View Reviews

Paper ID
14

Paper Title
Video-Guided Text-to-Music Generation Using Public Domain Movie Collections

Track Name
Papers

Reviewer #1

Questions

2. I am an expert on the topic of the paper.
Disagree

3. The title and abstract reflect the content of the paper.
Strongly agree

4. The paper discusses, cites and compares with all relevant related work
Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")
n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

7. The paper adheres to ISMIR 2025 submission guidelines (uses the ISMIR 2025 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Agree

9. Scholarly/scientific quality: The content is scientifically correct.
Agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")
n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
disagree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Disagree

15. Please explain your assessment of reusable insights in the paper.
Application paper describing dataset creation and model training

16. Write ONE line (in your own words) with the main take-home message from the paper.
There's a new dataset of movie clips, corresponding
soundtracks, and mood annotations called Open Screen Sound Library (OSSL)

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.

Agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines

Weak accept

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

First, my subjective impression of the demo in the supplementary material was that the system doesn't work. Compared to the excellent ground truth trailers and their soundtracking, all systems were awful. Particularly, I have a difficult time perceiving any meaningful differences between the base MusicGen models vs. the text finetune vs. the image adapter.

I thus wonder if it wouldn't be better to rework the paper to exclusively focus on the dataset and its creation, and leave any downstream predictive model for an isolated follow-up publication instead that builds on top of OSSL.

However, the paper is well-written and flows great so I don't think this is a needed change but perhaps a thought for future projects.

Strengths:

- Introduces a valuable new dataset (OSSL) with mood annotations for film music generation.

- Interesting to see concrete attempts of extending MusicGen with a video adapter for multimodal conditioning

- Comprehensive evaluation using both objective and subjective metrics is appreciated

- Ethical and reproducible approach, using public domain data and committing to open release

- Well-considered prompt design using mood and LLM-generated captions

Weaknesses:

- Small human evaluation limits subjective result (fairly acknowledged in conclusion)

- Model innovation is modest (not necessarily bad however), with primarily an integration of existing systems. Again, might be best to split out all model work into a follow-up paper.

- Reliance on public domain data may limit stylistic diversity and realism but it's a fair limitation ofc.!

- Some implementation details (e.g. training times, costs) could be added

- Prompt effects are not ablated or isolated for clearer impact analysis

Reviewer #2

Questions

2. I am an expert on the topic of the paper.
Disagree

3. The title and abstract reflect the content of the paper.
Strongly agree

4. The paper discusses, cites and compares with all relevant related work
Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")
n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Strongly agree

7. The paper adheres to ISMIR 2025 submission guidelines (uses the ISMIR 2025 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.
Strongly agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")
n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Agree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Strongly agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Agree (Novel topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Strongly agree

15. Please explain your assessment of reusable insights in the paper.
The insights into combining video information with text prompts for film music generation are highly reusable across domains like gaming, cinematic production, and interactive media. The dataset construction methodology (especially soundtrack extraction, chroma matching, and mood annotation pipeline) is valuable and can inform similar efforts in MIR and beyond.

16. Write ONE line (in your own words) with the main take-home message from the paper.
The Open Screen Sound Library and video-adapted text-to-music modeling show that adding visual context improves music generation quality for cinematic applications.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Strongly agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Weak reject

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
Strengths:
* The OSSL fills a key gap in MIR by providing ethically sourced, well-annotated film-video-music data at scale, with mood labels.
* The introduction of a lightweight video adapter into an autoregressive text-to-music model (MusicGen) seems efficient.
* The paper is very readable, all major steps are well explained, and there is a strong commitment to open-sourcing the dataset, models, and evaluation metrics.

Suggestions for Improvement:
* The subjective evaluation only involved 15 participants, which is understandable given resource constraints, but future work could aim for larger user studies to strengthen conclusions. In future work, consider expanding subjective evaluations (e.g., 30–50 participants) to strengthen the confidence in user study results, particularly given the variability in human judgments for mood and genre.
* It was good that the authors provided examples along with the paper. However for some of the examples the background sound was higher in volume compared to the dialogues. That might affect the subjective evaluations.
* Additionally, the examples of nervous were more than happy or peaceful. How about sad scenes? How does the model perform for those?
* While MusicGen baselines are appropriate, a comparison to other multimodal generative models (e.g., VidMuse, V2Meow) would further position the contribution. Even a qualitative discussion would help.
* The negative effect of adding genre labels to prompts is an interesting observation. It would be great if future versions explored if certain genre labels are consistently harmful, neutral, or helpful.

* It would be insightful to show a few qualitative examples where the model fails (e.g., a mismatch between video and generated mood), helping readers better understand common failure modes and future improvement directions.

Reviewer #3

Questions

2. I am an expert on the topic of the paper.
Agree

3. The title and abstract reflect the content of the paper.
Strongly agree

4. The paper discusses, cites and compares with all relevant related work
Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")
n/a

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Strongly agree

7. The paper adheres to ISMIR 2025 submission guidelines (uses the ISMIR 2025 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

9. Scholarly/scientific quality: The content is scientifically correct.
Strongly agree

10. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen, otherwise write "n/a")
n/a

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
disagree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Strongly agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Strongly agree

15. Please explain your assessment of reusable insights in the paper.
The paper introduces a we introduce a 36 hour dataset of public domain film clips paired with high-quality soundtracks and mood annotations.

