

MuseChat

Un sistema conversacional de
recomendación de música para videos

Zhikang Dong, Xiulong Liu, Bin Chen, Paweł Polak, Peng Zhang

<https://dongzhikang.github.io/musechat/>

<https://github.com/Dongzhikang/MuseChat-dataset>

¿Qué es un sistema de recomendación musical?

Herramienta que permite a sus usuarios encontrar automáticamente su tipo de música deseada adecuada al contexto proporcionado (canción actual, video, estado de ánimo, etc.)

- Mejora la experiencia del espectador
- Personaliza la música según el contexto
- Ahorra tiempo en la selección manual



Sistemas existentes resuelven el problema



Son expertos en seleccionar pistas que se alinean con la temática de un vídeo, [pero...](#)

Los sistemas ya existentes a menudo descuidan las preferencias individuales de los usuarios



Retos principales

■ *Información disponible*

Falta de datasets que vinculen video, música y texto en un entorno conversacional

■ *Aprendizaje Multimodal*

Se debe combinar texto, video y música en una sola representación

■ *Razonamiento explicativo*

El sistema debe justificar por qué hacer cierta recomendación



Estado del arte: recomendación musical

■ *MusicCaps, MuLan*

Para representación musical.

■ *CLIP y AST*

Para procesamiento de video y audio.

■ *Limitaciones*

No consideran el aspecto conversacional ni explicativo.



