This study investigated the effects of multimodal (audiovisual) and unimodal (visual-only and audio-only) presented information of film scenes, arousal, and valence on perceived duration. The aim of the study was to gain further insights into these effects. Results suggest that visual-only stimuli led to the longest duration estimations. Higher arousal caused durations as being estimated longer and this effect was modality specific (interaction effect).

The manuscript is well written with a comprehensible method and data processing section. The experiment controlled for the most important aspects. It lists the most important literature regarding time perception in audiovisual contexts. I particularly like the analysis approach of pupil size data. However, the manuscript's comprehensibility would benefit from a few adjustments and by adding some more information. Furthermore, in my opinion reporting on time series analysis of pupil size data should be shortened. Please see below for more detailed commenting.

Introduction:

page 7 line 48: In this section, the authors could mention recent findings on how human movement at different tempos

Participants:
Page 10, line 10 ff: Was the musical and film editing/making experience of the participants assessed? Instrumental practice of participant is mentioned which suggests some kind of measure for music related experiences. If so, please report this and did you find any relation to duration estimation or arousal measures? Did you control for potential familiarity of the music or film scenes? Furthermore, you report on excluding one participant due to large SDs. Was this choice informed some kind of outlier detection or was it just a subjective decision? Please comment.

Stimulus clips:
Page 10, line 50 ff: Having a brief look at the stimulus pool, I noticed that the clips largely differ in terms of aspect ratios. Were these clips presented on full screen and thus, stretched? You mention that audio signals were normalized which is good. Was any comparable procedure applied to the video signals to account for different qualities and resolutions as this might have affect your results, especially arousal measure. Please add further details concerning this.

Procedure:
Page 12, line 3 ff: You mention the monitor size for the presentation of visual signals but information on how audio signals were presented are missing. For example, did you use headphones or speakers and was the volume fixed or adjustable by each participants? Please add this important information.

Page 12, line 35: What informed your choice of using a slider for duration estimations and capping it at 25 seconds?

Eye tracking data acquisition and preprocessing
I really like the approach of using the pupil data from after the stimulus presentation. Nicely done!

Temporal analysis of pupil size
Page, line 48 ff: This is an interesting approach and the reported results are also informative. However, in my opinion the time series analysis of pupil data is prominently positioned throughout the manuscript, yet this is not the main aim of the study. Thus, I suggest reducing the amount of detail here and only report the most important aspects. The rest could be moved to supplementary materials. Just a suggestion to keep the hypothesis testing at the center of the manuscript.

Principal component analysis of arousal measurements
Page 19: This is an interesting approach but I think it would still be interesting to know if there is a difference between perceived and physiological arousal regarding durations estimations. Did you check what arousal measure better predicted your IV?

Page 19, line 27:Furthermore, you mention correlated fluctuations of these measures. This could be checked using repeated-measures correction. Could you add the correction coefficient? Please see:

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in psychology*, *8*, 456.

Hypotheses testing:

Page 20, line 39 f: It is a bit tricky understanding the equations (some goes for the equation for the time series analysis) as you include the terms "AV:AO" and „AV:VO“ but don't explain what the mean. This is explained later on page 21 starting from line 15. I suggest moving parts of this paragraph up so it is easier to follow your approach.

Page 21, line 15: Please mention what kind of centering was used, e.g. grand-mean or cluster-wise centering.

Page 21, 34 ff: "We chose the maximal number of random-effects parameters that (1) controlled for between-participant differences in average performance and (2) afforded converging parameter estimates across multiple iterations of model-fitting.“ I am not sure what is meant by maximal number of random-effects parameters. Reading this I thought it meant including all fixed effects as random effects as well.

Page 21: Why did you not included the groups as a random effect as participants are nested within these three groups. I think this would more accurately represent you data structure. Please provided reasoning for not using a participants nested within groups random effect structure.

Results:

Page 22, line 8 ff: I understand that explaining and reporting the approach and results of the pupil size data is important. Yet as mentioned above, I think you should report results regarding your hypotheses first. Although very interesting, the time analysis does not add much to the main aim of the study but rather concerns methodological questions. In my opinion this could go in the supps or reported much more briefly.

Additional findings:

Page 27, line 18: Did you also have a look at the audio signals? For example, the tempo of music has been shown to be the driving factor regarding effects on duration estimation and that tempo is closely related to induced arousal as well (Hammerschmidt & Wöllner, 2020; Wöllner & Hammerschmidt, 2021). It would be interesting if you could compare the tempo and intensity (e.g. perceived loudness) of the audio signals as you did with the number of cuts for the video signals.

Discussion:

Page 29, line 7: Please start the paragraph by repeating the main aim of the study. Makes it easier to follow the discussion when this is recapped here.

Page 34, line 11 ff: Regarding the dominance of audition in temporal tasks results of these two recent studies on tempo and duration perception in audiovisual contexts could be discussed as well:

Allingham, E., Hammerschmidt, D., & Wöllner, C. (2021). Time perception in human movement: Effects of speed and agency on duration estimation. *Quarterly Journal of Experimental Psychology*, *74*(3), 559-572.

Wang, X., Wöllner, C., & Shi, Z. (2021). Perceiving Tempo in Incongruent Audiovisual Presentations of Human Motion: Evidence for a Visual Driving Effect. *Timing & Time Perception*, *1*, 1-21.

Page 29 ff: The Discussion would benefit if for each section the corresponding hypothesis would be mentioned.