



Universiteit  
Leiden  
The Netherlands

# Opleiding Informatica

Evaluating Emotional and Contextual Art Interpretation with Large Multimodal Models

Ramizcan Umut Seçkin

Supervisors:  
Rob Saunders & Derya Soydaner

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)  
[www.liacs.leidenuniv.nl](http://www.liacs.leidenuniv.nl)

15/01/2026

## Abstract

Recent advances in large multimodal models (LMMs) have enabled artificial intelligence systems to generate natural language descriptions conditioned on visual input. While these models perform well on factual image captioning and visual question answering, their ability to interpret artworks in an emotionally and context-aware manner remains limited. Art interpretation involves subjective elements such as emotional response, atmosphere, and cultural context, which are difficult to capture using purely descriptive approaches.

This thesis investigates how effectively modern LMMs can generate emotionally aligned and context-aware interpretations of artworks. Using the ArtEmis dataset, which links artworks to human emotional reactions and written explanations, this thesis evaluates model outputs under a structured set of prompting strategies. These include descriptive, emotional, art-historical and simplified interpretations. The study compares outputs from BLIP-2 and LLaVA models using a combination of automatic metrics, emotion alignment analysis and qualitative evaluation.

The results provide insight into the strengths and limitations of current multimodal models in emotional art interpretation and to identify prompt-engineering strategies that improve interpretive quality and consistency. Ultimately, this work contributes toward the development of more human-like AI systems for art understanding and interactive gallery applications.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Questions . . . . .	2
<b>2</b>	<b>Background and Related Work</b>	<b>2</b>
2.1	Emotional Interpretation of Art . . . . .	2
2.2	Large Multimodal Models . . . . .	2
<b>3</b>	<b>Methodology</b>	<b>3</b>
3.1	Dataset . . . . .	3
3.2	ArtEmis v1 and v2: Strengths and Limitations . . . . .	3
3.3	Models . . . . .	4
3.4	Evaluation Metrics . . . . .	5
3.5	Limitations of linguistic similarity metrics. . . . .	5
<b>4</b>	<b>Results</b>	<b>6</b>
4.1	Baseline . . . . .	6
4.1.1	Text-Only Baseline: DistilBERT . . . . .	6
4.1.2	Image-Only Baseline: ResNet-32 . . . . .	7
4.1.3	Summary of Baseline Findings . . . . .	7
4.2	Multimodal Models: BLIP-2 and LLaVA . . . . .	8
4.2.1	Quantitative Results . . . . .	8
4.2.2	Prompt Sensitivity Analysis . . . . .	9
4.2.3	Qualitative Observations . . . . .	10
4.3	Qualitative Case Study: <i>The Starry Night</i> (Van Gogh, 1889) . . . . .	10
4.3.1	Analysis . . . . .	11
4.4	Out-of-Distribution Qualitative Example: Astrophotography . . . . .	12
4.4.1	Analysis . . . . .	13
4.5	Summary of Multimodal Results . . . . .	13
<b>5</b>	<b>Discussion</b>	<b>14</b>
<b>6</b>	<b>Conclusion</b>	<b>15</b>
	<b>References</b>	<b>18</b>
<b>A</b>	<b>Prompt Sensitivity Examples</b>	<b>19</b>

# 1 Introduction

Artificial intelligence systems have made rapid progress in interpreting visual content, particularly through the development of Large Multimodal Models (LMMs) that combine vision and language. Models such as BLIP-2 [1] and LLaVA[2] have demonstrated strong performance on tasks including image captioning, visual question answering and multimodal reasoning by integrating visual encoders with large language models. These systems are increasingly capable of generating fluent natural language descriptions grounded in visual input[2].

