REPORT: LLAVA-BASED IMAGE RATING TASK AND ANALYSIS

Yanheng Li
Renmin University of China
chrislee1067@outlook.com

1 METHODOLOGY

I select llava-v1.6-vicuna-13b (Liu et al. (2024)) as the model to execute the image rating task because of its availability and capability to process multi-modal tasks. Below I will introduce the methods used in the task, including the data, prompts design, implementation and analysis.

Dataset and Benchmark As provided in the given paper (Lehman et al. (2019)), the dataset is available online[1]. The dataset including 320 images to be processed. Besides, I used their uploaded sheet file containing average human ratings of relevance, arousal, and valence for each image as the benchmark to assess the performance of LLaVa rating.

Prompts Design Combining the previous prompts of the task and some other more popular templates used in many ACL conference works (e.g., Zhou et al. (2024)), I design three structured prompts for the LLaVa to output the rating as well as the model CoT (Chain of Thoughts) of "Relevance", "Arousal" and "Valence" in a relatively rigid template for easier parsing process. See prompts in Appendix A.

Implementation Details and Pseudocode I first deploy and test the LLaVA model locally **with setting `Temperature==0.3` to control the randomness**. Then, I integrate the image dataset with designated prompts for each rating aspect, formatting the prompts into a "jsonl" structure as model input. Each image input consists of the image itself and the corresponding prompt. Next, I run the model to perform the rating task, generating ten outputs per image to ensure more effective results and provide additional analysis options (e.g., evaluating model robustness). To analyze the results, I extract all ratings from the output "jsonl" files and compile them into a new ".xlsx" file. I then conduct statistical and empirical analyses, as detailed in the accompanying ".ipynb" files. For further reference, see the pseudocode in Appendix C.

Analysis To evaluate the performance of LLaVa image encoder, I set two parts for the analysis. Firstly, **the standard deviation, the Pearson correlation and the Intra-Class Correlation (ICC)** of all LLaVa ratings (10 ratings for each aspect per image) are computed to evaluate the robustness of the model itself. Secondly, to compare the average ratings between "LLaVa" and "Human", I design a three-phases evaluation. See the statistical details in Appendix C.

- Test whether there is significant difference as well as the direction of the difference between Human ratings and LLaVa ratings, using **Wilcoxon Signed-Rank Directional Test**.
- If there is no significant difference between some rating results, test whether there is statistical equivalence by **TOST**.
- Calculate **the errors (i.e., the MAEs, the MSEs), the correlations and the variances** between LLaVa and Human ratings across different rating aspects (i.e., relevance, arousal, valence) to quantitatively further analyze the difference between LLaVa and Human encoder.

[1] https://affectiveclimateimages.weebly.com/

2 Analysis Results

LLaVa Collectively, the relatively high standard deviation (Figure 1) and the low ICC (Figure 2) in LLaVA's ratings for each aspect per image indicate suboptimal model robustness. This suggests significant variability in LLaVA's ratings, which may necessitate either a more capable model or a fine-tuning process before encoding. Additionally, the correlation matrix (Figure 3) demonstrates the independence of LLaVA's ratings across the three aspects, potentially reflecting the model's ability to distinguish between different rating criteria.

LLaVa Versus Human Benchmark I set the average human rating as the benchmark for comparison against LLaVA's average ratings across the three aspects per image. As shown in Table 1, LLaVA's *Relevance* ratings are significantly lower than human ratings, whereas its *Valence* ratings are significantly higher, both at the 0.05 significance level. In contrast, the *Arousal* rating exhibits no significant difference between LLaVA and human ratings, as confirmed by the TOST analysis in Table 2, suggesting statistical equivalence. Consistently in Figure 4, the variance of *Arousal* among human and LLaVA ratings is the smallest, approaching $zero$ despite the presence of some outliers. In comparison, the variances of *Relevance* and *Valence* deviate further from $zero$, indicating greater discrepancies between LLaVA and human ratings in the two aspects.

Table 3 presents the Mean Absolute Error (MAE) and Mean Squared Error (MSE) values between ratings from LLaVa and human raters across multiple scenarios, ranging from using a single LLaVa rating to the average of up to ten LLaVa ratings. For all dimensions (Relevance, Arousal, and Valence), averaging more LLaVa ratings (moving from Group 1 to Group 10) results in a clear improvement in both MAE and MSE. This supports that aggregating multiple ratings can mitigate individual outliers or inconsistencies in predictions, leading to more stable and accurate results. It is worth noting that the *Arousal* dimension consistently shows the smallest MAE and MSE, suggesting that LLaVa may have a better ability to predict arousal ratings, aligned with previous results. While *Relevance* still exhibits relatively higher MAE and MSE compared to Valence and Arousal, which could indicate that relevance ratings are more variable or context-dependent, posing a greater challenge for LLaVa to predict accurately.

Figures 5 – 14 consistently show relatively high correlations between *Relevance* and *Valence* ratings from LLaVa and human raters across various scenarios (i.e., using one LLaVa rating, the average of two LLaVa ratings, and so on, up to ten LLaVa ratings). These results align with the earlier test findings, as correlation alone does not fully capture model performance. However, the observed negative correlation between *Arousal* ratings from human raters and LLaVa raises caution and suggests a more conservative assessment of the model's performance of this aspect.

