HSPI The Open Access Publisher

JAIRI ISSN 3069-762X JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION

Research Article

Drive-to-Music: Context-Aware Generative Audio for In-Vehicle Experiences

Cosmin Dragoiu* and Nooshin Nabizadeh

Mercedes-Benz Research & Development North America, 2708 Orchard Pkwy, San Jose, CA 95134, USA

More Information

***Corresponding author:** Cosmin Dragoiu, Mercedes-Benz Research & Development North America, 2708 Orchard Pkwy, San Jose, CA 95134, USA, Email: cosmin.dragoiu@mercedes-benz.com

Submitted: October 15, 2026
Accepted: October 22, 2026
Published: October 23, 2026

Citation: Dragoiu C, Nabizadeh N. Drive-to-Music: Context-Aware Generative Audio for In-Vehicle Experiences. J Artif Intell Res Innov. 2026; 2(2): 160-164. Available from: https://dx.doi.org/10.29328/journal.jairi.1001029

Copyright license: © 2026 Dragoiu C, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Check for updates

OPEN ACCESS

Abstract

In-vehicle music can serve as an adaptive interface to enhance driver experience, attention, and well-being. We present Drive-to-Music, a context-aware system that generates music in real time from multimodal driving signals. Using dashcam imagery and vehicle telemetry, the system extracts scene semantics and driving context, maps them to high-level musical descriptors, and conditions generative audio models to produce contextually aligned soundtracks. The architecture combines perception and generative components to translate visual and kinematic inputs into structured musical attributes and synthesize audio with low latency. It supports smooth transitions as driving conditions evolve, and to ensure robustness and deployment readiness, we incorporate constraint-based controls and safety checks across the generation pipeline.

Our results demonstrate the feasibility of real-time, context-aware music generation in automotive settings, providing a foundation for personalized and adaptive in-vehicle audio experiences.

Introduction

AI-driven music generation has become a prominent area of research, with broad applications across diverse do-mains [1-3]. Systems that generate music from text, images, or video inputs are widely used across con-tent creation industries such as digital production and stock music libraries [4]. Despite these advances, the automotive industry remains an unexplored domain for AI music generation technology.

Music, long considered a passive companion during travel, has untapped potential as an adaptive and emotion-ally resonant medium that can enhance both focus and overall experience. Sounds Ride [5], an in-vehicle audio augmented reality system links environmental affordances to musical events using continuous prediction and real-time mixing. Mood-aware music recommendation for drivers [6] infers driver mood and selects music intended to improve comfort and driving behavior. Studies on music effects on driving [7] show that music influences driver's cognition and style.

Despite these advances, most systems rely on selecting, reordering, or remixing existing music, and many evaluations are simulator-based prototypes rather than deployments in real vehicles. In this paper we introduce Drive-to-Music, a novel AI-driven system that dynamically generates personalized music based on real-time vehicle contextual data. Our system leverages dashcam imagery and vehicle telemetry to synthesize custom music tracks. By translating multimodal input into high-level musical descriptors, Drive-to-Music uses generative AI models to create audio output that is synchronized with the driving context.

In this paper, we outline the system architecture, multi-modal input preprocessing and model selection. Our work contributes to the emerging field of context-aware generative music and sets the stage for future in-vehicle AI companions that can enhance driver well-being through personalized, intelligent soundscapes.

System architecture

The Drive-to-Music architecture, shown in Figure 1, consists of several stages that process multimodal vehicle data to produce contextually relevant music alongside complementary media elements such as cover art and song name. The main input to the system is a front-facing camera image that captures the driving scene. Additional vehicle sensor data, such as vehicle speed and outside temperature is used to further enhance the music generation process.

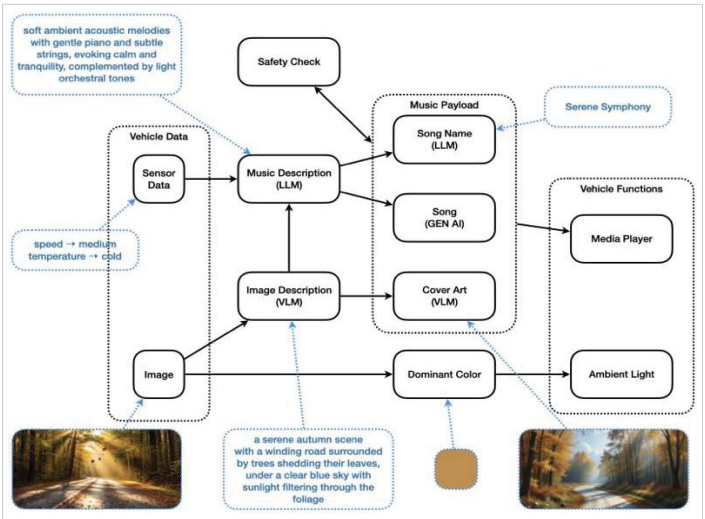


Figure 1: The Drive-to-Music system architecture showcasing the main components and the data flow.