However, when applied to artworks, many existing approaches still focus primarily on factual description or object-level recognition rather than emotional or contextual interpretation [3]. Artworks often involve ambiguity, symbolism, and subjective experience, making them fundamentally different from the everyday images used in standard vision-language benchmarks [4, 5]. Recent studies have highlighted that LMMs may rely heavily on learned linguistic priors, sometimes producing confident but weakly grounded interpretations or hallucinated details, particularly in visually complex domains [6, 7]. These limitations are empirically examined later in this thesis through quantitative emotion alignment metrics and qualitative case studies.

Art interpretation is inherently subjective and emotional [8]. Human viewers do not only describe what they see, but also relate artworks to feelings, atmosphere, cultural background, and personal associations[9]. Capturing these aspects is important for applications such as digital museums, educational tools and interactive gallery assistants. Despite their impressive language fluency, current AI systems often struggle with this challenge, producing outputs that are emotionally shallow, stylistically inconsistent, or poorly aligned with human emotional responses [3].

This thesis investigates whether modern LMMs can move beyond surface-level description and generate emotionally aligned and context-aware interpretations of artworks. Using the ArtEmis dataset [3], which links artworks to human emotional reactions and written explanations, this thesis systematically evaluates how different prompting strategies influence the quality, emotional accuracy and consistency of AI-generated interpretations. Specifically, the descriptive, emotional, art-historical, and simplified prompts are compared across multiple models and artworks.

The broader motivation behind this research is the development of an interactive “gallery buddy”: an AI companion that can accompany a visitor through a physical or digital art gallery and adapt its explanations to different emotional tones, levels of expertise, or audiences. By studying how prompt design and model choice affect emotional interpretation, this work aims to contribute to more human-like, reliable and emotionally aware AI systems for art understanding.

## 1.1 Research Questions

The main research question addressed in this thesis is:

*How effectively can current LMMs generate emotionally aligned and context-aware interpretations of artworks, compared to human-written interpretations from the ArtEmis dataset?*

To answer this question, the following sub-questions are investigated:

- How do different prompting strategies (descriptive, emotional, art-historical and simplified) influence the quality of generated art interpretations?
- To what extent can prompt design improve emotional accuracy and contextual richness in AI-generated interpretations?
- How consistent are model interpretations across visually and thematically similar artworks?

## 2 Background and Related Work

### 2.1 Emotional Interpretation of Art

Early research on computational art understanding primarily focused on tasks such as object recognition, style classification and factual description. While these approaches capture visual structure and content, they do not address how humans emotionally experience artworks [10]. The ArtEmis dataset introduced by Achlioptas et al. [3] represents a major step toward affective art understanding by linking artworks to emotion labels and natural-language explanations written by human annotators. Rather than asking annotators to describe images objectively, ArtEmis captures subjective emotional reactions, making it particularly suitable for studying emotional interpretation.

### 2.2 Large Multimodal Models

LMMs such as BLIP-2 [1] and LLaVA [2] combine visual encoders with Large Language Models (LLMs) to generate natural language conditioned on image input. By integrating pretrained vision backbones with powerful language models, these systems have achieved state-of-the-art performance on tasks including image captioning, visual question answering, and multimodal dialogue [5, 11]. Earlier work on vision–language alignment, such as CLIP [4], demonstrated that large-scale contrastive pretraining can effectively connect visual representations with natural language, laying the foundation for subsequent multimodal systems. More recent models, including KOSMOS-1 [12], further extend this by unifying perception, language understanding, and reasoning within a single framework.

Despite these advances, the majority of existing evaluations of LMMs focus on factual correctness, object recognition, or structured reasoning tasks. Comparatively little attention has been paid to subjective and interpretive tasks, such as emotional understanding, aesthetic judgment, or

contextual explanation in artworks. These tasks require models to move beyond surface-level visual features and integrate abstract concepts, affective language, and cultural context, which are only weakly grounded in the visual signal itself.

Recent studies have begun to highlight important limitations of LMMs in open-ended and interpretive settings. In particular, LMMs may generate confident, fluent, and emotionally expressive language even when visual evidence is ambiguous or insufficient, raising concerns about grounding and reliability [13]. Follow-up analyses show that grounding errors and hallucinations remain a persistent issue even in large-scale instruction-tuned vision–language models, including LLaVA [7]. Such behavior is especially relevant in the domain of art interpretation, where emotional responses are inherently subjective and visual cues may support multiple plausible readings.