Comparison	Wilcoxon Stat.	p-value (Higher)	p-value (Lower)
Relevance	8,114.00	1.00	≈ 0.00
Arousal	23,086.50	0.90	0.10
Valence	34,042.00	≈ 0.00	1.00

Table 1: Wilcoxon Directional Test Results (Human ratings as comparison benchmark)

Comparison	Mean Diff.	p-value(Left)	p-value(Right)	Equivalent
Relevance	-0.92	0.99	≈ 0.00	No
Arousal	-0.07	0.88	0.12	Yes
Valence	0.28	0.99	≈ 0.00	No

Table 2: TOST Equivalence Test Results

3 Case Study

3.1 Instability

To better understand the performance of LLaVa, a case could be illustrated by two ratings of *Relevance* both from "pic001" with the same prompts, see Appendix C for the pic001 (Figure 15)

Grp.	Avg_Relevance	Avg_Arousal	Avg_Valence
01	1.61 / 3.75	1.06 / 1.80	1.10 / 1.90
02	1.44 / 3.10	0.98 / 1.53	0.98 / 1.59
03	1.38 / 2.89	0.94 / 1.43	0.90 / 1.31
04	1.32 / 2.70	0.94 / 1.40	0.86 / 1.19
05	1.30 / 2.65	0.93 / 1.36	0.84 / 1.10
06	1.30 / 2.60	0.92 / 1.34	0.82 / 1.05
07	1.28 / 2.54	0.89 / 1.29	0.81 / 1.03
08	1.29 / 2.51	0.89 / 1.29	0.80 / 1.00
09	1.28 / 2.51	0.89 / 1.27	0.79 / 0.98
10	1.28 / 2.50	0.89 / 1.27	0.79 / 0.99

Table 3: MAE and MSE Results for Different Groups (Rounded to Two Decimal Places). Numbers ahead of the slash represent the MAE, while after the slash represent the MSE.

and the prompts. Two answers of LLaVa are as below:

> $AnswerNo.1$ {Rating: 9. Reason: This image depicts two polar bears swimming across an area of water. Climate change is a major factor affecting the Arctic ecosystem, and as sea ice melts due to climate change, polar bears face increasing challenges in finding food and breeding. The image serves as a powerful visual reminder of the impacts of climate change on the environment and wildlife.}

> $AnswerNo.4$ {Rating: 5. Reason: The image shows two polar bears, which are animals directly affected by climate change. Their habitat is changing due to rising temperatures and melting ice, leading to challenges in their survival. However, the image does not explicitly depict any direct effects of climate change on the environment. It also doesn't show clear evidence of human activity or infrastructure that contributes significantly to climate change. Therefore, while it is relevant to climate change as an indicator species, its direct impact on human-induced climate change is less visible in this image alone.}

The above responses demonstrate significant inconsistencies in LLaVA's rating mechanism and reasoning. For example, $AnswerNo.1$ describes the image as a "powerful visual reminder" of climate change, emphasizing **the symbolic significance of polar bears** and justifying a high relevance score, whereas $AnswerNo.4$ downplays this symbolism and argues that the image does not provide direct **evidence of human activity**, leading to a much lower rating. This suggests that the model applies stochastic evaluation criteria across different runs.

3.2 Discrepancies against Human Rating

Since LLaVa presents its weakness when it comes to *Relevance*(LLaVa<Human) and *Valence*(LLaVa>Human). For *Arousal*, though LLaVa shows best performance on it across three aspects, there still remains uncertainty as for its low correlation. Here, three typical cases are selected for understanding these discrepancies.

3.2.1 Relevance: LLaVa<Human

Pic212 (see Figure 16) has the largest negative difference between *Relevance* ratings from LLaVa (averaging from 10 ratings) and Human ($Avg.LLaVa(1.10) - -Avg.Human(5.28) = -4.18$). Most answers from LLaVa output the rating 1 or 2. One LLaVa answer is as follows.

> *Rating: 1. Reason: The image depicts a group of soldiers in trenches, which is not directly related to climate change. Climate change typically involves environmental science, policy, and the impact on ecosystems and human societies, none of which are represented in this historical military scene.*

3.2.2 VALENCE: LLAVA>HUMAN

Pic026 (see Figure 17) has the largest positive difference between *Valence* ratings from LLaVa (averaging from 10 ratings) and Human ($Avg.LLaVa(5.20) - -Avg.Human(2.12) = 3.08$). Most answers provide ratings ranging from 4 to 7. One LLaVa answer is as follows.

> *Rating: 7. Reason: The image depicts a beach scene with boats, which is generally considered positive. However, the presence of algae or seaweed on the shore could be seen as negative due to its potential impact on water quality and the environment.*

3.2.3 AROUSAL: AN OUTLIER

Pic100 (see Figure 18) has the largest negative difference between *Arousal* ratings from LLaVa (averaging from 10 ratings) and Human ($Avg.LLaVa(1.30) - -Avg.Human(5.40) = -3.90$). Most answers provide a rating 1. One LLaVa answer is as follows.