Estado del arte: sistemas recomendacionales y LLM

■ *Chat-rec*

Recomendación conversacional

■ *GPT, Vicuna*

Generación de lenguaje

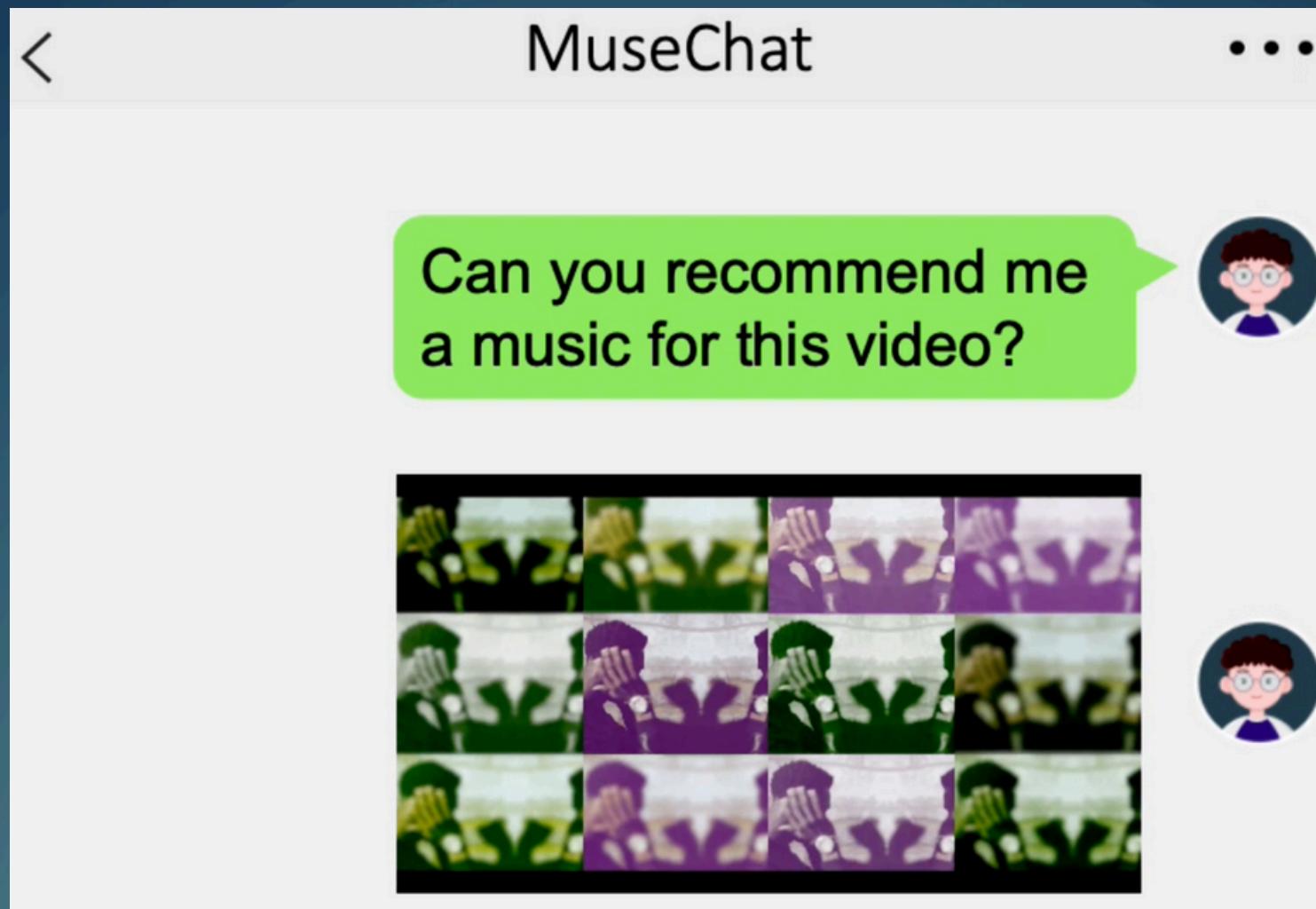
■ *Limitaciones*

Integrar LLMs y modelos
multimodales para tareas
explicativas



MuseChat

Propone un mecanismo basado en feedback conversacional para mitigar estas limitaciones



< MuseChat ...

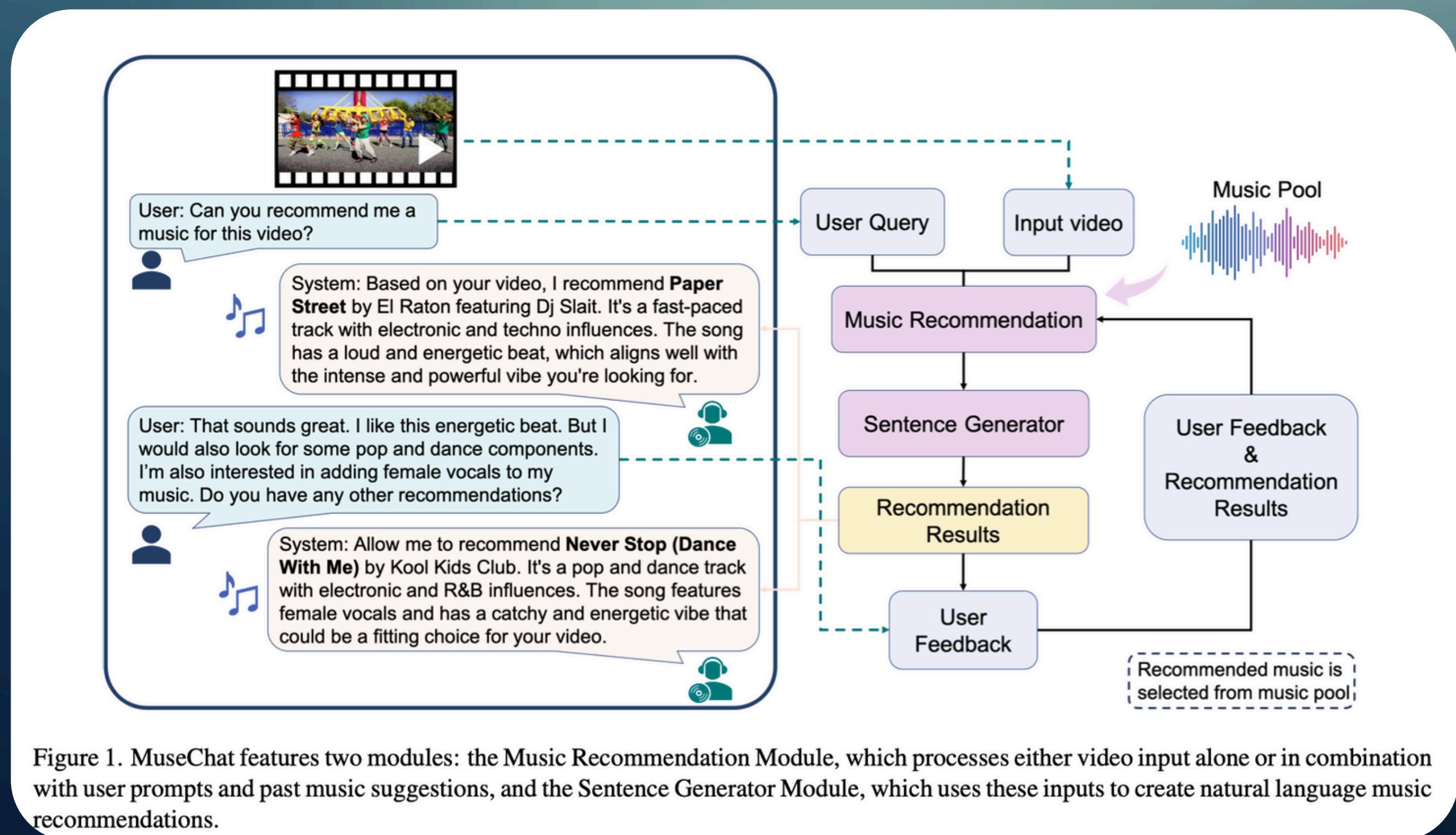





I suggest **Deliverer** by Weston Skaggs. The song combines elements of rock, indie, and alternative with a touch of pop. It features guitar and male vocals, which align with your preferences. The track has a distinctive sound that could be a fitting choice for your video.