For a true in-vehicle experience, all generative AI models run in the cloud and a dedicated in-vehicle service is responsible for gathering the relevant data from the vehicle, sending it to the cloud, and receiving the generated music and media elements back to the vehicle for playback. For such a setup, the only requirement is an internet connection that enables periodic communications between the vehicle and the cloud. The in-vehicle service collects camera imagery and vehicle telemetry, which are passed to the generative models for scene understanding, musicaldescriptor generation, media synthesis, and safety evaluation. The complete pipeline, from image capture to delivery of the generated music experience, completes in approximately 12 seconds end to end, with the music track itself generated in approximately 6 seconds.

Image Description Stage: To obtain a scene-level understanding of the driving context, we use a vision-language model (VLM) to generate an image description from the camera input. The prompt (see Table 6) is designed to ex-tract information like surroundings, visible landmarks, vegetation, sky conditions, and weather. This description pro-vides a compact semantic representation of the visual environment and serves as the basis for downstream generation.

Music Description Stage: The scene description is passed to a large language model (LLM), which combines it with additional vehicle sensor data (see Table 6) and generates a set of music-oriented descriptors that best match the observed environment. The music descriptors capture at-tributes such as mood, atmosphere, and style, and are used to condition a music-generation model.

Music Generation Stage: Motivated by driver safety, we generate exclusively instrumental music, as research demonstrates that instrumental content reduces cognitive and emotional demands on drivers. To reduce transition frequency while maintaining responsiveness to changing driving conditions, we set the generated track duration to 90 seconds. In parallel with music generation, the system produces additional media elements to provide a more complete playback experience. A cover-art image is generated from the previously extracted scene description using an image-generation model. The objective is to preserve the overall mood and semantic content of the original scene while allowing for a more stylized and flexible visual out-put (see Table 6). A separate LLM call is used to generate a short music title based on the music description.

Dominant Color Extraction Stage: To further enhance the in-vehicle experience, the system also extract the dominant color of the captured scene. This component uses a histogram-based approach. Muted sky regions and low-information gray areas, such as roads and building, are masked. The remaining pixels are mapped to a supported in-vehicle color palette, and the resulting dominant color is extracted. This color is used to dynamically adjust the ambient lighting in the vehicle by setting the interior lighting elements to the extracted color. The ambient light color gets updated with every new song to create a more immersive and harmonious environment that complements the generated music and visual elements.

Safety Check Stage: All AI-generated artifacts are validated against specific safety and quality requirements. The song name is screened with rule-based filters for explicitly prohibited content and evaluated by an LLM configured as a structured safety classifier. The generated mu-sic is evaluated with deterministic digital-signal-processing (DSP) checks, including integrated loudness, true-peak level, and short-term loudness jumps computed following the calibrated thresholds of ITU-R BS.1770-5 [8] and EBU R128 [9], combined with the Audio Setpretrained Efficient AT model [18] to detect unexpected vocal content in the otherwise instrumental tracks. The cover art is screened to ensure that it does not contain offensive visual content.

In case a component fails the safety checks, a backup mechanism is triggered to replace the failed component with a pre-generated alternative.

To maintain continuous playback, generation of the next track begins once the current 90-second track reaches the 70-second mark, since the full pipeline, from image capture to completion of the safety evaluation, takes approximately 12 seconds. The remaining 20-second window is sufficient to complete the generation and safety check process.

Model selection and evaluation

Across all components, model and algorithm selection is guided by key system-level constraints, including minimizing end-to-end latency, maintaining acceptable output quality, and reducing overall resource usage, as well as limiting deployment and API cost. The estimation of different met-rics is based on evaluating on average about ten samples per model.

To select the best VLM for our system, we conducted a benchmarking study of several candidate models, including Liquid AI LFM 2.5 VL 1.6B [11], Qwen 3 VL 2B Instruct [12], and Gemini 2.5 Flash Lite [13]. We evaluated these models based on their inference latency, output quality, including the CLIP score [14], and resource usage. The latency is measured as the average time to process a single request. The results are summarized in Table 1.