> *Rating: 1. Reason: The image depicts a scene of destruction and disarray, with piles of rubble, debris, and trash. This type of imagery is typically associated with chaos or distress, which would likely evoke feelings of unease or concern in participants rather than calmness or excitement.*

In all the cases above, it is consistent that LLaVa did not connect the symbolic elements in the images to climate change, likely due to a mismatch between the content of the image and the typical associations the model has with climate change. Additionally, LLaVa seems to exhibit less imaginative reasoning than humans in certain contexts. While some elements in images may be intuitive and easily understood by humans, they can be more abstract for LLaVa. This suggests that certain iconic symbols, whose meanings are acquired through intuition or cultural context and are obvious to humans, pose greater challenges for LLaVa to reason about.

4 DISCUSSION

The results suggest that while LLaVA captures *Arousal* in a manner closely aligned with human perception, its assessments of *Relevance* and *Valence* exhibit systematic differences in its understanding of visual relevance and emotional intensity, as well as a lack of rating robustness. To be considered, fine-tuning process or more reliable prompts may be beneficial to the reliability of LLaVA's image encoder.

REFERENCES

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8:194855061769717, 05 2017. doi: 10.1177/1948550617697177.

Betsy Lehman, Jessica Thompson, Shawn Davis, and Joshua M. Carlson. Affective images of climate change. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00960. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00960.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351. ELRA and ICCL, May 2024. URL https://aclanthology.org/2024.lrec-main.816/.

A PROMPTS

1. **Relevance Rating Prompt:**
   Now you are an image encoder for some images. You will be given an image. You must observe and analyze the image carefully, then produce an answer considering:

   Rating:
   Please provide a rating of this image to describe ̈how relevant or irrelevant it was to climate change~~,~~ on a scale of 1 to 9 (1 means the most irrelevant, 9 means the most relevant.)

   Your answer must rigidly follow this template:
   Rating: an integer on a scale to 1 to 9.
   Reason: a short reason of the given rating.

2. **Arousal Rating Prompt:**
   Now you are an image encoder for some images. You will be given an image. You must observe and analyze the image carefully, then produce an answer considering:

   Rating:
   Please provide a rating of this image to describe ̈how calming or exciting it made participants feel" on a scale of 1 to 9 (1 means the most calm, 9 means the most excited.)

   Your answer must rigidly follow this template:
   Rating: an integer on a scale to 1 to 9.
   Reason: a short reason of the given rating.

3. **Valence Rating Prompt:**
   Now you are an image encoder for some images. You will be given an image. You must observe and analyze the image carefully, then produce an answer considering:

   Rating:
   Please provide a rating of this image to describe ̈how negative or positive the images appeared~~,~~ on a scale of 1 to 9 (1 means the most negative, 9 means the most positive.)

   Your answer must rigidly follow this template:
   Rating: an integer on a scale to 1 to 9.
   Reason: a short reason of the given rating.

B FIGURES FOR ANALYSIS

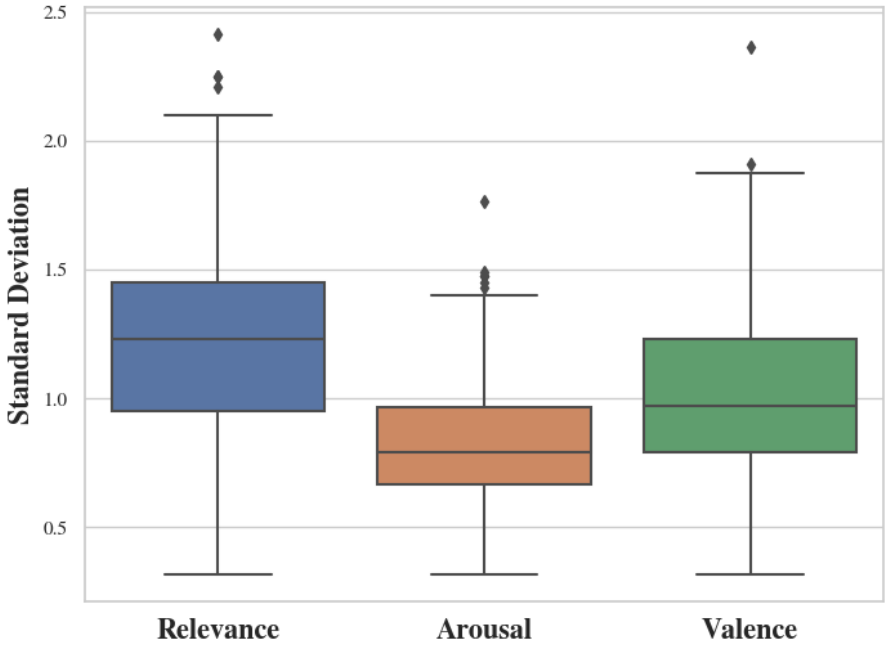

Figure 1: Standard deviation of ten LLaVA ratings per image across three aspects.

