Comparative Study of Window Views' Distinctive Impact on Human and VLM Impressions

Anonymous ACL submission

Abstract

This paper investigates the subjective dimensions of window view impressions by comparing human participants' verbal responses with image descriptions generated by seven state-of-the-art vision-language models (VLMs). We analyze a dataset (Cho et al., 2023b, 2025a,b) of transcribed impressions—2100 utterances collected in two separate virtual reality (VR) experiments—and compare it against synthetic texts from several high-performing VLMs. Using the combined dataset, we compare human and machine responses based on three key criteria: (1) most frequent N-grams, (2) clustering structure, and (3) sentiment. Our findings reveal significant differences across all three dimensions and highlight distinctive patterns in human perceptions of window views.

1 Introduction

Access to a window view strongly shapes occupant comfort, satisfaction, well-being, and spatial perception (Markus, 1967; van Esch et al., 2019; Gerhardsson and Laike, 2021). Consequently, assessing perceived view quality has become an important goal in applied architectural research. Most existing studies pursue this goal by manually recording participants' subjective impressions, typically through questionnaires, image and VR-based rating scales, interviews, or physiological assessments (Abd-Alhamid et al., 2023; Cho et al., 2023a; Matusiak and Klöckner, 2016; Aries et al., 2010). These protocols are time consuming and susceptible to inconsistency and human error, since each response requires manual annotation. As a result, there is growing interest in automating the estimation of perceived window view quality. Recent computer vision studies already extract key view metrics: (Xia et al., 2021) predict sky view factor as a proxy for openness; (Ranftl et al., 2021) estimate monocular depth to recover viewing distance; and (Gong et al., 2018) use attention-based segmentation to map a scene's semantic composition. While these pipelines quantify visual features, they still do not generate a direct textual appraisal of perceived view quality.

Vision-language models, on the other hand, can process an image and generate concise, factually accurate textual description (Cheng et al., 2025). Thus, VLMs offer a promising route for predicting the textual impression of the window view. Yet it remains unclear how closely their outputs capture the many facets of human window view perception. This gap motivates our guiding question:

> **Q.** How do human participants' impressions of window views compare with the descriptions produced by state-of-the-art vision-language models?

In this study, we focus on office-window views evaluated by university students and staff (Cho et al., 2023b, 2025a,b). Cho et al. collected 2100 transcribed descriptions covering 50 scene-condition combinations. These scenes were captured on a university campus and were presented to participants in VR in either an image or a video format. Building on this dataset, we conduct an in-depth exploratory comparison between human descriptions and captions generated by seven state-of-the-art vision-language models.

Contributions. Our study makes two key contributions:

- Dataset extension: For each of the 35 scene-condition images, we added captions from seven state-of-the-art VLMs: 6 captions per baseline model and 20 from the best-performing model, yielding 910 machine-generated descriptions that sit alongside the 2100 human descriptions and can be queried by scene, condition, or model.

- Comparative analysis: To our knowledge, this is the first systematic comparison of human- and machine-generated descriptions of window views that jointly evaluates sentiment, lexical choice, and content saliency.

2 Related Work

In recent years, Large Language Models (LLMs) have gained significant popularity, with several studies validating machine-generated responses against human texts (Guo et al., 2023; Herbold et al., 2023; Ha and O'Donoghue, 2024). Notably, (Guo et al., 2023) proposes a RoBERTa-based ChatGPT detector that distinguishes between human answers and responses from GPT-3.5 with an F1 score of 98.78 across a variety of knowledge domains. When analyzing the linguistic differences between humans and GPT-3.5, (Guo et al., 2023) reports that GPT-3.5 tends to produce longer answers but with a smaller vocabulary. Further, they note that GPT-3.5 generations exhibit a more formal style, greater objectivity, and less emotion. (Ha and O'Donoghue, 2024) notes a similar trend in Llama-2 generations; the authors report that machine-generated text tends to have a more positive sentiment than the human-authored equivalent. Meanwhile, (Herbold et al., 2023) investigated the output of a more modern GPT-4 and reported greater lexical diversity in the model's essays when compared to human texts. However, several linguistic characteristics still distinguish GPT-4, including fewer discourse markers, more nominalizations, and higher syntactic complexity.

A parallel line of research compares the output of vision-enabled LLMs with human image descriptions. (Cheng et al., 2025) reports that OpenAI's GPT-4o reaches or even surpasses human performance in terms of precision and level of detail. However, the authors explicitly consider only the factual correctness of the captions, not their style or linguistic characteristics.

In this study, we compare the responses of vision-enabled LLMs with human impressions of window views collected in (Cho et al., 2023b, 2025a,b). In these studies, the authors conducted two independent VR experiments, each with 42 participants and identical hardware and protocol. The first experiment presented 15 campus views twice (once as a static image and once as a matched video), yielding 30 scene–format combinations and 1260 verbal impressions. The second experiment revisited 10 of those locations under clear and overcast skies, producing 20 scene-sky combinations and a further 840 impressions from a new cohort. Each of the aforementioned campus views is shown in Tables 2, 3, and 4. The present paper pools all 2100 transcribed utterances from these studies; analysis of machine-generated impressions for on-campus videos is deferred to future work.

3 Dataset Construction

To obtain synthetic window view impressions in the form of textual descriptors for each image, we used commercially available vision-language models through their web APIs. The model settings and generation hyperparameters are documented in Appendix A. We used the same prompt shown to human participants, enabling a direct, side-by-side comparison between machine- and human-generated responses.

4 Model Selection

We used the CapArena (Cheng et al., 2025) benchmark to select top-performing vision-enabled LLMs accessible via web APIs. Our selection includes both reasoning models (Gemini 2.5 Pro, Claude Sonnet 4, o4-mini) and non-reasoning models (Gemini 2.5 Flash, Claude 3.5 Haiku, Qwen 72B VL, and GPT-4.1). Exact model IDs can be found in Appendix B. We then used BERTScore (F1) (Zhang et al., 2019) to compute the semantic similarity between the model-generated texts and the transcribed human impressions. Furthermore, we computed intragroup BERTScore (F1) to assess internal consistency among human and VLM responses. A detailed description of the procedure for computing BERT scores is outlined in Appendix C.

Figure 1 shows that, for all evaluated VLMs, model/human similarity is lower than human/human similarity. The strongest alignment with the human texts is achieved by GPT-4.1. This finding motivated us to further investigate its image descriptions. To this end, we sampled 20 generations for each scene-condition pair, as a compromise between output diversity and computational cost (Theodoropoulos et al., 2025). The resulting GPT-4.1 generations have a significantly higher intragroup BERTScore (F1) than human impressions (0.6168 vs. 0.3500). This indicates that human responses are more variable than GPT-4.1 texts.

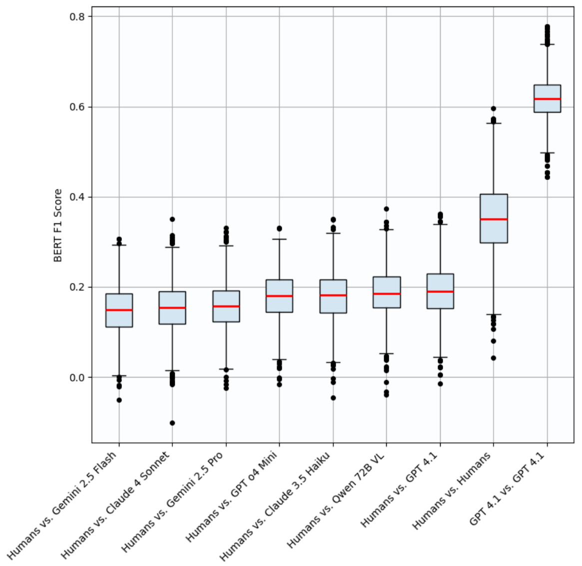


Figure 1: BERTScore (F1) similarity for pairs of human-human and human-GPT-4.1 texts.