All models provide comparable output quality, with CLIP scores in the 26–27 range. However, there are significant differences in inference latency. Gemini 2.5 Flash Lite achieves the lowest latency at approximately 0.6 seconds per request, while LiquidAI LFM2.5 VL 1.6B and Qwen3 VL 2B Instruct require around 1.2 and 3.9 seconds, respectively. Based on these results, we selected LiquidAI LFM2.5 VL 1.6B as the scene-description component of our pipeline. For the music description generation stage, we bench-marked the following LLM candidates: LiquidAI LFM 2.5 1.2B Instruct [11], Qwen 3.5 2B [12], and Gemini 2.5 Flash Lite [13]. Output quality was assessed on ten outputs per model, each independently rated for stability and relevance on a 5-point scale. All models provide high output quality, with LiquidAI LFM 2.5 1.2B Instruct and Gemini 2.5 Flash Lite achieving a perfect score of 5/5. In terms of latency, LiquidAI LFM 2.5 1.2B Instruct is the fastest at approximately 0.7 seconds per request, followed by Gemini 2.5 Flash Lite at 1.2 seconds, and Qwen 3.5 2B at 2.4 seconds.

The complete benchmark results are summarized in Table 2. Based on these results, the LiquidAI LFM 2.5 1.2B In-struct model stands out and therefore was selected as the music description generation component of our pipeline. The same model is also used for the song name generation, as it provides a good balance of latency and output quality for both tasks.

The music generation stage is at the core of the Drive-to-Music system and is the most computationally intensive component. The selection of the music generation model relies on a combination of factors. We prioritize models that can produce instrumental music with no vocals, as vocals can be a distracting factor for drivers. We also require that models be able to generate songs of 90 seconds or longer to minimize the frequency of song transition and to provide a more immersive experience.

From all available models, we selected ElevenLabs [15], Mubert [16] and Stable Audio 2.5 [17] as the most promising candidates for further evaluation. We benchmarked the selected models against latency and several audio quality criteria, which are listed in Table 3.

The MuQ-MuLan score [18] measures the prompt–audio semantic alignment. It relies on a joint music–text embedding model trained contrastively for music-text similar-ity. The Technical Audio Quality Score (TAQS) is a cus-tom 0–100 score that summarizes the engineering quality of an audio track based on standardized loudness and true-peak measurements defined in ITU-R BS.1770-5 [8] and EBU R128 [9], combined with clipping, and sudden loud-ness jumps. A higher score means the track is technically cleaner, safer for playback, and more production-ready.

Based on these criteria, we selected Stable Audio 2.5, which achieved the highest MuQ-MuLan score and the low-est latency among the evaluated models, while its TAQS score was the second highest and remained suitable for consistent playback, as the music generation model for our sys-tem.

In addition to the audio track, the system generates a companion cover art image. The goal is to preserve the essence of the captured front-facing camera scene without reproducing the original frame. To meet these requirements, we use a dedicated VLM that is optimizied for image gener-ation. We tested several models, including Stable Diffusion 3.5 Large Turbo [19], Qwen Image [20] and Gemini 2.5 Flash Image [13] (Table 4).

Similar to previous stages, the models were evaluated based on their inference latency, output quality, including the GenEval score [21], and resource usage. The GenEval score

Table 1: Benchmark results for image understanding VLM candidates. The CLIP score is used to evaluate the semantic alignment between the generated description and the input image.

Model	Capacity	Latency	CLIP Score
LiquidAI LFM 2.5 VL 1.6B	1.6B	1.2s *	26.43
Qwen 3 VL 2B Instruct	2B	3.9s *	26.59
Gemini 2.5 Flash Lite	150B	0.6s †	27.25

* Local hosting on NVIDIA T4 (16GB), †API call

Table 2: Benchmark results for music description LLM can-didates. The output quality is evaluated based on stability and relevance based on a set of curated test cases on a 5-point scale.

Model	Capacity	Latency	Quality
LiquidAI LFM 2.5 1.2B Instruct	1.2B	0.7s *	05-May
Qwen 3.5 2B	2B	2.4s *	04-May
Gemini 2.5 Flash Lite	150B	1.2s †	05-May

*Local hosting on NVIDIA T4 (16GB), †API call.

Table 3: Benchmark results for music generation model can-didates. The MuQ-MuLan score measure the prompt–audio semantic alignment. A high score means the music matches the text semantically. The TAQS score summarizes the en-gineering quality of an audio track. A higher score means a higher quality.

Model	Latency *	MuQ-MuLan Score	TAQS Score
ElevenLabs	15.7s	0.09	72.22
Mubert	7.4s	0.29	91.26
Stable Audio 2.5	6.2s	0.36	78.03

* API call

Table 4: Benchmark results for cover art VLM candidates. The GenEval score measures how reliably a text-to-image model fol-lows the prompt composition.