Motivated by these observations, the present work focuses explicitly on emotional and context-aware interpretation rather than descriptive accuracy alone. By systematically evaluating BLIP-2 and LLaVA under controlled prompting strategies and comparing their outputs to human-written emotional explanations from the ArtEmis dataset, this thesis aims to provide a clearer understanding of the strengths, limitations, and failure modes of current multimodal models in affective art interpretation.

## 3 Methodology

### 3.1 Dataset

The primary dataset used in this study is ArtEmis, which contains over 80,000 artworks annotated with human emotional reactions. Annotators selected one of nine emotion categories and provided a short textual explanation describing their emotional response. Two versions of the dataset are considered: the original ArtEmis dataset and a contrastive version designed to emphasize emotional distinctions between visually similar artworks.

### 3.2 ArtEmis v1 and v2: Strengths and Limitations

The ArtEmis dataset is provided in two distinct versions [3, 14], referred to as ArtEmis v1 (original) and ArtEmis v2 (contrastive). While both versions aim to capture human emotional reactions to artworks, they differ significantly in structure, emotional distribution and intended use cases. Understanding these differences is essential for interpreting experimental results and for selecting evaluation strategies.

ArtEmis v1 consists of natural emotional reactions collected from human annotators viewing artworks independently. For each artwork, multiple annotators selected one of nine emotion categories and provided a short textual explanation describing their emotional response. As a result, many artworks in v1 exhibit mixed emotional distributions, where no single emotion is dominant. A major strength of ArtEmis v1 is its validity. The dataset reflects how people naturally respond to art, often experiencing multiple or ambiguous emotions simultaneously. This makes v1 particularly suitable for studying nuanced emotional interpretation and multi-emotion modeling.

However, this realism also introduces challenges. The dataset shows a strong bias toward positive emotions such as contentment, awe, and amusement, while negative emotions are underrepresented. In addition, the frequent presence of mixed emotions makes automatic evaluation more difficult, as there is often no clear ground-truth dominant emotion. These properties can lead to ambiguity when assessing emotional alignment and may obscure differences between model outputs.

ArtEmis v2 was introduced to address some of the limitations of the original dataset. In this version, artworks are paired in a contrastive manner, such that visually similar images are associated with clearly different emotional reactions. This design leads to sharper emotional distributions, with a higher proportion of artworks exhibiting a dominant emotion.

The primary advantage of ArtEmis v2 is its clarity. Emotional labels are more decisive, making the dataset better suited for supervised learning, automatic evaluation and benchmarking. The stronger presence of negative emotions such as sadness, fear and disgust also results in a more balanced emotional spectrum.

At the same time, the contrastive sampling strategy introduces its own limitations. Emotional reactions in v2 are less natural than in v1 and may exaggerate emotional differences between artworks. This can reduce realism and may encourage models to overfit to strong emotions rather than capturing subtle or ambiguous emotional responses.

In this thesis, both ArtEmis v1 and v2 are used to evaluate multimodal model behavior under different emotional conditions. ArtEmis v1 is primarily used to analyze how models handle natural, mixed and ambiguous emotional interpretations, while ArtEmis v2 enables clearer quantitative evaluation of emotional alignment and consistency.

By considering both dataset versions, this study aims to balance realism, evaluability and to provide a more complete picture of how LMMs interpret emotional and contextual aspects of art.

### 3.3 Models

Two LMMs are evaluated in this study:

- **BLIP-2**: Used to extract visual representations and generate intermediate captions, followed by language-based interpretation.
- **LLaVA**: Used to directly generate detailed interpretations conditioned on image input and prompt instructions.

Both models are evaluated under identical prompting conditions to ensure fair comparison.