5 GPT Detection

Having observed a significant difference in both intra- and inter-group BERTScore similarity between human impressions and GPT-4.1 outputs, we set out to test whether a BERT-style classifier could distinguish between the two.

Using the ChatGPT detector proposed by (Guo et al., 2023), we achieved an out-of-the-box F1 score of 0.947 on a sample of 206 human impressions and 70 GPT-4.1 responses. In the following sections, we will investigate the exact differences between the human and GPT texts that could explain such strong separation between the two groups.

6 N-Grams

To obtain a high-level overview of both the transcribed verbal responses and the GPT-generated text, we identified the most frequent N-grams (contiguous sequences of N words). N-grams were extracted by sliding a window of size N over the input sequence to form tuples of N consecutive words. We analyzed bigrams (N = 2) and trigrams (N = 3), which capture the most common two- and three-word sequences in the corpus. Among the 50 highest-frequency combinations (full list in Appendix D), four broad semantic classes emerged: (1) sentiment, (2) content, (3) grammatical fillers, and (4) location. The five most frequent bigrams in each class are shown in Table 1. We use color coding to depict the frequency of each bigram; higher color intensity reflects a more commonly occurring word sequence.

Several differences stand out. In the sentiment class, GPT-4.1 adopts more formal and less emotionally polarized descriptors, favoring terms such as *calm* and *scenic*, whereas human respondents rely on informal adjectives like *nice*. In the content class, model outputs gravitate toward abstract qualities like *modern*, and *urban*, and explicitly mention seasons (*autumn*). Participants, on the other hand, mention concrete elements visible in the scene, e.g., *construction site*, *mountains*, and *lake*. Synthetic responses also reference *sky* far more often than human comments. Finally, fewer location-related bigrams appear in the model's top-50 list. In contrast, participants frequently situate features with spatial adverbs, such as *left*, *right*, and *front*. Location-oriented trigrams are likewise scarce in GPT output, confirming this pattern.

To test whether N-grams can summarize participant impressions scene by scene, we extracted the ten most frequent content-related N-grams for each of the window views (Tables 2, 3, and 4). When these phrases are read alongside the corresponding images, they capture many of the scenes' salient visual details, indicating strong representational power for a simple frequency analysis. This observation motivated the subsequent use of N-gram features in our text-clustering workflow.

7 Text Clustering

The strong correspondence between the most salient visual elements in the window views and the highest-frequency N-grams prompted us to base our explainable clustering on content-related bigrams. The full procedure is summarized in Algorithm 1.

Figure 2 shows that this N-gram approach yields a moderate clustering structure (a silhouette score of 0.475 with 0.315 to 0.535 95% empirical confidence interval(CI) for participant impressions, and 0.442 score with 0.257 to 0.539 95% empirical CI for machine-generated responses). The empirical confidence intervals were constructed by running single-stage bootstrap resampling with replacement. More details on the bootstrapping methodology along with the resulting co-occurrence matrix are given in Appendix E.

In the human transcripts, bigram frequency separates the data into two main clusters: one dominated by *building* descriptors, the other by *moun-*

Bigram type	Study participants	GPT-4.1
Sentiment	is nice don't like nice to like that not nice nice and	calm and a scenic and inviting a lively a vibrant
Content	construction site the construction the mountains mountains and the lake	a modern modern urban urban or late autumn autumn or
Grammatical	a lot lot of the view is not like the	This image image depicts depicts a The overall overall atmosphere
Location	in front the left the right front of the back	the background, along the either side side of foreground, there

Table 1: Five most frequent bigrams by category for study participants vs. GPT-4.1

tain references, plus an outlier (Scene 9) characterized by the word *construction*. Interestingly, within the *building* cluster, participants frequently mention site-specific entities, such as *Rolex* and *Point Vélo*. These terms do not appear in the GPT-4.1 completions, presumably because they reflect campus-specific jargon familiar to the participants but underrepresented in the model's training corpus. Additionally, the *building* cluster includes Scene 5, whose dominant bigram is *the mountains*. However, because the frequency of this bigram is very low, Scene 5 lies near the border yet remains in the *building* cluster.

GPT-4.1, by contrast, sorts its responses into three groups: (1) a cluster centered on the bigram with the adjective *modern*, (2) a heterogeneous *miscellaneous* cluster, and (3) a single outlier, Scene 9, characterized by the word *scaffolding*. The dominant bigrams in these clusters differ markedly from those in the human text, indicating that the model foregrounds visual features other than the ones participants find most noteworthy. This divergence underscores a distinct pattern in human perception of window views that is not fully captured by the language model. At the same time, Scene 9 appears as an outlier for both study participants and GPT-

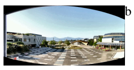
No.	Scene	Study participants	GPT-4.1
1	 a	the trees the buildings the road and cars trees and buildings are people and and buildings nature and and trees	late autumn autumn or or winter. a modern winter. The or office The buildings few cars the road, urban or
2	 a  b  c	the building buildings and the trees the buildings the road trees and grey buildings and trees people walking with people	The sky sky is an urban modern buildings autumn or a fisheye late autumn urban scene few people urban or
3	 a	the mountains mountains and the building building in big building building is of cars cars and and people people and	or research Polytechnique Fédérale Fédérale de few people a modern, a modern urban campus concrete and university or modern urban
4	 a  b  c	the mountains the trees the buildings trees and buildings are to work the sky people walking of people the tree	a modern, campus or or business The sky sky is business park open campus The area a wide-angle wide-angle or
5	 a  b  c	the mountains the building buildings and the buildings mountains in front of the window to work open space of people	a modern campus or modern campus or institutional The sky sky is windows and The area contemporary buildings buildings with
6	 a  b  c	the mountains mountains and the lake the buildings lake and mountains in and mountains to work and lake open space	campus or a modern contemporary buildings modern campus sky is The sky or institutional with contemporary few people people walking

Table 2: Ten most frequent content-bearing bigrams extracted from participants' descriptions of each window view and GPT-4.1 generations. Scenes 1-6 under three sky conditions: (a) any sky, (b) clear sky, and (c) overcast sky.

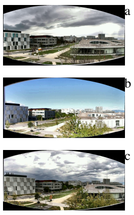
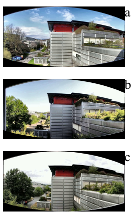
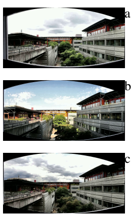
No.	Scene	Study participants	GPT-4.1
7		the mountains mountains and the building the rolex buildings and mountains in the road trees and building on colors and	modern campus campus or few people a clear a bright, or institutional mountains under and outdoor buildings on trees and
8		the building buildings and the buildings the road and cars buildings are to work is grey the cars cars passing	a modern modern urban urban or or campus lines and few people The sky sky is is overcast, clean lines
9		construction site the construction the building the road a construction buildings are very grey grey and site and of noises	scaffolding and a modern modern urban a curved a parked with scaffolding under construction a person a crane urban or
10		the mountains the building the rolex buildings and the buildings buildings are roof of the roof open space an open	a modern campus or modern campus The sky sky is or research contemporary buildings or business The buildings lines and
11		the building the trees the buildings front of building in trees and building is the window to work greeneries and	a modern building with a wide-angle wide-angle or a rooftop modern building trees and The sky or fisheye panels and
12		the building buildings and the trees the buildings trees and buildings are the colors to work the bridge the sun	a modern or office The sky sky is trees and university or a university windows and The buildings campus or

Table 3: Ten most frequent content-bearing bigrams extracted from participants' descriptions of each window view and GPT-4.1 generations. Scenes 7-12 under three sky conditions: (a) any sky, (b) clear sky, and (c) overcast sky.