Model	Capacity	Latency	GenEval Score
Stable Diffusion 3.5 Large Turbo	8B	1.4s *	0.66
Qwen Image	20B	40.4s *	0.91
Gemini 2.5 Flash Image	150B	6.5s †	0.96

*Local hosting on NVIDIA A100 (80GB), †API call

measures how reliably a text-to-image model follows prompt composition, like object presence, count, color, position, and attribute binding, using automatic detectors in-stead of human judgment. Gemini 2.5 Flash Image and Qwen Image achieve the highest GenEval scores, with 0.96 and 0.91, respectively, while Stable Diffusion 3.5 Large Turbo achieves a score of 0.66.

In terms of latency, Stable Diffusion 3.5 Large Turbo is the fastest model, with an average latency of 1.4 seconds per request when running locally, followed by Gemini 2.5 Flash Image at 6.5 seconds, and Qwen Image at more than 40 seconds. Based on these results, we selected Stable Diffusion 3.5 Large Turbo as the cover art generation component. Al-though it achieves a lower GenEval score than the other two candidates, it offers the most favorable overall trade-off among latency, quality, resource usage, and deployment cost.

We also conducted a survey of the music generation per-formance in different driving scenarios. Of all 50 partici-pants, 68% expressed interest in music that adapts to the driving situation, particularly during long drives, while 72% preferred to retain control over the genre and instrumentation. These findings suggest that an adaptive music system that preserves user control through personalized musical preferences is desirable.

Conclusion

In this paper, we introduced Drive-to-Music, a context-aware system that generates personalized music from multi-modal driving signals. By combining vision-language models for scene understanding, language models for mapping context to musical descriptors, and generative audio models for synthesis, the system translates driving conditions into adaptive soundtracks in real time, with representative examples provided in Table 5.

Through systematic evaluation, we selected components that balance latency, quality, and resource constraints, demonstrating the feasibility of integrating generative AI into in-vehicle experiences. Limited user studies showed positive reception, with participants appreciating the con-textual relevance and personalization of the generated mu-sic. Future work will extend the system with additional signals, such as traffic, navigation, and user preferences, to further enhance personalization and contextual relevance.

A. Examples of Generated Content

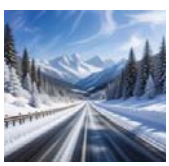
Camera Image	Vehicle Data	Image Description	Music Description	Cover Art
	low speed and warm outside	a bustling urban street scene with a clear blue sky, tall buildings, and a tree-lined sidewalk, suggesting a sunny day in a city environment	upbeat indie folk with gentle rhythms, soft acoustic guitar, and subtle electronic undertones to match the lively yet relaxed atmosphere of a sunny urban	
	high speed and hot outside temperature	a highway with multiple lanes, surrounded by concrete overpasses, under a clear blue sky with some clouds and sunlight shining through	high-energy electronic or synthwave tracks with driving rhythms, vibrant basslines, and uplifting melodies to match the speed and atmosphere	
	medium speed and warm outside temperature	a serene autumn scene with a winding road surrounded by trees shedding their leaves, under a clear blue sky with sunlight filtering through the foliage	soft ambient acoustic melodies with gentle piano and subtle strings, evoking calm and tranquility, complemented by light orchestral tones	
	high speed and warm outside temperature	a two-lane highway with a white truck traveling in the left lane, surrounded by green fields and power lines under a clear blue sky	high-energy electronic rock with driving rhythms, intense bass, and uplifting melodies to match the speed and atmosphere	
	medium speed and cold outside temperature	a snowy mountain road with a single lane and a guardrail, surrounded by snow-covered trees and mountains, under a clear blue sky	soft ambient acoustic melodies with gentle piano and subtle strings, evoking serene winter landscapes, paired with light electronic textures for a calm, reflective mood	

Table 5: Examples of Drive-to-Music outputs showing scene descriptions mapped to musical characteristics and generated cover art images.

B. Model Prompts

Table 6: Prompts used for each model component as part of the Drive-to-Music pipeline.

Model	User Prompt
Image Description VLM	Describe the image by focusing on surroundings, landmarks, vegetation, and weather; ignore people and offensive content in the image; limit the response to one sentence.
Cover Art VLM	Generate a colorful image depicting <image description>; do not generate people or offensive content; return an image with an aspect ratio of one.
Music Description LLM	What kind of music would fit best <image description>? Take into account that the car drives at <speed level> speed and outside is <temperature level>; return only the music description, no other text.
Song Name LLM	Generate a short music title that captures the essence of <music description>; the title should be in Title Case style; return only the music title, no other text.
Song Generation GenAI	Generate <music description>.