### 3.4 Evaluation Metrics

Model outputs are evaluated using a combination of quantitative and qualitative methods designed to capture both linguistic similarity and emotional alignment:

- **Linguistic similarity metrics:** BLEU [15], ROUGE [16], and BERTScore [17] are used to measure textual overlap and semantic similarity between model-generated interpretations and human-written ArtEmis explanations.
- **Emotion alignment:** A RoBERTa-based classifier [18], trained on ArtEmis emotion labels, is used to predict emotion distributions from generated text.
- **Consistency analysis:** Sentence-level embeddings [19] are used to measure semantic similarity between interpretations of visually or thematically related artworks.
- **Qualitative analysis:** Manual inspection is conducted to identify common failure patterns such as emotional misalignment, superficial explanations, hallucinated details or inappropriate tone.

In contrast to the original ArtEmis baseline, which primarily evaluates models based on the dominant emotion label, this study explicitly accounts for secondary emotions present in the annotation distributions. Rather than treating emotional interpretation as a single-label classification task, model outputs are evaluated against full emotion distributions where possible.

To capture both distributional similarity and ranking quality, a combination of complementary metrics is used. Jensen–Shannon (JS) divergence [20] measures the similarity between predicted and human emotion distributions, providing a symmetric and bounded comparison of probabilistic outputs. In addition, Hit@2 [21] and nDCG@3 [22] are reported to assess whether models correctly prioritize dominant and secondary emotions, following standard practices in information retrieval and recommendation systems.

### 3.5 Limitations of linguistic similarity metrics.

While BLEU and ROUGE are widely used for evaluating natural language generation, they primarily measure surface-level n-gram overlap and are therefore limited in assessing emotional interpretation quality. In the context of art interpretation, multiple emotionally valid explanations may exist that differ lexically from reference texts, leading to artificially low scores despite meaningful emotional alignment. BERTScore partially alleviates this issue by incorporating contextual embeddings, but remains insensitive to emotional correctness when semantically similar text conveys a different tone. For this reason, linguistic similarity metrics are treated as complementary signals rather than primary indicators of emotional interpretation quality in this study.

Together, these evaluation metrics directly support the research questions by enabling a systematic analysis of how different prompting strategies and model choices affect emotional alignment, contextual richness and consistency of generated art interpretations relative to human responses.

## 4 Results

### 4.1 Baseline

This subsection presents baseline results obtained using text-only and image-only models. These baselines serve two purposes: first, to assess how much emotional information is present in textual and visual representations individually, and second, to provide reference points for interpreting the performance of multimodal models in subsequent experiments.

Model	Dataset	JS ↓	Hit@2 ↑	nDCG@3 ↑
Text-DistilBERT	v1	0.223	0.650	0.688
Text-DistilBERT	v2	<b>0.097</b>	<b>0.703</b>	<b>0.781</b>
Image-Baseline	v1	<b>0.141</b>	0.889	<b>0.593</b>
Image-Baseline	v2	0.156	<b>0.904</b>	<b>0.593</b>

Table 1: Baseline emotion prediction results on ArtEmis v1 and v2. Best values are shown in bold within each model across

Table 1 reveals clear differences between modalities and dataset versions. Text-only models achieve substantially lower Jensen–Shannon (JS) divergence than image-only models, indicating closer alignment with full human emotion distributions. In contrast, image-only models obtain relatively high Hit@2 scores, suggesting that they often identify a plausible dominant emotion, but exhibit weaker alignment with full emotion distributions, as reflected by consistently lower nDCG@3 values.

While Hit@2 measures whether the dominant human emotion appears among the top-ranked predictions, nDCG@3 additionally accounts for ranking quality and secondary emotional signals. As a result, nDCG@3 provides a more sensitive measure of nuanced emotional alignment, particularly for artworks associated with mixed or ambiguous emotional responses.