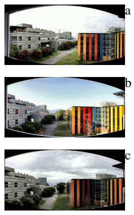
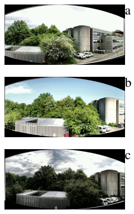
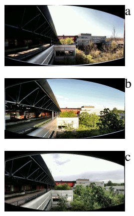
No.	Scene	Study participants	GPT-4.1
13		the mountains mountains and the lake the building the buildings mountains in building on buildings on the colors grey buildings	a modern vertical stripes stripes in orange, red, red, and modern urban campus or buildings with building with construction or
14		the building point velo the trees front of trees and buildings are to work and trees trees are trees in	a modern, a fenced green trees trees and building with wide-angle or a wide-angle white vehicles concrete building or institutional
15		the building the trees the buildings trees and of trees the window to work nature and the sun the red	a modern a university university or or office covered walkway modern architectural trees and wide-angle or a covered greenery and

Table 4: Ten most frequent content-bearing bigrams extracted from participants' descriptions of each window view and GPT-4.1 generations. Scenes 13-15 under three sky conditions: (a) any sky, (b) clear sky, and (c) overcast sky.

4.1, and both use construction-related terminology at high frequency.

8 Sentiment analysis

Next, we examine how scene content and sky condition shape the sentiment in both human transcriptions and GPT-4.1 responses. Sentiment is quantified as a continuous *Average Sentiment Score* score derived from a RoBERTa-based tripolar classifier (Loureiro et al., 2022) with (*positive*, *negative*, and *neutral* classes (see Appendix F.1 for details).

8.1 Effect of scene content

We first compare the *Average Sentiment Score* across the N-gram clusters (Figure 2). Sentiment is regressed on cluster ID, using *buildings* as the baseline for humans and *miscellaneous* for GPT-4.1. Ordinary Least Squares (OLS) coefficients show that, relative to the baseline, human texts are significantly more negative for the *construction* cluster

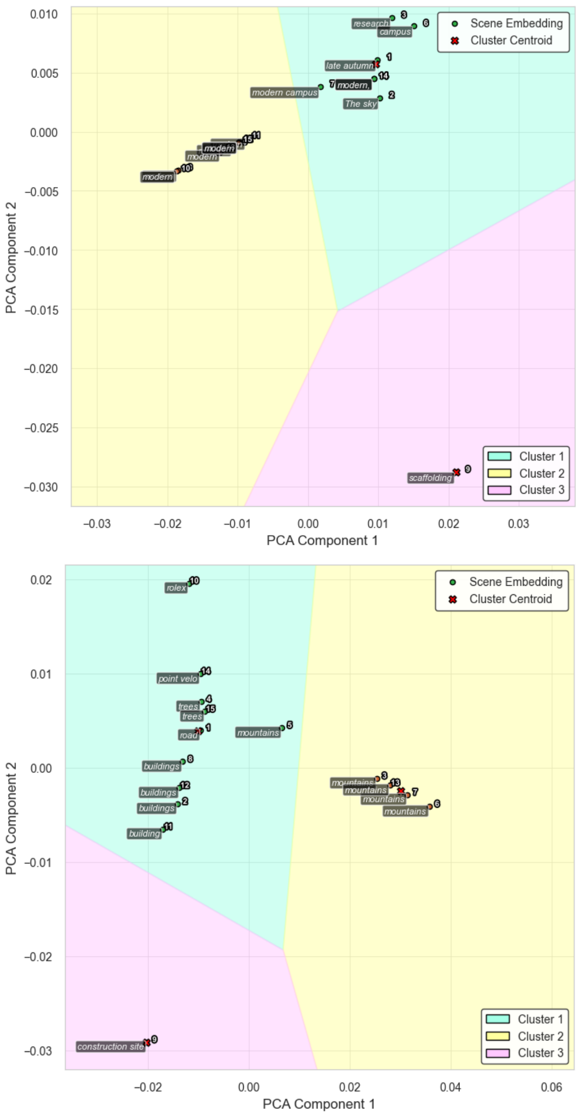


Figure 2: Bigram-based clustering of GPT-4.1 responses (top) vs. transcribed participant utterances (bottom)

($p < 0.001$, $c = -0.885$). GPT-4.1 completions for the same cluster show a similar negative shift ($p < 0.001$, $c = -0.341$). Captions that frequently include the adjective *modern* exhibit a small but significant positive offset in GPT-4.1 ($p = 0.022$, $c = 0.044$). Full coefficients are reported in Appendix G.

For a finer view-by-view analysis, we fit an OLS regression of the *Average Sentiment Score* on scene number (Appendix G), taking Scene 12, whose median sentiment approximates the overall median in both corpora, as the reference. Human and GPT outputs align on eleven of the thirteen scenes with significant coefficients. Both rate Scenes 4-7 (characterized by trees, open spaces, or mountains) above Scene 12, and Scenes 1-3, 9, and 14 (dominated by cars and/or limited openness) below it.

Algorithm 1 N-gram-based Clustering

Require: A set of scenes $S = \{s_1, s_2, \ldots, s_{15}\}$, Word2Vec model W2V
Ensure: Embeddings $\{v_s\}_{s \in S}$ and their clusters

for each scene $s \in S$ **do**
  // 1. Group impressions for s
  $T \leftarrow$ concatenate all responses for s
  // 2. Extract bigrams
  $B \leftarrow$ extract all bigrams from T
  Compute frequency $f(b)$ for each $b \in B$
  // 3. Get the most frequent content bigram
  $b^* \leftarrow \arg \max_{\substack{b \in B \\ \text{content-related}(b)}} \text{freq}(b)$
  // 4. Compute bigram embedding
  Let $W_{b^*} \leftarrow$ non-stopwords in bigram b^*
  $e_{b^*} \leftarrow \frac{1}{|W_{b^*}|} \sum_{w \in W_{b^*}} \text{W2V}(w)$
  // 5. Scene vector is frequency-normalized embedding of the most frequent bigram
  $F_{b^*} \leftarrow \frac{f(b^*)}{\sum_{b \in B} f(b)}$ // relative frequency
  $v_s \leftarrow F_{b^*} \cdot e_{b^*}$
end for
/* 5. Cluster scene vectors */
Apply agglomerative clustering to $\{v_s\}_{s \in S}$
/* 6. Visualization */
Reduce $\{v_s\}$ to 2D with PCA

Divergence occurs on Scenes 8 and 11, which feature a road and limited open space. Participants rate them below Scene 12, whereas GPT-4.1 rates them above, revealing a human-specific aversion to vehicle roads absent in GPT output.

Figure 3 presents violin plots of the *Average Sentiment Score* by scene, colored by N-gram cluster and ordered by increasing median sentiment. Two clear patterns emerge. First, human responses are much more polarized: they span the full -1 to $+1$ range, and Scenes 1-3 and 8-10 all have negative medians. Second, GPT-4.1 captions are skewed toward the positive; while they occasionally register negative values, the lowest score is only -0.092, whereas the most negative human score reaches -0.953. Together, these patterns indicate that GPT-4.1 is far less inclined than human observers to voice strongly negative impressions of the window views.