Limitations and future work

Participants in our preliminary user study reported recurring melodies, limited variation in tempo and instrumentation, and similarities in genre and style across consecutive tracks, suggesting that future work should address the diversity, structure, and continuity of generated tracks.

The contextual safety framework should also be further validated across a broader range of driving scenarios. Priorities for future work include evaluating the decision thresh-olds against human safety judgments, followed by controlled driving-simulator and real-world in-vehicle studies under varied road, traffic, weather, and vehicle conditions, to measure driver distraction, loudness perception, and the effectiveness of the failure-handling mechanisms under realistic driving conditions.

References

1. Mitra R, Zualkernan IA. Music generation using deep learning and generative AI: a systematic review. IEEE Access. 2025;13:18079-18106. Available from: https://doi.org/10.1109/ACCESS.2025.3531798
2. Yang LC, Lerch A. On the evaluation of generative models in music. Complete journal or conference publication details are required to finalize this reference.
3. Jaini S, Katikireddi PM. Music and art generation using generative AI. Int J Sci Res Comput Sci Eng Inf Technol. 2022;8(6):684-690.
4. Plut C, Pasquier P. Generative music in video games: state of the art, challenges, and prospects. Entertain Comput. 2020;33:100337. Available from: https://doi.org/10.1016/j.entcom.2019.100337
5. Kari M, Grosse-Puppendahl T, Jagaciak A, Bethge D, Schütte R, Holz C. SoundsRide: Affordance-synchronized music mixing for in-car audio augmented reality. In: Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology. New York: Association for Computing Machinery; 2021. Available from: https://doi.org/10.1145/3472749.3474780
6. Çano E, Morisio M. Mood-based on-car music recommendations. In: Proceedings of the ACM conference. Association for Computing Machinery; 2020. Full conference proceedings title, page range, and DOI need verification to complete the NLM reference.
7. Karageorghis CI, Payre W, Howard LW, Kuan G, et al. Influence of music on driver psychology and safety-relevant behaviours: a multi-study inductive content analysis. Theor Issues Ergon Sci. 2022;23(6):643-662. Available from: https://doi.org/10.1080/1463922X.2021.2009933
8. International Telecommunication Union. Algorithms to measure audio programme loudness and true-peak audio level. Recommendation ITU-R BS.1770-5. Geneva: International Telecommunication Union; 2023. Available from: https://www.itu.int/rec/R-REC-BS.1770-5-202311-I
9. European Broadcasting Union. Loudness normalisation and permitted maximum level of audio signals. Technical Report EBU R 128, Version 5.0. Geneva: European Broadcasting Union; 2023.
10. Schmid F, Koutini K, Widmer G. Efficient large-scale audio tagging via transformer-to-CNN knowledge distillation. In: Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. p. 1-5. Available from: https://doi.org/10.17023/xe7z-1q61
11. Amini A, Banaszak A, Benoit H, Böök A, Dakhran T, Duong S, et al. LFM2 technical report. arXiv [Preprint]. 2025. arXiv:2511.23404. Available from: https://arxiv.org/abs/2511.23404
12. Qwen Team. Qwen3 technical report. 2025. Publication details and a verified source URL should be confirmed.
13. Google Gemini Team. Gemini 2.5 technical report. 2025. Publication details and a verified source URL should be confirmed.
14. Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y. CLIPScore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. p. 7514-7528. Available from: https://aclanthology.org/2021.emnlp-main.595/
15. ElevenLabs. ElevenLabs [Internet]. Available from: https://elevenlabs.io/
16. Mubert. Mubert [Internet]. Available from: https://mubert.com/
17. Stable Audio. Stable Audio [Internet]. Available from: https://stableaudio.com/
18. Zhu H, Zhou Y, Chen H, Yu J, Ma Z, Gu R, et al. MuQ: Self-supervised music representation learning with Mel residual vector quantization. arXiv [Preprint]. 2025. Available from: https://arxiv.org/abs/2501.01108
19. Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, et al. Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. Proc Mach Learn Res. 2024;235:12606-12633. Available from: https://proceedings.mlr.press/v235/esser24a.html
20. Wu C, et al. Qwen-Image technical report. 2025. Additional publication details and a verified source URL should be confirmed.
21. Kamath A, Chang KW, Krishna R, Zettlemoyer L, Hu Y, Ghazvininejad M. GenEval 2: Addressing benchmark drift in text-to-image evaluation. arXiv [Preprint]. 2025. Available from: https://arxiv.org/abs/2512.16853