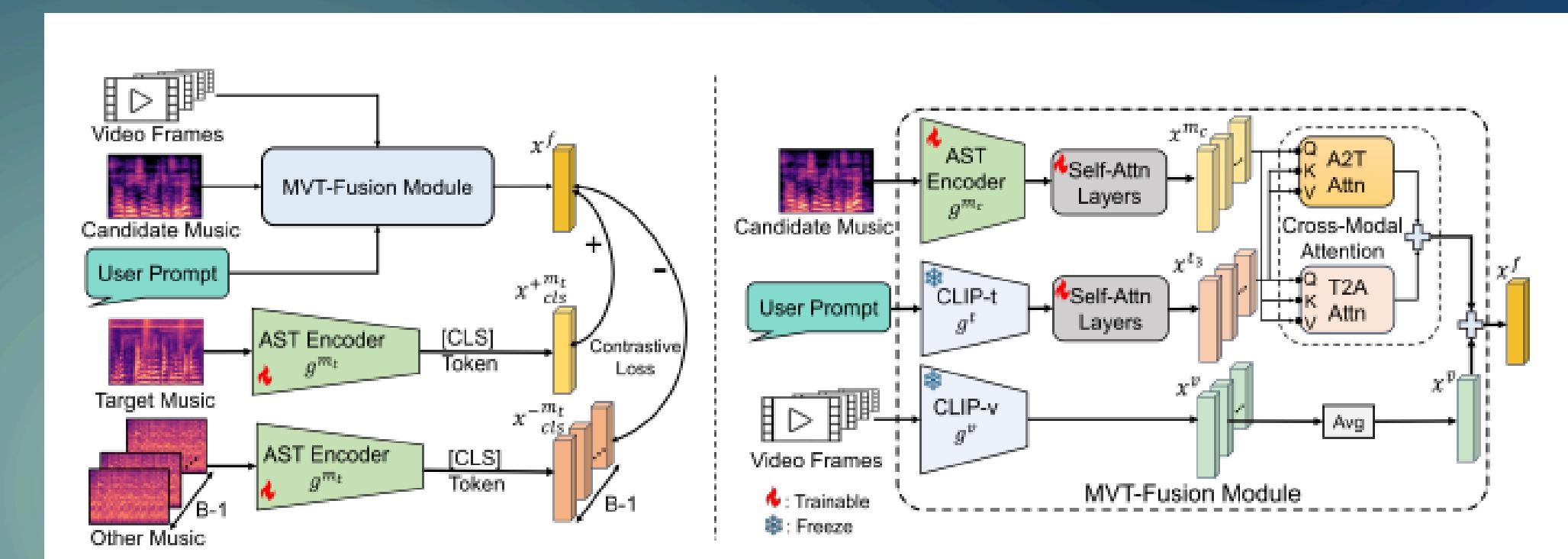
¿Cómo lo resuelve?

Pipeline general



Módulo de Recomendación (MVT-Fusion)

- Modalidades: CLIP + AST (video), BERT (texto), MuLan (música)
- Aprendizaje contrastivo para alinear embeddings
- La fusión multimodal es clave para el matching de contenido



Módulo de Explicación

- Input: embedding musical + metadatos
- Modelo Vicuna-7B con LoRa
- Output: Explicación en lenguaje natural del por qué de la recomendación



Resultados experimentales

Detalles de implementación: Training Set

- Compuesto por **88,000 clips** de videos musicales alocados al training set, **120s c/u.**
- Cada video se divide en **12 fragmentos**, compuestos por segmentos de **10s c/u.**
- De cada segmento, se capturan **5 frames por segundo.**



Resultados experimentales

Detalles de implementación: Training Process

Existen 4 elementos clave por cada muestra del dataset:

Clip principal

- Un segmento de video que necesita una recomendación musical. (10s)

Clip Target

- La pista musical considerada como la más adecuada para acompañar ese video. (10s)

Second Candidate

- Una pista alternativa que no es tan adecuada, usada para entrenamiento contrastivo.