8.2 Effect of sky condition

We also examined whether the *Average Sentiment Score* varies with weather, comparing the clear ver-

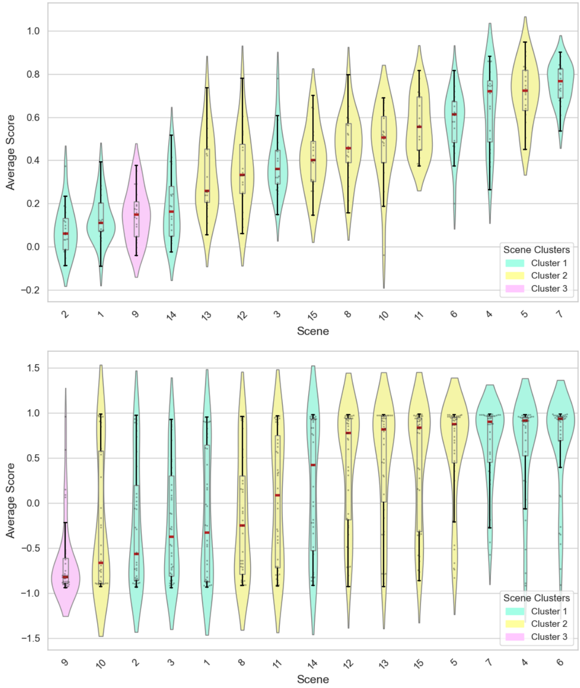


Figure 3: Sentiment score per scene number for GPT-4.1 responses (top) and study-participant impressions (bottom)

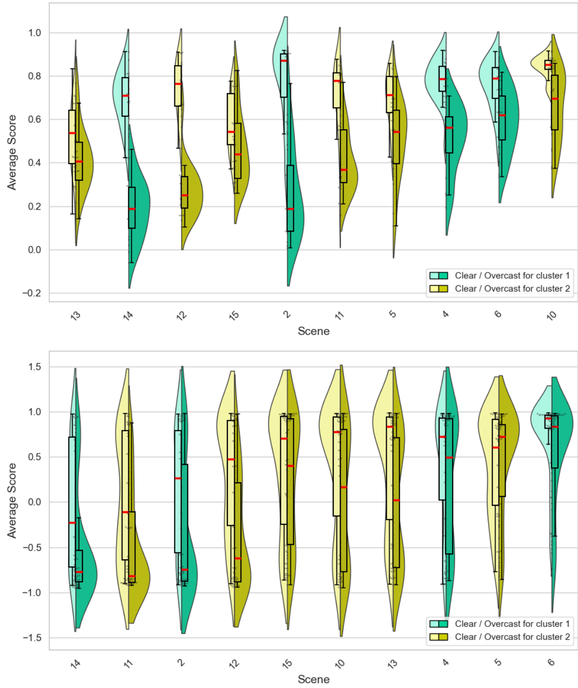


Figure 4: Sentiment by scene number and type for GPT-4.1 responses (top) and transcribed participant utterances (bottom)

sus overcast skies visible in the captured window scenes. The violin plots for both GPT completions and participant texts show lower sentiment scores for overcast scenes (Figure 4). This pattern is confirmed by a statistically significant negative coefficient for the *overcast* indicator in GPT-generated descriptions ($p < 0.001$, $c = -0.281$) and an even larger negative coefficient in the human responses ($p < 0.001$, $c = -0.326$); see Appendix G for full results. Thus, sky condition accounts for additional variance in sentiment beyond scene identity, with a markedly stronger impact on human language than on GPT output. Additionally, the R^2 statistic is higher for GPT-4.1 generations than in human responses (0.513 vs. 0.190), indicating that scene number and scene type explain a larger proportion of variance in text sentiment for machine-generated texts.

8.3 Word-level sentiment extraction

Finally, to reveal how both study participants and the GPT-4.1 model encode sentiment, we applied an ablation-based word-level sentiment identification method (see Appendix F.2.2 for details on the ablation procedure and performance comparison against DecompX (Modarressi et al., 2023) and Randomized Path-Integrations (Barkan et al., 2024)). Tables 5 and 6 list the top 10 most influential words for the sentiment classification. The analysis spans all scenes and weather conditions, spotlighting the terms that contribute most strongly—positively or negatively—to overall text sentiment.

For positively rated scenes, GPT-4.1 adopts a relatively formal style, emphasizing *striking architecture* and a *peaceful* atmosphere in Scenes 13 and 5. The absolute word-level importance scores (Appendix F.2.1) are roughly three times smaller than those in participants' texts, indicating milder phrasing. By contrast, human participants favor plainly positive adjectives, such as *nice*, *beautiful*, and *great*.

For the negatively rated scenes, human texts continue to use strongly charged adjectives, with *boring*, *ugly*, and *ruining* contributing the most to negative sentiment. Participants also negate otherwise positive descriptors, for instance, describing Scene 2 as *less pleasant* and mentioning that Scene 8 *wouldn't be an ideal place to work*. GPT-4.1, however, tends to choose intrinsically negative adjectives, such as (*muted* and *subdued*).

Figure 5 highlights the difference in the distribution of word importance between human texts and GPT-4.1 responses. Removing up to five words with the strongest sentiment from the transcribed human

Study participants	GPT-4.1
pleasant	striking
peaceful	peaceful
nice	pleasant
beautiful	lush
interesting	spacious
shining	calm
love	greenery
like [this view]	well-maintained
really [like the mountains]	innovative
great	day

Table 5: Words with strongest impact on sentiment in positively rated scenes

Study participants	GPT-4.1
boring	contrast
ugly	overcast
less [pleasant]	muted
ruining	obscuring
nothing [particularly interesting]	distorted
uncomfortable	subdued
depressing	overall
special	grey
wouldnt [be an ideal place]	metal
grey	cloudy

Table 6: Words with strongest impact on sentiment in negatively rated scenes

utterances causes a larger drop in accuracy than for GPT-4.1-generated text, implying that GPT-4.1 spreads sentiment more evenly across its generated tokens.

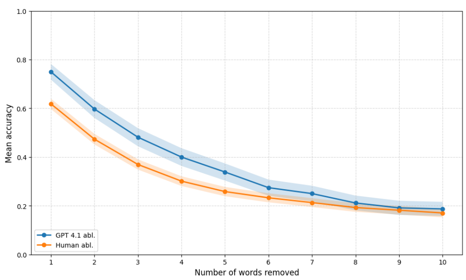


Figure 5: The average sentiment classification accuracy after removing the top 1–10 most impactful words, as identified by the ablative token attribution method. Shaded regions indicate 95% confidence intervals. The blue line represents GPT-4.1 generations, and the orange line denotes human-written texts.

9 Discussion and Conclusion

This work analyzed the open-ended descriptions collected in (Cho et al., 2023b, 2025a,b), and compared them with GPT-4.1 completions for the same window-view scenes. The goal was to isolate aspects of view-out perception that are genuinely human and currently absent from a state-of-the-art multimodal transformer.

Unstructured texts were first explored through the most frequent bi- and trigrams, which fell naturally into four semantic categories. With a simple, explainable bigram-based clustering, we identified the objects that most shaped each account. Human responders referred most often to *mountains*, *lake*, and *construction*, whereas GPT-4.1 emphasized the *sky* and abstract architectural qualities like *modern* and *urban*. Further, GPT-4.1's descriptions never singled out mountains, and they omitted several named entities that appeared regularly in human speech.

Sentiment analysis with a RoBERTa classifier revealed far stronger polarity in the human texts. Participants expressed clear dislike for scenes containing construction sites, cars, or limited open space, and clear preference for those with nature or open spaces. GPT-4.1, in contrast, produced only mildly positive sentiment across all scenes. Deviations from the baseline Scene 12 were nevertheless directionally similar between the two corpora, except for Scenes 8 and 11, whose lower human sentiment was not matched by the GPT model. When sentiment was regressed on the weather, both corpora showed lower scores for overcast images, but the effect size was over 16% larger in the human data.

Word-level ablation confirmed these stylistic differences: GPT-4.1 relied on formal adjectives such as *spacious* or *well-planned*, with very small attribution weights; whereas participants injected emotion through everyday adjectives (*nice*, *great*) and especially through the negation of positive terms (*less pleasant*, *nothing interesting*). Together, the findings show that open space and natural elements (mountains, trees, lake) drive a positive affect, while construction, roads, and visual clutter depress it, and that the transformer model captures this pattern only partially. Therefore, while GPT-4.1 can capture the broad directional trends observed here, its muted tone and key omissions expose clear limits. At present, it cannot replace human judgment when nuanced appraisal of window-view quality is required.