16. Write ONE line (in your own words) with the main take-home message from the paper.
A novel dataset of high-quality video-music pairs is used to fine-tune a SOTA text-to-music generation model and demonstrates that the semantic alignment of the resulting music is improved.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

20. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer

guidelines at https://ismir.net/reviewer-guidelines

Strong accept

21. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.

The primary contribution of the paper is the dataset. The methodology used to construct the dataset is thoughtful and detailed, and I have high hopes for its quality.

In the title, I like the framing of "video-guided text-to-music generation" over "video to music generation" which other papers use to describe similar techniques.

L174: "To achieve the highest music quality by obtaining soundtrack stems without unnecessary noise, we download soundtracks, instead of source-separated music, for each film from YouTube, guided by IMDB metadata." I really like this approach to ensuring high-quality audio rather than settling for youtube-encoded, source-separated audio with audio degradations and other artifacts like distortion from processing like audio ducking for speech and sound effects. I think more researchers should follow suit rather.

L13: "To demonstrate the effectiveness of our dataset in improving the performance of pre-trained models on film music generation tasks..." Are there other tasks you can first use to show the quality of the dataset? To me there is a little bit of incongruity between the primary contribution of the paper - the dataset - and the focus of the paper - the methodology for fine-tuning and evaluating musicgen. A section analyzing the quality of the dataset and providing some statistics/distributions about the dataset compared to other SOTA datasets for the same task would go a long way in balancing things out.

View Meta-Reviews

Paper ID
14

Paper Title
Video-Guided Text-to-Music Generation Using Public Domain Movie Collections

Track Name
Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. I am an expert on the topic of the paper.
Agree

3. The title and abstract reflect the content of the paper.
Disagree

4. The paper discusses, cites and compares with all relevant related work.
Agree

5. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen, otherwise write "n/a")
I didn't carefully check this

6. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

7. The paper adheres to ISMIR 2025 submission guidelines (uses the ISMIR 2025 template, has at most 6 pages of technical content followed by “n” pages of references or ethical considerations, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

8. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.

Agree

9. Scholarly/scientific quality: The content is scientifically correct.
Agree

10. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chose, otherwise write "n/a")
I didn't check the experimental details carefully, but I didn't see anything obviously incorrect.

11. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Disagree

12. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

13. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated “Strongly Agree” and “Agree” can be highlighted, but please do not penalize papers rated “Disagree” or “Strongly Disagree”. Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

14. Reusable insights: The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Disagree

15. Please explain your assessment of reusable insights in the paper.
There are some potential reusable insights about being able to train a video-to-music system on public domain movies and still generalize to commercial movies; or about using video as an additional/alternative conditioning for generation over text. However both of these potential insights are not fully developed/convincing based on the evidence in the paper.

16. Write ONE line (in your own words) with the main take-home message from the paper.

A open source dataset of movies & their aligned soundtracks is released, and a model is trained to generate soundtracks using the video as input.

19. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Disagree

20. Overall evaluation (to be completed before the discussion phase): Please first evaluate before the discussion phase. Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines.
Weak reject

21. Main review and comments for the authors (to be completed before the discussion phase). Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
Strengths:
* The paper addresses an important and challenging problem: generating music for video.
* The creation and release of the OSSL dataset is a significant contribution. The authors have made a commendable effort towards reproducibility by promising to release the dataset, code, and evaluation sets.
* The paper is well-written and easy to follow.

Areas for Improvement:
* Misleading Title: The title suggests that the paper's focus is on a music generation system for films, but the paper focuses mainly on the dataset, plus a proof-of-concept model with limited experiments. Clarifying the emphasis in the title or abstract would improve alignment with the content.
* Lack of Clear Research Questions: The paper would benefit from explicitly stating its research questions for both the dataset and the model. Currently, the narrative reads somewhat like: “we made a dataset, and we trained a model on it,” without clearly articulated motivations or hypotheses. For the dataset: while its open-source nature is valuable, the authors could strengthen its contribution by identifying specific gaps it fills compared to existing datasets. For the model: what specific aspects of video-guided music generation are being explored? One possible direction mentioned—but not emphasized—is how well models trained on public domain data generalize to commercial data. Another could be the comparative benefits of video-guided over text-based soundtrack generation, which would benefit from a deeper literature-based argument and more thorough results analysis.
* Dataset Details:
- A more thorough comparison of OSSL to other existing film music datasets would improve the paper, including a discussion of dataset sizes and type/quality of the content.
- The characteristics of the public domain dataset could be discussed in more detail, explicitly