User Prompt

- Una instrucción o preferencia expresada por el usuario para guiar la recomendación.



Resultados experimentales

Detalles de implementación: Training Process

Extracción de features
para **video y texto**



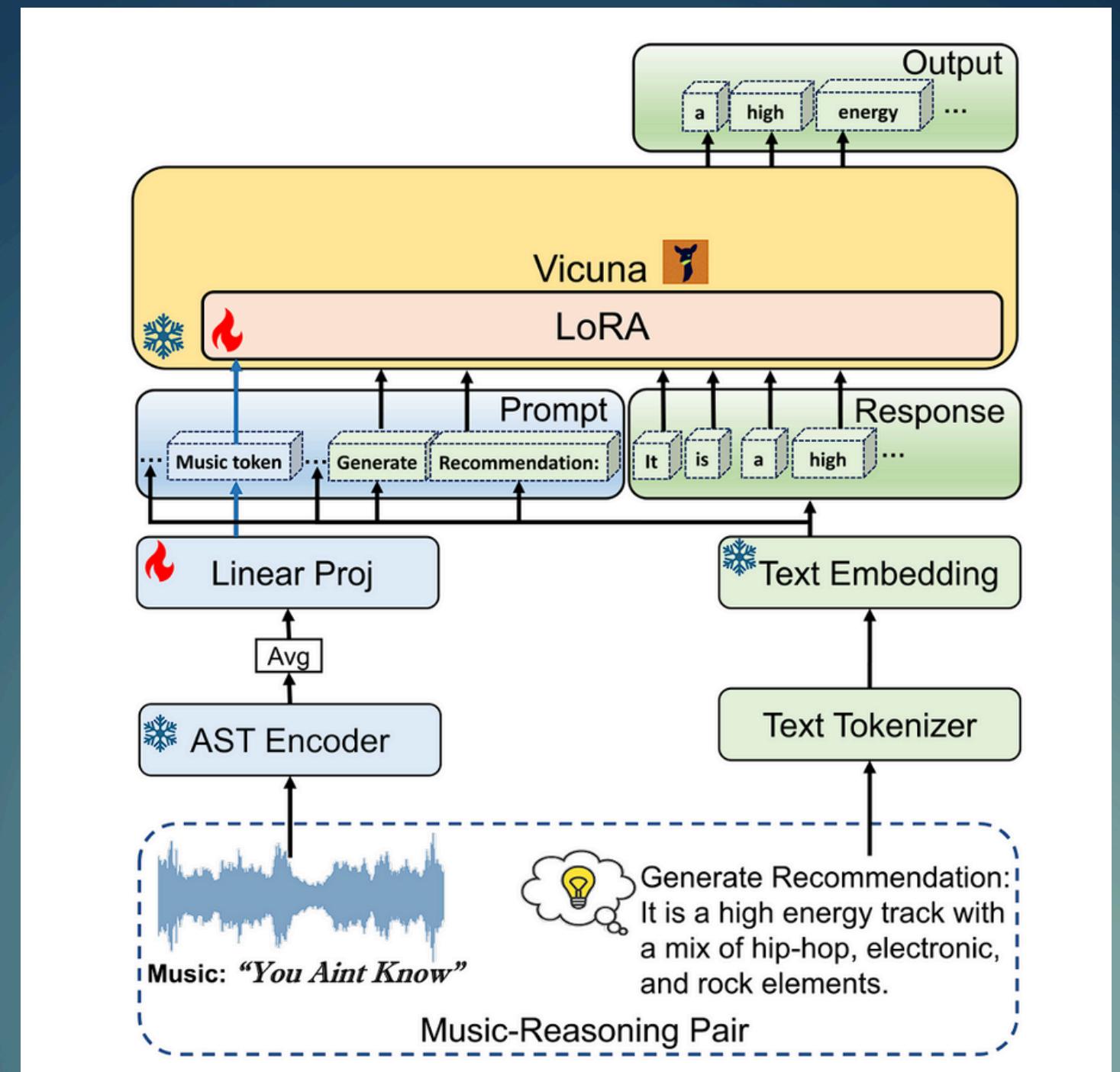
OpenAI Clip Model

Extracción de features
para **audio**



AST Model

Las características extraídas de video, música y texto
se proyectan a vectores de 256 dimensiones
mediante capas lineales



Resultados experimentales

Detalles de implementación: Baseline Comparisons

4 modelos base para comparar con Musechat

- **Music Video Pretrained (MVP)**
Sigue la misma arquitectura pero entrenando con Youtube-8M, en vez de datos propios
- **Sum-Fusion**
Fusión simple sumando vectores de video, música y texto.
- **Self-Attn Fusion**
Añade capas de self-attention antes de la fusión, para capturar dinámicas internas
- **Cross-Attn Fusion**
Usa atención cruzada entre texto y música, sin self-attention