10 Limitations

The present analysis is based on a relatively small corpus—fifteen distinct window-view locations and 2100 verbal responses collected across two VR experiments (Cho et al., 2023b, 2025a,b). Replicating the workflow on a larger, demographically broader sample and on more varied scenery (e.g., different climates, building typologies, and degrees of familiarity) will be essential before generalizing the findings.

In addition, our text-clustering pipeline has scalability issues. It still depends on manual labeling of salient N-grams; with hundreds of scenes, this step would become labor-intensive and susceptible to coder drift. Moreover, the current frequency-based clustering is sensitive to outlier strings: a participant who copies the same sentence repeatedly, or injects unrelated content, can distort the cluster geometry and bias sentiment estimates. Future versions should incorporate automated noise filtering and topic-modeling techniques that are less vulnerable to adversarial or low-effort inputs.

Furthermore, we have not yet explored varying the system and user prompts to better align the VLMs' responses with human window-view impressions. We hypothesize that prompt optimization techniques, such as TextGrad (Yuksekgonul et al., 2024), could yield more human-like completions, e.g., by prompting for "use colloquial language". This could reduce the divergence between model and human responses.

Finally, we note that the introduced ablative word-level sentiment attribution approach perturbs the syntax and can inflate the importance of function words.

References

Fedaa Abd-Alhamid, Michael Kent, and Yupeng Wu. 2023. Quantifying window view quality: A review on view perception assessment and representation methods. *Building and Environment*, 227:109742.

Myriam B.C. Aries, Jennifer A. Veitch, and Guy. R. Newsham. 2010. Windows, view, and office characteristics predict physical and psychological discomfort. *Journal of Environmental Psychology*, 30(4):533–541.

Oren Barkan, Yehonatan Elisha, Yonatan Toib, Jonathan Weill, and Noam Koenigstein. 2024. Improving LLM attributions with randomized path-integration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9430–9446, Miami, Florida, USA. Association for Computational Linguistics.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. *Preprint*, arXiv:2503.12329.

Yunni Cho, Caroline Karmann, and Marilyne Andersen. 2023a. Dynamism in the context of views out: A literature review. *Building and Environment*, 244:110767.

Yunni Cho, Caroline Karmann, and Marilyne Andersen. 2023b. A vr-based workflow to assess perception of daylit views-out with a focus on dynamism and immersion. *Journal of Physics: Conference Series*, 2600(11):112002.

Yunni Cho, Caroline Karmann, and Marilyne Andersen. 2025a. Daylight dynamics and view perception in virtual reality: the impact of sky conditions and weather variations.

Yunni Cho, Caroline Karmann, and Marilyne Andersen. 2025b. Perception of window views in vr: Impact of display and type of motion on subjective and physiological responses. *Building and Environment*, 274:112757.

Kiran Maini Gerhardsson and Thorbjörn Laike. 2021. Windows: a study of residents' perceptions and uses in sweden. *Buildings & Cities*, 2(1).

Fang-Ying Gong, Zhao-Cheng Zeng, Fan Zhang, Xiaojiang Li, Edward Ng, and Leslie K. Norford. 2018. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment*, 134:155–167.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

WingYin Ha and Diarmuid O'Donoghue. 2024. Comparing human and machine generated text for sentiment. pages 335–342.

S. Herbold, A. Hautli-Janisz, U. Heuer, and 1 others. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13:18617.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *CoRR*, cs.CL/0205028.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *Preprint*, arXiv:2202.03829.

Thomas A Markus. 1967. The function of windows—a reappraisal. *Building Science*, 2(2):97–121.

Barbara Szybinska Matusiak and Christian A. Klöckner. 2016. How we evaluate the view out through the window. *Architectural Science Review*, 59(3):203–211.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. Decompx: Explaining transformers decisions by propagating token decomposition. *Preprint*, arXiv:2306.02873.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. *Preprint*, arXiv:2103.13413.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Nikitas Theodoropoulos, Giorgos Filandrianos, Vassilis Lyberatos, Maria Lymperaiou, and Giorgos Stamou. 2025. Berttime stories: Investigating the role of synthetic story data in language pre-training. *Preprint*, arXiv:2410.15365.

Emmy van Esch, Robert Minjock, Stephen M Colarelli, and Steven Hirsch. 2019. Office window views: View features trump nature in predicting employee well-being. *Journal of environmental psychology*, 64:56–64.

Yixi Xia, Nobuyoshi Yabuki, and Tomohiro Fukuda. 2021. Sky view factor estimation from street view images based on semantic segmentation. *Urban Climate*, 40:100999.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic "differentiation" via text. *Preprint*, arXiv:2406.07496.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Text Generation

To produce machine-generated impressions for each of the scene-condition combinations, we used the following user prompt.

> **Prompt:** *"In a few sentences, could you describe your overall impressions of this image?"*

No system prompt was provided. As for the text generation parameters, we set the default temperature setting of 1.0. For Gemini 2.5 Pro, we set the thinking budget to 1024 tokens. Meanwhile, for the OpenAI models, the output length was capped at 512 tokens. We obtained multiple synthetic impressions per view-out scene by repeating the same request 20 times for GPT-4.1. For the other models, a single response was collected for each input image.

B Model IDs

In this section, we report the model identifiers, stable release dates, or API call dates, depending on the information available for each model. When a stable release date was not explicitly listed, we provided the most relevant alternative.
To avoid conflating identifiers with dates, we report for each model:

(i) The exact API model ID we used (when available),

(ii) The *reference type* indicating what the date represents (stable release vs. usage),

(iii) The ISO date itself (YYYY-MM-DD).

If the provider exposes a dated API model ID (i.e., the ID includes a YYYY-MM-DD suffix), we list that full ID and take the suffix as the reference date. If no dated ID is available but a stable release date is published, we report the stable release date. If neither is available, we report the first date we used the model in our experiments. Table 7 summarizes these details.

Model (family)	Exact API model ID	Reference type	Date
Claude 3.5 Haiku	`claude-3-5-haiku-20241022`	Dated model ID	2024-10-22
Claude Sonnet 4	`claude-sonnet-4-20250514`	Dated model ID	2025-05-14
GPT-4.1	`gpt-4.1-2025-04-14`	Dated model ID	2025-04-14
o4-mini	`o4-mini-2025-04-16`	Dated model ID	2025-04-16
Gemini 2.5 Pro	`gemini-2.5-pro`	Stable release	2025-06-17
Gemini 2.5 Flash	`gemini-2.5-flash`	Stable release	2025-06-17
Qwen2.5-VL-72B-Instruct	`Qwen/Qwen2.5-VL-72B-Instruct`	Usage date	2025-07-13

Table 7: Models, exact API IDs, and the date associated with each entry. "Dated model ID" means the ID itself carries the YYYY-MM-DD suffix, which we use as the reference date.

C BERT Score Calculation

In this study, we calculate intragroup similarity for human/human and GPT-4.1 / GPT-4.1 texts, along with inter-group similarity for VLM/human texts using BERTScore (F1). We use the latest version of HuggingFace's *distilbert-base-uncased* model available as of July 23, 2025, as the backbone. When computing intragroup similarity, we exclude pairs of identical texts. For instance, for a given human impression of scene 7 with a clear sky, we compute the similarity with every other human impression for this scene-condition combination. The full procedure for computing BERT Score similarity is outlined in Algorithm 2.