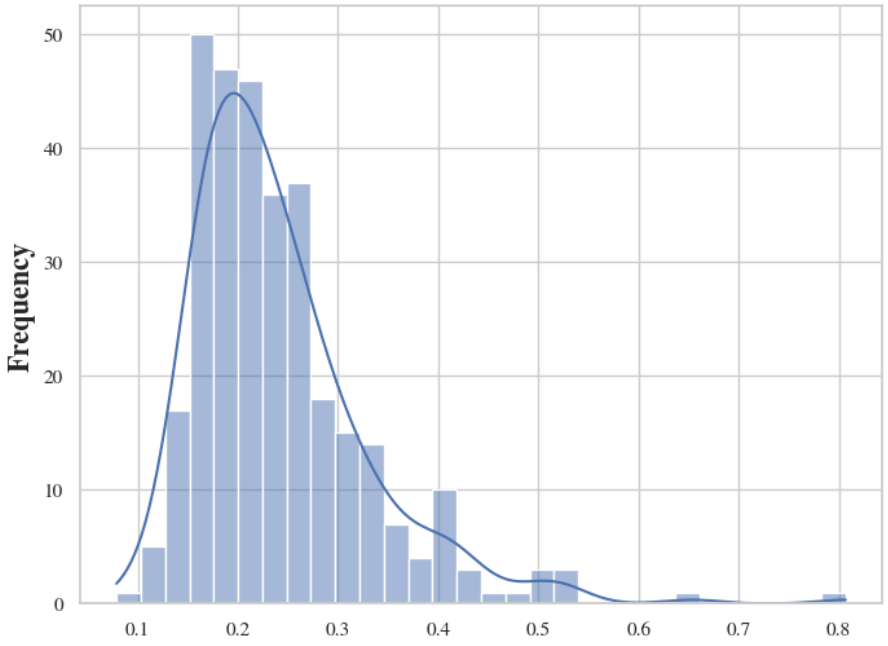

Figure 2: Consistency analysis (ICC) of ten LLaVA ratings per image across three aspects.

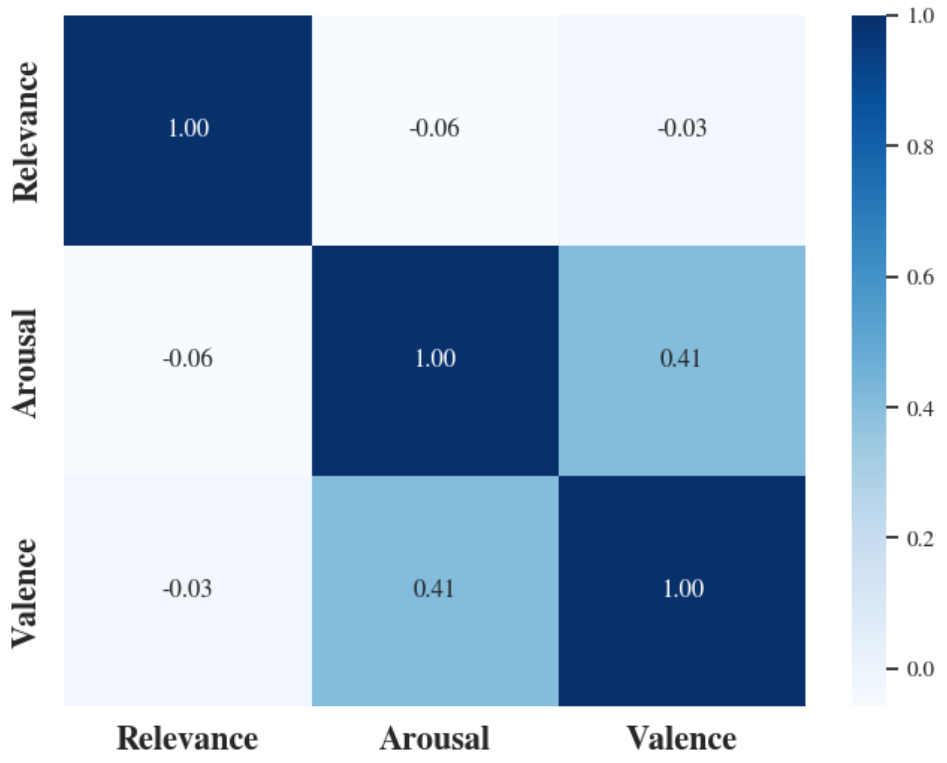

Figure 3: Pearson correlation coefficients of all LLaVA ratings per image among three aspects.

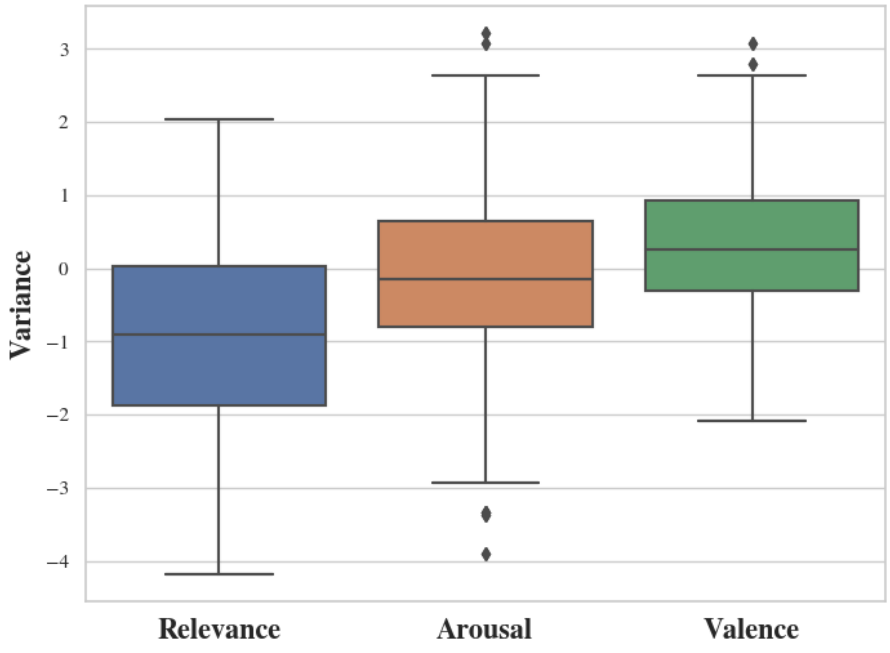

Figure 4: Variance among Human and LLaVA ratings per image across three aspects (Variance = LLaVA Rating − Human Rating).