#### 4.1.1 Text-Only Baseline: DistilBERT

As a text-only baseline, a DistilBERT-based model [23] was trained to predict emotion distributions from human-written ArtEmis explanations. The model was evaluated on both ArtEmis v1 and ArtEmis v2 using JS divergence, Hit@2, and nDCG@3.

The results show a clear performance difference between the two dataset versions, primarily driven by differences in emotional ambiguity and label sharpness. On ArtEmis v1, the model exhibits relatively high JS divergence, reflecting the presence of mixed and ambiguous emotional labels. In contrast, performance on ArtEmis v2 is substantially stronger across all metrics, with lower JS divergence and higher nDCG@3, indicating a closer statistical match between predicted and ground-truth emotion distributions.

These findings suggest that textual explanations in ArtEmis v2 contain clearer and more separable emotional signals than those in v1. Importantly, the relatively strong Hit@2 and nDCG@3 scores on both datasets indicate that the model often captures secondary emotions correctly, even when the dominant emotion is ambiguous.

### 4.1.2 Image-Only Baseline: ResNet-32

To assess how much emotional information can be inferred from visual input alone, an image-only baseline was implemented using a ResNet-32 architecture [24] trained to predict emotion distributions directly from artworks.

Overall, the image-only model performs worse than the text-only baseline in terms of distributional alignment and ranking quality, as reflected by higher JS divergence and lower nDCG@3 scores. While Hit@2 scores remain relatively high for both ArtEmis v1 and v2, indicating that the model can often identify a plausible dominant emotion, its alignment with full emotion distributions is substantially weaker.

An exception is observed on ArtEmis v1, where the image-only model achieves slightly lower JS divergence than the text-only baseline. This suggests that, in settings with highly mixed or ambiguous emotional distributions, coarse emotional signals may sometimes be inferred directly from visual features alone. Nevertheless, this advantage does not translate into improved ranking performance, reinforcing the conclusion that visual information by itself is insufficient for robust emotional interpretation.

A notable difference between dataset versions is that ArtEmis v1 yields slightly better JS divergence for the image-only model, whereas ArtEmis v2 results in higher Hit@2 scores. This suggests that v1 may be easier for visual models to approximate at the distribution level, while v2 encourages stronger alignment with a small number of dominant emotions.

### 4.1.3 Summary of Baseline Findings

Across both baselines, text-only models outperform image-only models in emotional prediction, highlighting the importance of language in expressing emotional interpretation. At the same time, the contrasting behaviors observed between ArtEmis v1 and v2 reinforce the need to evaluate models on both dataset versions.

Together, these baseline results motivate the use of LMMs, which combine visual and textual reasoning, as a potential way to bridge the gap between visual perception and emotionally grounded art interpretation.

## 4.2 Multimodal Models: BLIP-2 and LLaVA

This subsection presents results obtained using LMMs that combine visual and textual reasoning. In contrast to the baseline experiments, which isolate single modalities. These models generate natural language interpretations conditioned jointly on visual input and structured prompts.

Model	Prompt	JS ↓	Hit@2 ↑	nDCG@3 ↑
BLIP-2	Art-historical	0.288	0.400	<b>0.699</b>
BLIP-2	Descriptive	0.268	0.600	0.538
BLIP-2	Emotional	0.234	<b>0.800</b>	0.665
BLIP-2	Simplified	<b>0.202</b>	0.200	0.625
LLaVA	Art-historical	<b>0.145</b>	<b>0.600</b>	0.827
LLaVA	Descriptive	0.161	<b>0.600</b>	<b>0.912</b>
LLaVA	Emotional	0.292	0.400	0.413
LLaVA	Simplified	0.223	0.400	0.699

Table 2: Prompt-based emotional alignment for BLIP-2 and LLaVA on ArtEmis v1. Lower JS indicates better alignment with human emotion distributions, while higher Hit@2 and nDCG@3 indicate better ranking performance. Best values are shown in bold within each model across prompt types.