Algorithm 2 BERT Score Calculation

Require:
- H: set of human responses
- G: set of GPT-4.1 responses
- $V = \{V_1, V_2, \ldots, V_K\}$: sets of responses from K other VLMs
- C: set of scene-conditions
- $M : \left(H \cup G \cup \bigcup_{i=1}^{K} V_i\right) \rightarrow C$, a mapping which assigns each response its scene-condition
- Pretrained function $\mathrm{BERTScore}(r_a, r_b)$

Ensure: A dictionary $\mathcal{S}$ of BERT scores for selected group pairs

```
Initialize empty dictionary S
for all group pairs (X, Y) ∈ {(H, H), (G, G), (G, H)} ∪ {(V_i, H) | i = 1, ..., K} do
    Initialize S[X, Y] ← ∅
    for all responses r ∈ X do
        for all responses s ∈ Y with s ≠ r do
            if M(r) = M(s) then
                S[X, Y] ← S[X, Y] ∪ {BERTScore(r, s)}
            end if
        end for
    end for
end for
return S
```

D Top 50 Bi- and Trigrams

In Tables 8, and 9 we present the full set of 50 most common bi- and trigrams extracted from both human window view impressions and GPT texts. Each N-gram is color-coded to depict its frequency, with a higher saturation implying a more commonly occurring word sequence.

E Cluster Stability Estimation

To evaluate the robustness of the clustering patterns, we generated 1,000 bootstrap samples—each consisting of utterances or completions selected with replacement—from both human and GPT responses. Figure 6 presents a co-occurrence matrix whose entry $M_{(i,j)}$ stores the number of samples in which scenes i, j such that $i < j$ occurred in the same cluster.

Looking at the results for human texts, we can see

Trigram type	Study participants	GPT-4.1
Sentiment	which is nice it is nice	-
Content	the construction site the mountains and mountains in the and the lake mountains and the the lake and the point velo building in front building on the lot of cars buildings on the the roof of a construction site the mountains in of the rolex the building on roof of the see the mountains construction site and construction site which lot of trees lake and mountains mountains and lake cars and people the big building and the mountains	a modern urban late autumn or modern urban or urban or campus The sky is depicts a modern modern campus or scaffolding and a a modern campus a few people Polytechnique Fédérale de autumn or winter. vertical stripes in a university or with scaffolding and sky is overcast, campus or institutional campus or business building with a clean lines and parked along the a modern, open university or research shows a modern and a person a person walking a curved road, or winter. The or campus setting a modern architectural modern urban campus a modern building a wide-angle or or business park mountains under a a business or and a crane
Grammatical	a lot of I can see to look at it is not the view is are a lot with a lot can see the of the view there is not I don't like is not much with not much as well as like I am	This image depicts image depicts a The overall atmosphere atmosphere. In the This image shows image shows a suggesting it is There are several The scene is
Location	on the right on the left in front of building in front in the back front of the the left and in front and	In the background, In the foreground, along the street, On the left,

Table 8: 50 most frequent trigrams for study participants and GPT-4.1

Bigram type	Study participants	GPT-4.1
Sentiment	is nice	-
	don't like	
	nice to	
	like that	
	not nice	
	nice and	
Content	construction site	modern urban
	the construction	urban or
	the mountains	late autumn
	mountains and	autumn or
	the lake	or campus
	the building	lines and
	the rolex	scaffolding and
	buildings and	The sky
	point velo	modern campus
	the trees	campus or
	the buildings	sky is
	lake and	or research
	mountains in	or winter.
	the road	a curved
	and cars	or office
	the point	vertical stripes
	building in	few people
	a construction	building with
	big building	stripes in
	trees and	contemporary buildings
	buildings are	Polytechnique Fédérale
	building on	Fédérale de
		or business
		a parked
		a university
		university or
		with scaffolding
		under construction
		a person
		winter. The
		orange, red,
		red, and
		is overcast,
		a clear
		a bright,
		or institutional
Grammatical	a lot	This image
	lot of	image depicts
	the view	depicts a
	is not	The overall
	like the	overall atmosphere
	can see	atmosphere. In
	not much	image shows
	a bit	shows a
	see the	The scene
	to see	
	view is	
	to look	
	look at	
	it feels	
	are not	
	this view	
	I feel	
Location	in front	the background,
	the left	along the
	the right	
	front of	
	the back	

Table 9: 50 most frequent bigrams for study participants and GPT-4.1

that scene 9 doesn't co-occur with any other scene in over 60% of the bootstrap samples, highlighting the fact that the construction taking place in it sets this scene apart. Meanwhile, the pair of scenes 7 and 13 is the most frequently co-occurring, due to their shared references to mountains and their physical proximity. Similarly, scenes 11 and 12 often co-occur, as both are characterized by the presence of buildings and trees.

For GPT-generated responses, scenes 4 and 14 co-occur in 997 out of 1,000 bootstrap samples, reflecting their emphasis on the *modern* qualities of the university campus. The next most frequent pair is scenes 5 and 10, as their descriptions often refer to *modern* architectural styles.

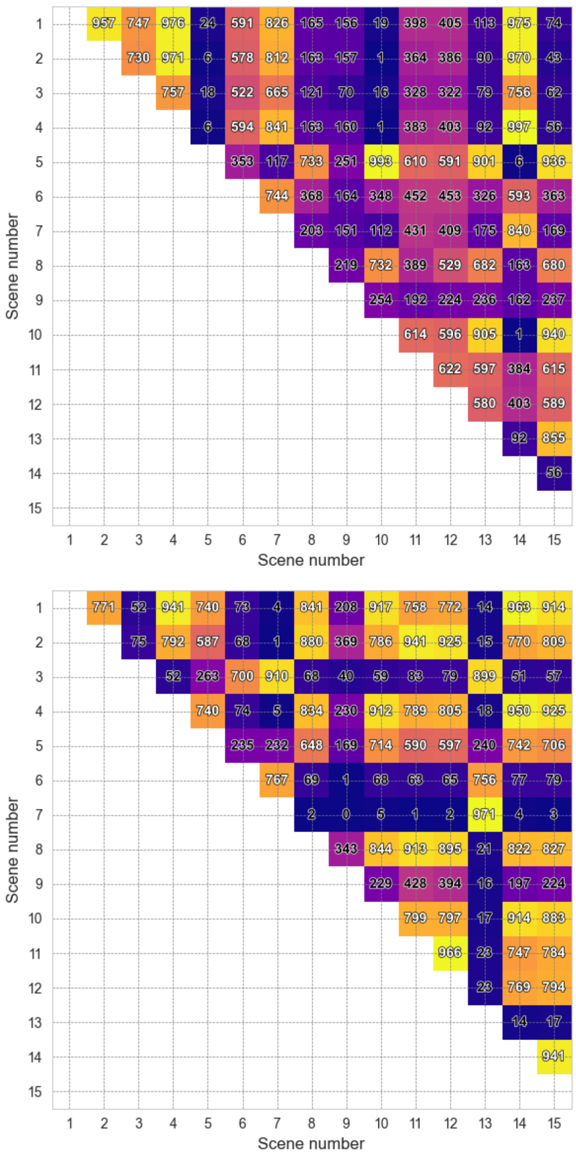


Figure 6: Bootstrap-based co-occurrence matrices for GPT-4.1 texts (top) and human impressions (bottom)

F Sentiment Analysis

F.1 Text-Level Sentiment

To provide a comprehensive assessment of the sentiment of a given text, we define the *Average Sentiment Score* as follows. Let x denote the input text and let $f_{\text{sent}}(x)$ be the tripolar RoBERTa sentiment classifier (Loureiro et al., 2022). We used the latest version (*cardiffnlp/twitter-roberta-base-sentiment-latest*) available on HuggingFace as of July 13, 2023. This classifier outputs a probability distribution over the positive, neutral, and negative sentiment labels. Specifically, let $p_{\text{pos}}(x)$ and $p_{\text{neg}}(x)$ denote the probabilities assigned to the positive and negative classes, respectively. The *Average Sentiment Score*, $S(x)$, is then defined as:

$$S(x) = p_{\text{pos}}(x) - p_{\text{neg}}(x)$$

This score captures the net polarity of the text, ranging from -1 (maximally negative) to 1 (maximally positive), thus providing a holistic measure of overall sentiment.

