaluating the online capabilities of onset detection methods.," in *ISMIR*, 2012, pp. 49–54.

[31] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[32] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13401–13412.

[33] Flavio Schneider, "Archisound: Audio generation with diffusion," *arXiv preprint arXiv:2301.13267*, 2023.

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