addressing the release year and quality of the films, and how they compare to the commercial dataset. (I mention this because I assume most public domain movies that you can find on YouTube are quite old? If that's the case, it would be interesting to discuss these particularities.)
- How good is the time-alignment between the soundtrack clips and the movie's audio? Also, are there ever subtle differences between the soundtrack audio and the music in the movie's audio?
* Model Details and Evaluation:
- The related work section on audio-domain music generation contains inaccuracies (e.g. that diffusion-based generation focuses on generating spectral representations - this is not the case for most systems). The section could simply just focus on introducing MusicGen and conditioned music generation, which is probably enough for this paper.
- The authors could discuss the strengths and weaknesses of other multimodal generation approaches in the related work, particularly in comparison to their proposed approach.
- It's unclear if or how script information is used in the model, given the claim in the introduction that video frames alone are insufficient to capture mood, and that scripts often have mood annotations.
- The paper lacks a comparison to existing video-to-music generation systems, making it difficult to assess the effectiveness of the proposed model. The baselines used are kind of apples-and-oranges comparisons, as the baseline does not use video information. A more direct comparison to other video-to-music models would be more compelling.
- The connection between the dataset and the proposed adapter architecture needs better motivation (i.e. why are these two sections part of the same paper?). It is not clear why this architecture is particularly suited for this dataset. Said differently - if the paper was only about the model and was based on a different video + soundtrack dataset, would anything be different about the approach / experiments?
- The human evaluation has a small number of participants & examples, which limits the significance of these results.
- The evaluation results are somewhat mixed in terms of trends, making it difficult to draw clear conclusions about the model or the dataset.
* Minor Points:
- The citation for pyAudioAnalysis is missing.
- There appears to be a typo in Table 2 ("green checkmark on m base") for "OSSL-finetuned" -- should that be a red X instead?
- The method used for mapping soundtrack clips to movie clips could be improved by using more robust fingerprinting techniques besides the one that was tested.
- As a general motivation for video-guided generation, one compelling reason that is not discussed in the paper is that text-based prompts cannot as easily capture important temporal musical elements (e.g. a musical surprise the moment the monster jumps out)

22. Final recommendation (to be completed after the discussion phase) Please give a final recommendation after the discussion phase. In the final recommendation, please do not simply average the scores of the reviewers. Note that the number of recommendation

options for reviewers is different from the number of options here. We encourage you to take a stand, and preferably avoid “weak accepts” or “weak rejects” if possible.
Weak reject

23. Meta-review and final comments for authors (to be completed after the discussion phase)
Paper Summary:
This paper introduces the Open Screen Sound Library (OSSL), a new dataset of movie clips from public domain films paired with soundtracks and mood annotations. The authors also propose a model which extends the MusicGen text-to-music generation model with a video adapter to incorporate video input. The model is evaluated on the evaluation data from OSSL, which includes both public domain and commercial film datasets.

There's a consensus among the reviewers that the OSSL dataset is a valuable contribution, addressing a gap in ethically sourced, well-annotated film-video-music data. All reviewers acknowledge the paper's readability and commitment to open-sourcing the dataset and code. However, I see several significant limitations, particularly concerning the model's evaluation and the paper's overall focus. The paper could be improved by addressing the following concerns:

- Lack of Clear Research Questions and Focus: The paper lacks explicitly stated research questions, particularly for the model. As I noted, the narrative feels like "we made a dataset, and we trained a model on it," without a clear hypothesis or motivation for the model's specific architecture or why it's particularly suited for this dataset. Reviewer 1 also suggests that the model work could be split into a separate follow-up paper, reinforcing the idea that the model's inclusion feels somewhat tacked on rather than central to a well-defined research inquiry.

- Insufficient Model Evaluation and Comparisons: A critical weakness is the absence of a robust comparison to existing video-to-music generation systems. The baselines used are text-to-music models, making it difficult to assess the actual benefit of the video adapter. This concern was raised by myself and Reviewer 2, who explicitly requested comparisons to models like VidMuse or V2Meow. Without such comparisons, the effectiveness of the proposed model, especially its video-guided aspects, remains unconvincing.

- Limited Significance of Model Results: As observed by myself and Reviewer 1, the subjective performance of the model appears to be quite poor. Reviewer 1's statement that "the system doesn't work" and that they had "a difficult time perceiving any meaningful differences between the base MusicGen models vs. the text finetune vs. the image adapter" is a strong indicator of the model's current limitations. The mixed trends in the evaluation results, as I noted, also make it hard to draw clear conclusions. While the human evaluation size is acknowledged as small by the authors, this further compounds the issue of drawing significant conclusions from the model's performance.

- Discrepancy Between Title and Content: The title, "Video-guided Text-to-Music Generation for Films," suggests a primary focus on the generation system. However, as noted in my review, the paper's emphasis leans heavily towards the dataset. This discrepancy, while minor in isolation, contributes to the overall impression that the model component is not as fully developed or justified as it should be for a paper with such a title.

- Unclear Connection Between Dataset and Model: The motivation for connecting the OSSL dataset with this specific adapter architecture is not clearly articulated. It's not evident why this architecture is uniquely suited for the OSSL or what specific insights about video-guided generation are gained by training on this particular dataset. This raises questions about the synergy between the two main contributions.