F.2 Word-Level Sentiment

In this study, we investigate which words have the greatest impact on sentiment classification of human texts and GPT-4.1 responses. To this end, we evaluate two existing state-of-the-art token attribution methods, namely Randomized Path-Integrations (Barkan et al., 2024) and DecompX (Modarressi et al., 2023), as well as an ablative sentiment attribution approach. To ensure the relevance of our analysis, we exclude English stop words as defined by the NLTK library (Loper and Bird, 2002).

F.2.1 Ablative Sentiment Attribution

To compute a context-aware sentiment attribution score for each word in a verbal response, we can use an ablation-based approach. For each word w_i in the response $R = (w_1, w_2, \ldots, w_n)$, we first compute the *Average Sentiment Score* of the full response, denoted $S(R)$. Then, we compute the *Average Sentiment Score* of the response with w_i removed, denoted $S(R_{\setminus i})$, where $R_{\setminus i}$ is the response with the i-th word omitted. We define the sentiment attribution score for w_i as the difference:

$$A(w_i) = S(R) - S(R_{\setminus i}),$$

where $A(w_i)$ quantifies the contribution of w_i to the overall sentiment of the response, in the context of the surrounding words. This attribution score reflects the extent to which each word influences the sentiment prediction, leveraging the contextual sensitivity of self-attention mechanisms (as in RoBERTa). The described ablative attribution method is closely related to the perturbation idea introduced in the Local Interpretable Model-agnostic Explanations (LIME) framework (Ribeiro et al., 2016).

As a result of applying the ablative sentiment attribution method, we obtain the per-word scores in human and GPT texts presented in Tables 10, and 11.

Study participants		GPT-4.1	
Word	**Score**	**Word**	**Score**
pleasant	+1.54	striking	+0.39
peaceful	+1.38	peaceful	+0.35
nice	+1.31	pleasant	+0.34
beautiful	+1.29	lush	+0.34
interesting	+1.26	spacious	+0.34
shining	+1.16	calm	+0.31
love	+1.15	greenery	+0.30
like [this view]	+1.09	well-maintained	+0.25
really [like the mountains]	+1.04	innovative	+0.25
great	+1.04	lush	+0.24

Table 10: Words with the strongest impact on sentiment in positively rated scenes.

Study participants		GPT-4.1	
Word	**Score**	**Word**	**Score**
boring	-1.62	contrast	-0.35
ugly	-1.59	overcast	-0.25
less [pleasant]	-1.52	muted	-0.23
ruining	-1.51	overcast	-0.20
nothing [particularly interesting]	-1.44	obscuring	-0.19
uncomfortable	-1.30	distorted	-0.19
depressing	-1.27	subdued	-0.18
special	-1.23	overall	-0.18
wouldnt [be an ideal place]	-1.22	grey	-0.17
grey	-1.20	metal	-0.17

Table 11: Words with the strongest impact on sentiment in negatively rated scenes.

F.2.2 Comparison with other token attribution methods

To evaluate the different token attribution methods, we assess sentiment classification accuracy after sequentially removing the top 1–10 most impactful words as identified by each method. For each removal step, we report both the mean accuracy and the 95% confidence interval. Lower accuracy after the word removal suggests that the corresponding attribution method more effectively pinpoints words that are crucial for sentiment classification. Panels (**E**) and (**F**) in Figure 7 demonstrate that the ablative sentiment attribution approach consistently yields lower classification accuracy than all other methods for the removal of the first five words. DecompX ranks second, while the Randomized Path-Integration (RPI) methods perform considerably worse: comprehensiveness-based RPI occupies third place and sufficiency-based RPI fourth. This ranking is observed for both human responses and GPT-4.1 outputs.

Moreover, panels (**A**) to (**D**) reveal that, across all four token attribution methods, classification accuracy declines more rapidly for human texts than for GPT-4.1 responses during the removal of the first five words. This observation suggests that human-written texts tend to concentrate sentiment within a few key words, whereas GPT-4.1 distributes sentiment more evenly across the text.

G Regression Analysis

To investigate the relationship between *Average Sentiment Score* and various categorical predictors, we conduct a series of Ordinary Least Squares (OLS) regression analyses. The categorical predictors are one-hot encoded. We consider three different predictor combinations:

- **Cluster ID:** Each unique scene-condition pair corresponds to one of three clusters.
- **Scene Number:** Analysis restricted to scene-condition combinations with condition fixed to *any sky* and human responses collected during the first experimental session reported in (Cho et al., 2023b, 2025a,b).
- **Scene Number and Scene Type:** Regression restricted to scene-condition pairs with condition limited to *clear* or *overcast* and human responses collected during the second experimental session conducted by (Cho et al., 2023b, 2025a,b).

The estimated coefficients and corresponding significance levels for each regression model are summarized in Tables 12, 13, and 14. Results are reported separately for human participants and the GPT-4.1 generations.

H Licensing

We use the dataset of human impressions of office-window views collected by (Cho et al., 2023b, 2025a,b), which is distributed under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license.[1] Consistent with this license, we credit the creators, link to the license, and note all modifications we make to the data. We will release our augmented dataset under CC BY 3.0, accompanied by a LICENSE file and an explicit TASL attribution (Title, Author, Source, License). Our code will be released under the MIT License to facilitate reuse.[2]

I Computing Infrastructure

All experiments reported in this work were performed on a single laptop machine. We used an Apple MacBook Pro equipped with the Apple M4 system-on-chip, and an integrated GPU. The machine has 16 GB of unified memory and a 512 GB solid-state drive. Further, we used Python 3.10. On this setup we ran:

1. **GPT detection** via the BERT-style classifier of Guo *et al.* (Guo et al., 2023).
2. **Similarity scoring** using BERTScore (Zhang et al., 2019).
3. **Sentiment analysis** with a RoBERTa-based classifier following Loureiro and Chen (Loureiro et al., 2022).
4. **Token attribution** methods, including the discussed ablative approach, DecompX (Modarressi et al., 2023) and Randomized Path Integrations (Barkan et al., 2024).

Because the M4 SoC does not support CUDA, all computations were run on the CPU. Typical end-to-end processing of the combined dataset completed within 24 hours per experiment.

[1] https://creativecommons.org/licenses/by/3.0/
[2] https://opensource.org/license/MIT

Term	Coefficient	p-value
Intercept	**0.2101**	< 0.001
Cluster 1 (vs. 0)	-0.0596	0.069
Cluster 2 (vs. 0)	**-0.8854**	< 0.001

Term	Coefficient	p-value
Intercept	**0.4881**	< 0.001
Cluster 1 (vs. 0)	**0.0439**	0.022
Cluster 2 (vs. 0)	**-0.3413**	< 0.001

Table 12: Estimated coefficients and significance levels from regressing *Average Sentiment Score* on cluster ID, for human responses (left) and GPT-4.1 generations (right).