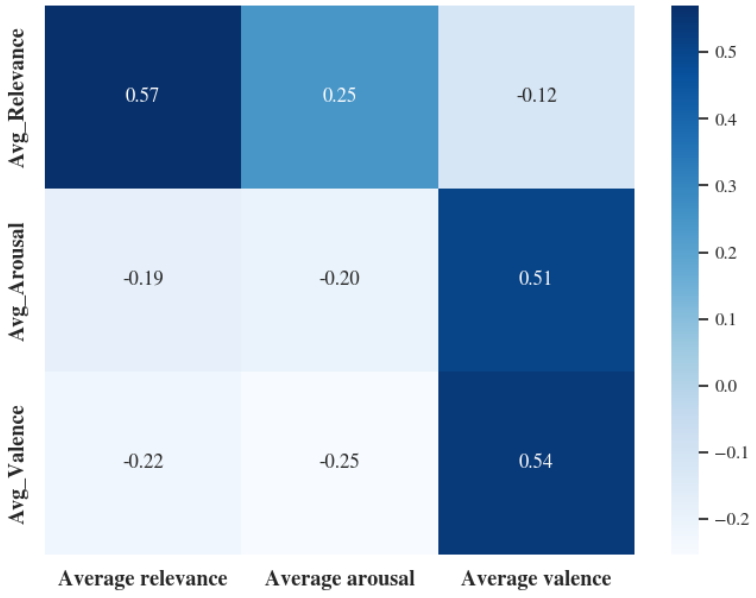

Figure 5: Pearson correlation matrix of Only One LLaVa Rating Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

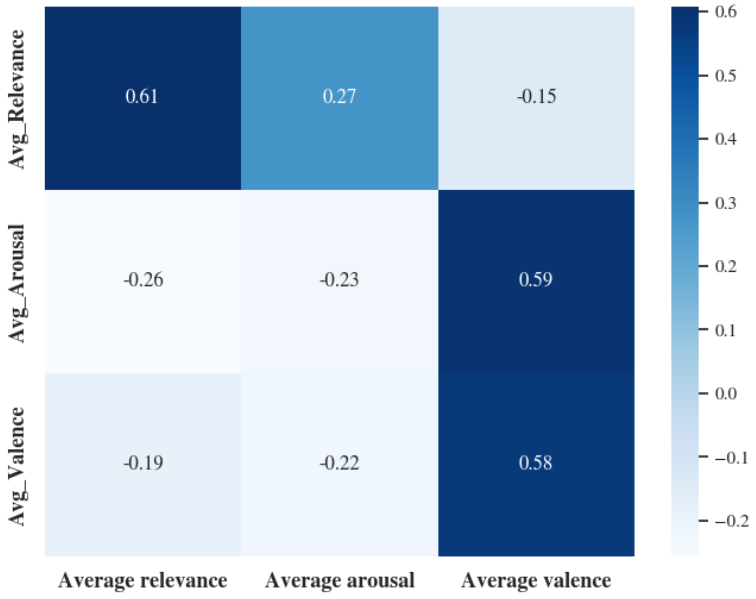

Figure 6: Pearson correlation matrix of The Average of Two LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

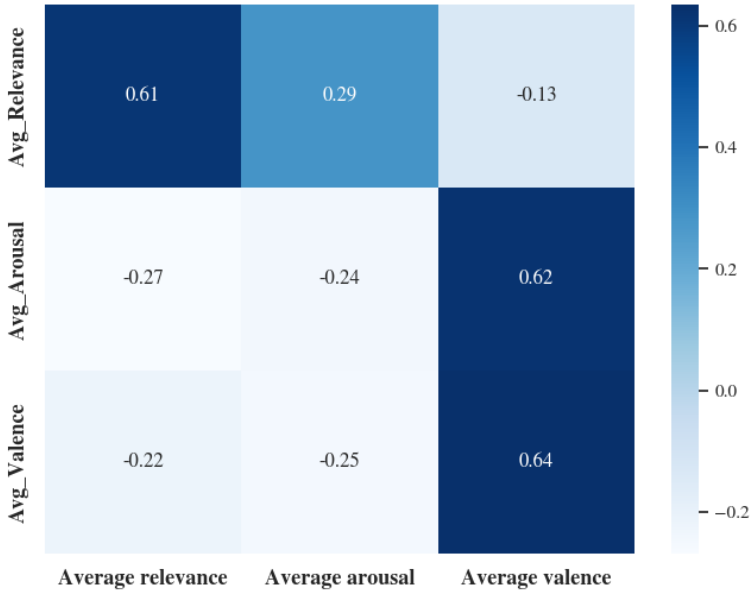

Figure 7: Pearson correlation matrix of The Average of Three LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