Model	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑	SR ↑
Sum-Fusion	17	10.58	28.47	40.19	21.70
MVP	7	20.71	48.89	63.14	20.71
Self-Attn Fusion	5	21.40	50.83	65.29	28.37
Cross-Attn Fusion	3	25.71	56.97	71.07	31.48
MuseChat (ours)	2	32.79	63.92	76.53	40.49
Chance	250	0.20	1.00	2.00	0.40

Table 1. Music retrieval results for different baseline models v.s MuseChat.

Resultados experimentales

Detalles de implementación: Ablation Studies

Se evalúa mediante estudios de ablación la importancia de cada modalidad (video, música candidata, texto).

Model	Modality	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑	SR ↑
MVP	Video → Music	7	20.71	48.89	63.14	20.71
MuseChat w/o Video	(Music, Text) → Music	19	8.12	24.53	37.11	8.12
MuseChat w/o Candidate Music	(Video, Text) → Music	5	22.67	51.53	65.42	26.02
MuseChat (ours)	(Video, Music, Text) → Music	2	32.79	63.92	76.53	40.49
Chance	-	250	0.20	1.00	2.00	0.40

Table 2. Ablation Studies: Performance comparison when removing modality branches from MuseChat and training corresponding models from scratch.

Resultados experimentales

Detalles de implementación: Semantic Comparison

Comparaciones de las salidas generadas por los tres modelos usados
usando cuatro métricas automáticas de similitud semántica

Model	Input Modality	BertScore [50] (f1) ↑	AB Divergence [6] ↓	\mathcal{L}_2 Distance ↓	Fisher-Rao Distance [6] ↓
Vicuna-7B	Music Title	0.9453	3.93	0.382	2.11
Vicuna w/ Music	Music Embeddings	0.9526	2.68	0.279	2.02
MuseChat (ours)	Music Title + Embeddings	0.9676	1.51	0.208	1.47

Table 3. Comparison of semantic similarity between output and simulated conversations using various metrics. BERTScore [50] assesses token-level similarity, while AB Divergence, \mathcal{L}_2 Distance, and Fisher-Rao Distance are derived based on InfoLM [6].

Resultados experimentales

Detalles de implementación: Human Evaluation

Correctitud: Nombre de canción/artista

Musicalidad: Descripción de características musicales

Claridad: Coherencia y comprensibilidad

Model	Correctness	Musicality	Clarity	Overall
Vicuna-7B	3.07	2.54	3.60	3.07
Vicuna w/ Music	1.24	3.50	4.05	2.93
MuseChat (ours)	4.63	4.22	4.54	4.46

Table 4. Human evaluation scores for music reasoning outputs.

Conclusiones de los resultados

Contribuciones principales

- **Nuevo conjunto de datos**

Se introdujo un nuevo conjunto de datos diseñado específicamente para recomendaciones musicales guiadas por diálogo y razonamiento en el contexto de videos. 98,206 quartetos! (Video, música original, música del candidato y conversaciones de 2 turnos)

- **Sistema conversacional**

Capacidad de generar explicaciones claras y lógicas para sus recomendaciones musicales

- **Nueva arquitectura**

Se presenta una arquitectura en forma de triodo para el emparejamiento entre música y video, que incorpora entradas textuales como parte del modelo

- **Capacidad de comprensión**

Comprensión profunda de las características musicales mediante un módulo generador de frases (basado en modelos de lenguaje de gran escala (LLM).)



Trabajos relacionados



Etiquetado automático de música

Asignar etiquetas que describen aspectos como emoción, género o tema.

Descripción musical en lenguaje natural libre

Describir pistas musicales usando oraciones completas, no solo etiquetas.

Modelos LLM y multimodalidad

Procesar múltiples tipos de datos (texto, imágenes, video, audio) con modelos de lenguaje.

Sistemas de recomendación conversacionales

Interacción por diálogo con el usuario para entender y adaptar recomendaciones.

Recomendación musical para video

Recomendar música según atributos del video.

Referencias clave

- Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778, 2021.
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI 16, pages 776-794. Springer, 2020.
- 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.

MuseChat

Un sistema conversacional de
recomendación de música para videos

Zhikang Dong, Xiulong Liu, Bin Chen, Paweł Polak, Peng Zhang

<https://dongzhikang.github.io/musechat/>

<https://github.com/Dongzhikang/MuseChat-dataset>