Term	Coefficient	p-value
Intercept	**0.1260**	0.016
Scene 1 (vs. 12)	**-0.2121**	0.020
Scene 2 (vs. 12)	**-0.2240**	0.002
Scene 3 (vs. 12)	-0.1765	0.052
Scene 4 (vs. 12)	**0.3444**	< 0.001
Scene 5 (vs. 12)	**0.3857**	< 0.001
Scene 6 (vs. 12)	**0.5295**	< 0.001
Scene 7 (vs. 12)	**0.5925**	< 0.001
Scene 8 (vs. 12)	**-0.1887**	0.038
Scene 9 (vs. 12)	**-0.8014**	< 0.001
Scene 10 (vs. 12)	-0.1146	0.121
Scene 11 (vs. 12)	**-0.2486**	0.001
Scene 13 (vs. 12)	0.0824	0.266
Scene 14 (vs. 12)	**-0.2886**	< 0.001
Scene 15 (vs. 12)	**0.1469**	0.047

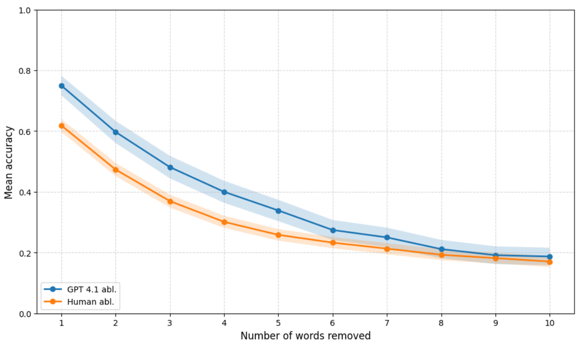
Term	Coefficient	p-value
Intercept	**0.3541**	< 0.001
Scene 1 (vs. 12)	**-0.2255**	< 0.001
Scene 2 (vs. 12)	**-0.2795**	< 0.001
Scene 3 (vs. 12)	0.0347	0.466
Scene 4 (vs. 12)	**0.2931**	< 0.001
Scene 5 (vs. 12)	**0.3586**	< 0.001
Scene 6 (vs. 12)	**0.2248**	< 0.001
Scene 7 (vs. 12)	**0.4012**	< 0.001
Scene 8 (vs. 12)	**0.1125**	0.019
Scene 9 (vs. 12)	**-0.2073**	< 0.001
Scene 10 (vs. 12)	**0.1215**	0.011
Scene 11 (vs. 12)	**0.2084**	< 0.001
Scene 13 (vs. 12)	-0.0236	0.619
Scene 14 (vs. 12)	**-0.1641**	0.001
Scene 15 (vs. 12)	0.0538	0.258

Table 13: Regression coefficients and significance levels for predicting *Average Sentiment Score* by scene number, based on human responses (left) and GPT-4.1 generations (right). Analyses are restricted to impressions of images with any sky condition.

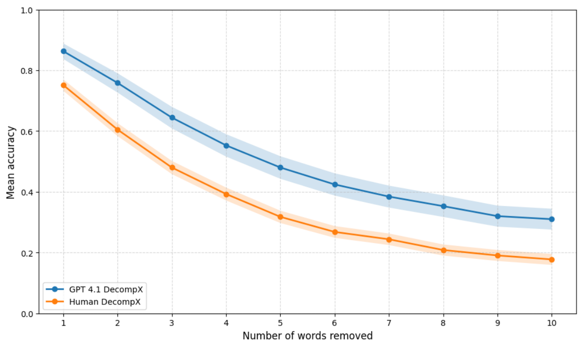
Term	Coefficient	p-value
Intercept	**0.4437**	< 0.001
Overcast (vs. Clear)	**-0.3261**	< 0.001
Scene 2 (vs. 15)	**-0.3689**	< 0.001
Scene 4 (vs. 15)	0.0330	0.747
Scene 5 (vs. 15)	0.1431	0.162
Scene 6 (vs. 15)	**0.3559**	0.001
Scene 10 (vs. 15)	-0.0668	0.514
Scene 11 (vs. 15)	**-0.4963**	< 0.001
Scene 12 (vs. 15)	**-0.3042**	0.003
Scene 13 (vs. 15)	-0.0775	0.448
Scene 14 (vs. 15)	**-0.5399**	< 0.001

Term	Coefficient	p-value
Intercept	**0.6623**	< 0.001
Overcast (vs. Clear)	**-0.2813**	< 0.001
Scene 2 (vs. 15)	-0.0056	0.881
Scene 4 (vs. 15)	**0.1209**	0.001
Scene 5 (vs. 15)	**0.0759**	0.042
Scene 6 (vs. 15)	**0.1550**	< 0.001
Scene 10 (vs. 15)	**0.2274**	< 0.001
Scene 11 (vs. 15)	0.0374	0.316
Scene 12 (vs. 15)	-0.0180	0.630
Scene 13 (vs. 15)	-0.0537	0.150
Scene 14 (vs. 15)	**-0.0782**	0.036

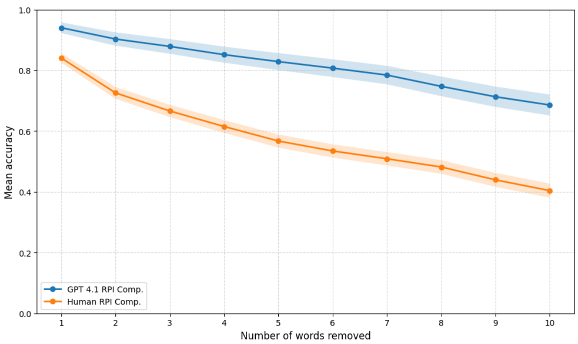
Table 14: Regression coefficients and significance levels for predicting *Average Sentiment Score* by scene number and type, based on human responses (left) and GPT-4.1 generations (right). Analyses are restricted to impressions of images with clear and overcast sky conditions.



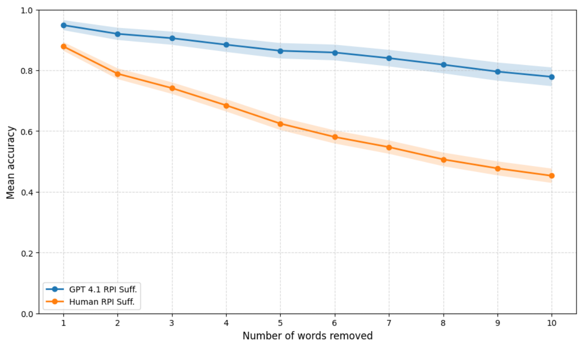
(**A**) Ablative method



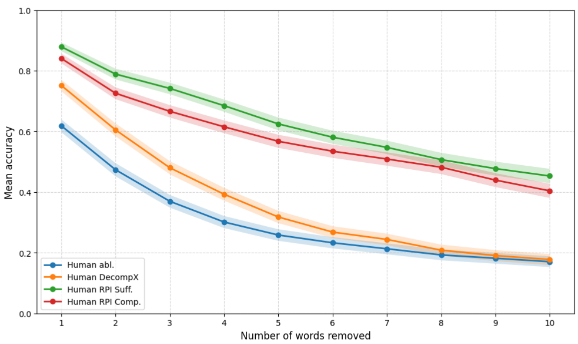
(**B**) DecompX method



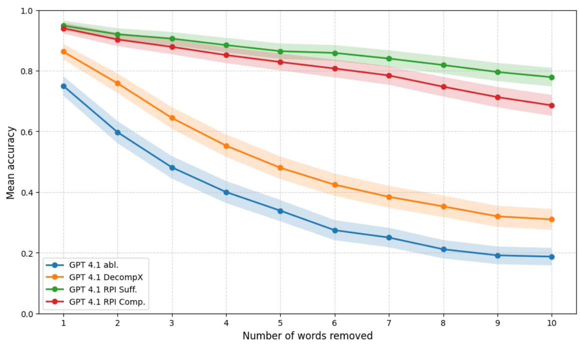
(**C**) RPI, maximizing Comprehensiveness



(**D**) RPI, maximizing Sufficiency



E) Human texts: all methods



F) GPT-4.1 texts: all methods

Figure 7: **Comparison of token attribution methods for sentiment analysis in human and GPT-4.1 texts.** Each panel shows the average sentiment classification accuracy after sequentially removing the top 1–10 most impactful words, as identified by four attribution methods: (**A**) Ablative, (**B**) DecompX, (**C**) Randomized Path-Integrations (Comprehensiveness), and (**D**) Randomized Path-Integrations (Sufficiency). Shaded regions indicate 95% confidence intervals. In panels (A)–(D), blue lines represent GPT-4.1 generations and orange lines represent human-written texts. Panels (E) and (F) summarize all four attribution methods for human and GPT-4.1 datasets, respectively.