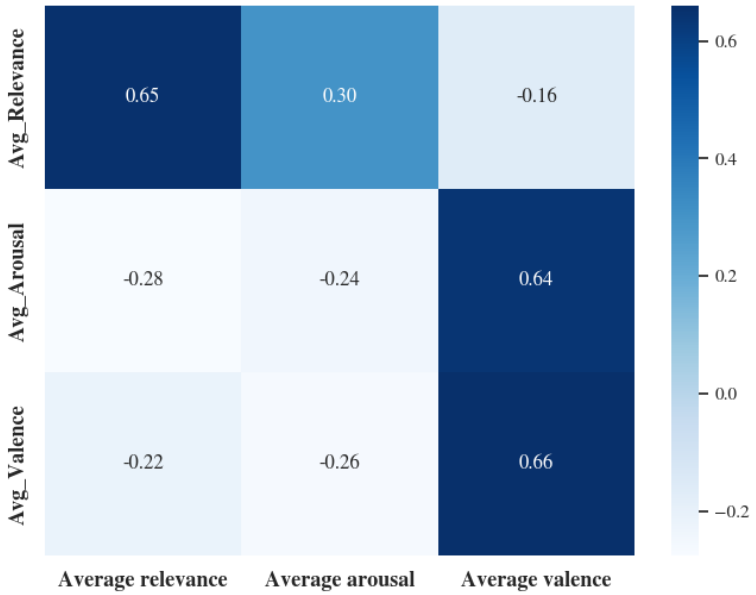

Figure 8: Pearson correlation matrix of The Average of Four LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

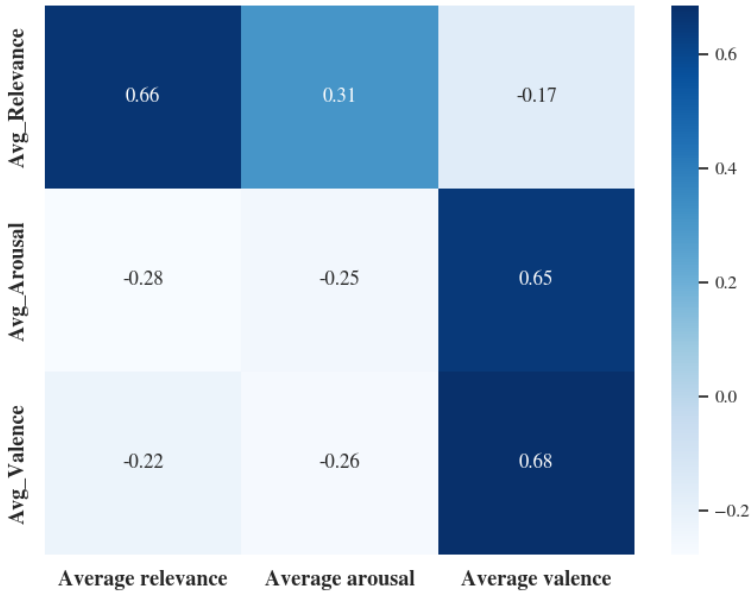

Figure 9: Pearson correlation matrix of The Average of Five LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

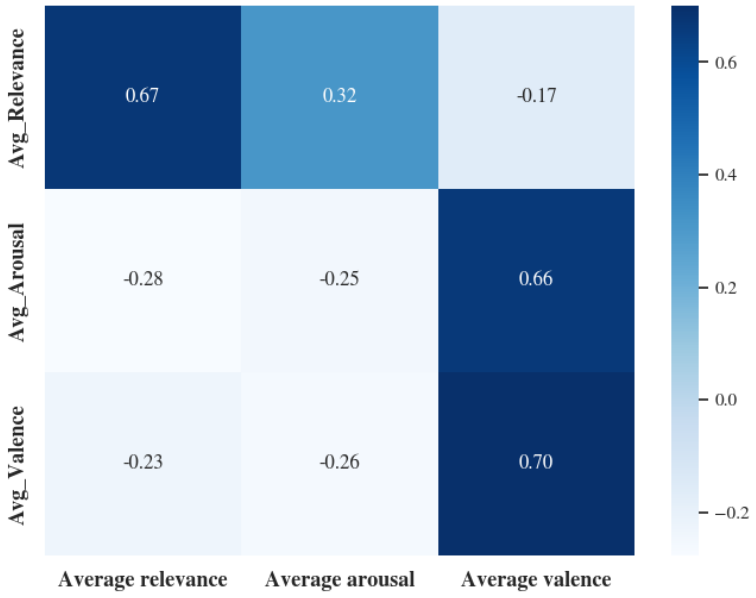

Figure 10: Pearson correlation matrix of The Average of Six LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

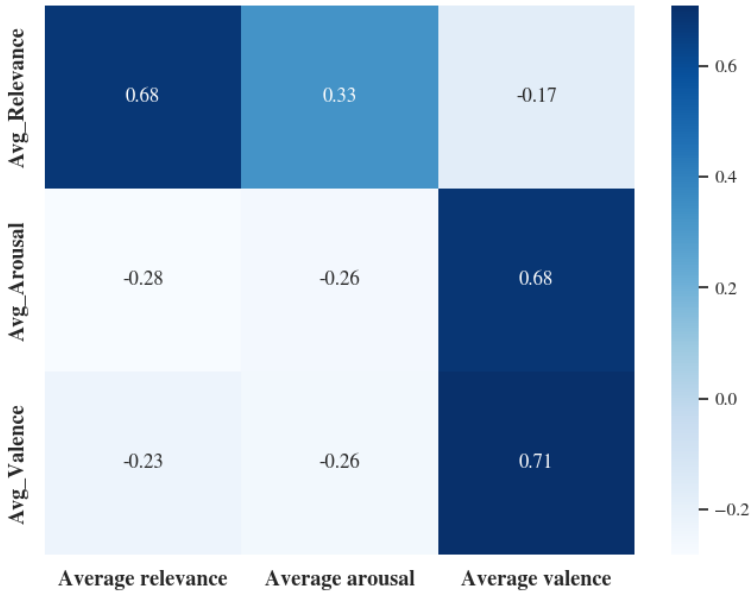

Figure 11: Pearson correlation matrix of The Average of Seven LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

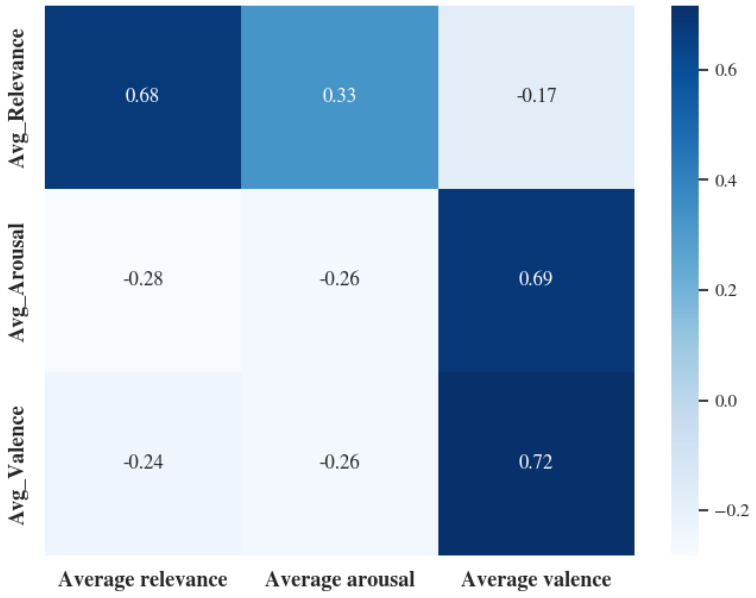

Figure 12: Pearson correlation matrix of The Average of Eight LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

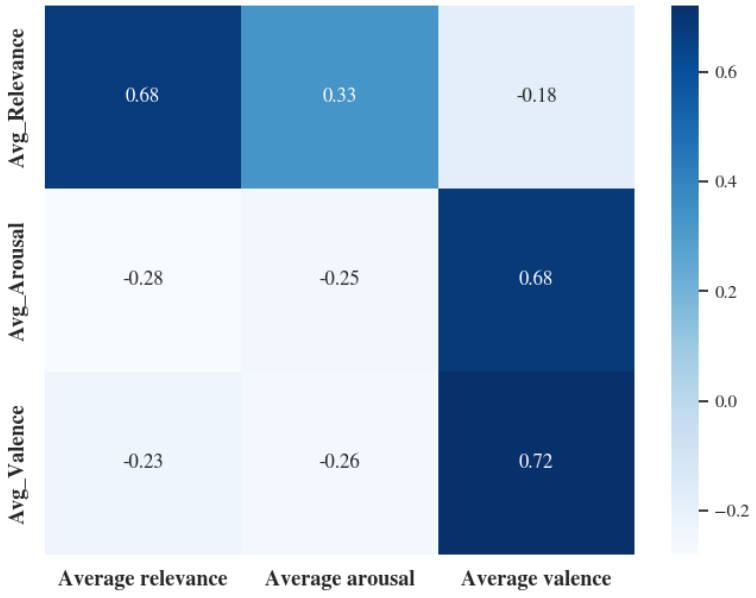

Figure 13: Pearson correlation matrix of The Average of Nine LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

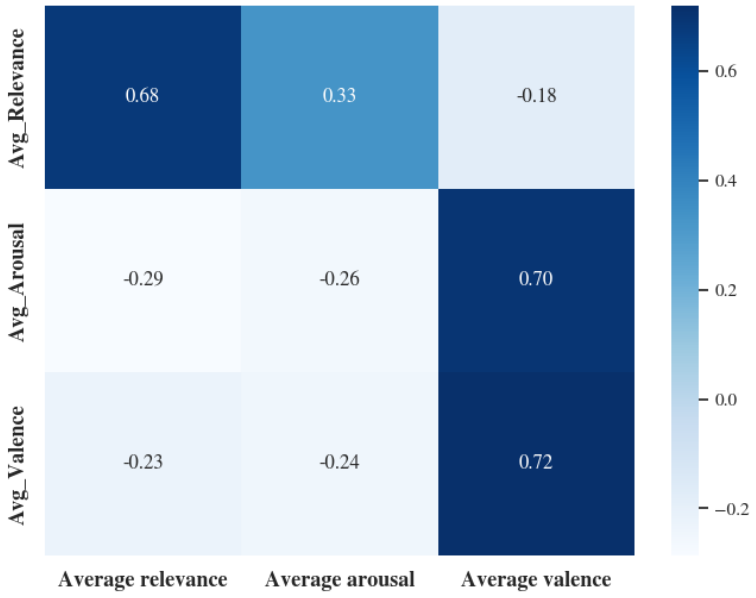

Figure 14: Pearson correlation matrix of The Average of Ten LLaVa Ratings Against Human Rating. The x-axis represents ratings from human, while the y-axis represents ratings from LLaVa.

C SUPPLEMENTS

Intra-Class Correlation The ICC is computed as:

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

where: $\sigma_B^2$is the **between-group variance**, representing variability among different groups (e.g., different images). σ_W^2 is the **within-group variance**, representing variability within each group (e.g., different model ratings for the same image).

The between-group variance σ_B^2 is given by:

$$\sigma_B^2 = \frac{1}{k}\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

where: k is the number of groups (images), n_i is the number of ratings in group i, $\bar{X}_i$ is the mean of ratings in group i, $\bar{X}$ is the grand mean (overall mean across all groups).

The within-group variance σ_W^2 is given by:

$$\sigma_W^2 = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

where: N is the total number of ratings, X_{ij} is the rating of the jth observation in the ith group.

$ICC \approx 0$ indicates low agreement (random variation dominates).
$ICC \approx 0$ indicates high agreement (most variation is between groups).

Wilcoxon Signed-Rank Test Let (b_i, p_i) be the baseline and LLaVa ratings for instance i, and define $d_i = b_i - p_i$. We discard any i where $d_i = 0$. For $d_i \neq 0$, let r_i be the rank of $|d_i|$ among these nonzero differences, with $r_i = 1$ for the smallest and $r_i = N$ for the largest (where N is the number of nonzero differences). We then define:

$$W^+ = \sum_{i:d_i>0} r_i, \quad W^- = \sum_{i:d_i<0} r_i.$$

We use W^+ to test $H_0 : \text{median}(d_i) \leq 0$ against $H_A : \text{median}(d_i) > 0$, and an analogous procedure tests $H_0 : \text{median}(d_i) \geq 0$. Under the null, W^+ follows a known distribution for small N, and an asymptotic normal approximation for large N:

$$z = \frac{W^+ - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}.$$

If z exceeds a critical value z_α, we reject H_0, indicating a significant directional shift.

Equivalence (TOST) If the Wilcoxon Signed-Rank Test indicates no significant difference, we turn to determine whether $|\bar{d}|$ is sufficiently small for practical equivalence. Specifically,

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i,$$

$$\sigma_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2},$$

$$\text{SE} = \frac{\sigma_d}{\sqrt{n}}.$$

and an equivalence margin δ is defined. We compute

$$t_\ell = \frac{\bar{d} + \delta}{\text{SE}}, \quad t_u = \frac{\bar{d} - \delta}{\text{SE}}.$$

We reject $H_0^-(\delta) : \bar{d} \leq -\delta$ if $t_\ell > t_{\alpha,n-1}$, and $H_0^+(\delta) : \bar{d} \geq +\delta$ if $t_u < -t_{\alpha,n-1}$. Rejecting both implies $\left|\bar{d}\right| < \delta$—no meaningful difference (Lakens, 2017).

C.1 PSEUDOCODE

Algorithm: Image Rating on LLaVa

1. **Input**:
   - Image data $\mathcal{P}$;
   - LLaVa Rating function $M(\cdot, \cdot)$;
   - Prompts settings $\mathcal{R}$.
2. **Implement**:
   - For each pair (p_i, r_i), let the LLaVa model output be $o_{\text{output}} = M(p_i, r_i)$;
   - Repeat 10 times, the model outputs $o_{\text{output}} = \{M_t(p_i, r_i)\}_{t=1}^{10}\}$.
3. **Benchmark Analysis** (Δo):
   - Use average human rating results as benchmark o_{base};
   - Use the average LLaVa rating Avg $o_{\text{output}} = \frac{1}{10}\sum_{i=1}^{10} o_{\text{output}_t}$ to be compared with benchmark o_{base};
   - Compute the variance $\Delta o =$ Avg $o_{\text{output}} - o_{\text{base}}$.
4. **Analyze Results**:
   - Compare Δo with 0;
   - Evaluate the performance of LLaVa by some test methods, e.g., directional test.

D CASE STUDY

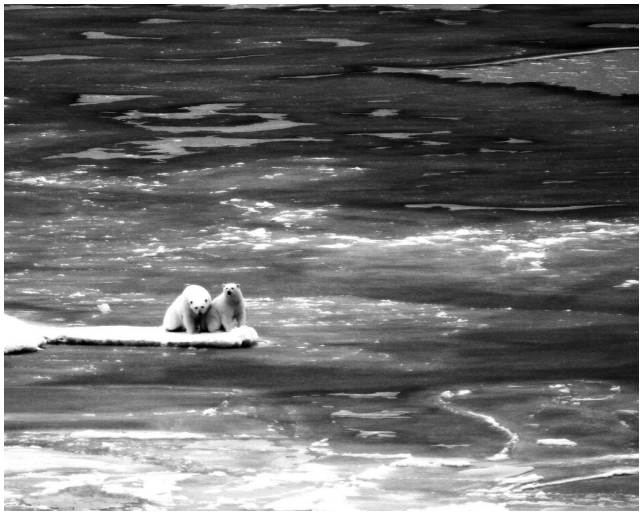

Figure 15: pic001

Figure 16: pic212

Figure 17: pic026

Figure 18: pic100