Model	Prompt	JS ↓	Hit@2 ↑	nDCG@3 ↑
BLIP-2	Art-historical	0.609	0.333	0.108
BLIP-2	Descriptive	0.384	<b>0.250</b>	<b>0.453</b>
BLIP-2	Emotional	0.436	0.417	0.449
BLIP-2	Simplified	<b>0.276</b>	0.167	0.594
LLaVA	Art-historical	0.476	0.213	0.338
LLaVA	Descriptive	0.430	0.328	0.366
LLaVA	Emotional	0.444	0.328	0.396
LLaVA	Simplified	<b>0.382</b>	<b>0.426</b>	<b>0.400</b>

Table 3: Prompt-based emotional alignment for BLIP-2 and LLaVA on ArtEmis v2. Lower JS indicates better alignment with human emotion distributions, while higher Hit@2 and nDCG@3 indicate better ranking performance. Best values are shown in bold within each model across prompt types.

### 4.2.1 Quantitative Results

Tables 2 and 3 show that both BLIP-2 and LLaVA substantially outperform the image-only baseline across most prompt types, confirming that integrating language with visual input is critical for emotionally grounded interpretation. Compared to the text-only baseline, multimodal models achieve comparable or improved alignment on ArtEmis v2, while exhibiting more variable behavior on ArtEmis v1.

On ArtEmis v2, both models demonstrate stronger and more consistent emotional alignment across prompts, reflecting the clearer emotional structure of the contrastive dataset. For both BLIP-2 and LLaVA, simplified prompts yield the lowest JS divergence, indicating that reducing linguistic complexity can help stabilize emotion prediction when emotional signals are well separated.

In contrast, results on ArtEmis v1 show greater sensitivity to prompt formulation. Emotional and art-historical prompts occasionally improve ranking performance but also introduce higher variance, particularly for artworks associated with mixed or ambiguous emotional responses. This variability highlights the challenge of interpreting natural human emotion distributions.

Across both datasets, LLaVA tends to achieve higher nDCG@3 scores than BLIP-2 for descriptive prompts, suggesting stronger ranking of secondary emotions. However, this advantage does not consistently translate into lower JS divergence, indicating that more expressive or detailed language does not necessarily imply better overall emotional alignment. BLIP-2, while often producing shorter outputs, exhibits more stable behavior across prompt types.

Finally, similar to the text-only baseline, multimodal models frequently assign secondary emotions that correspond to human annotation distributions, particularly on ArtEmis v1. This further supports the use of distribution-based evaluation rather than relying only on dominant emotion labels when assessing emotional interpretation quality.

#### 4.2.2 Prompt Sensitivity Analysis

Prompt design has a strong influence on multimodal model behavior. Descriptive prompts tend to produce factual summaries with limited emotional depth, while emotional prompts tend to increase alignment with dominant ArtEmis emotion labels. Art-historical prompts often provide contextual detail but sometimes reduce explicit emotional focus, particularly in ArtEmis v1. Simplified prompts produce shorter outputs that are easier to interpret but occasionally oversimplify complex emotional content. The effect of prompt choice is more pronounced in ArtEmis v1, where ambiguous emotional distributions allow greater interpretive freedom. In ArtEmis v2, the influence of prompts is more constrained by the dataset’s sharper emotional labels.

In addition to prompt type, it is observed that the length and level of detail of the prompt have a noticeable impact on the style and content of model outputs. Short and underspecified prompts tend to show more generic and formulaic responses, often resembling standard captioning behavior. In contrast, longer and more detailed prompts encourage more elaborate, contextualized, and human-like interpretations, including affective language and metaphorical expressions.

This effect is particularly visible in LLaVA outputs, where extended emotional prompts frequently lead to richer atmospheric descriptions, while minimal prompts produce safer and more neutral responses. These findings suggest that prompt sensitivity in multimodal models is not only a function of semantic intent (e.g., emotional vs. descriptive), but also of prompt granularity and instruction richness. Concrete examples illustrating how different prompt formulations lead to systematically different interpretations for the same artwork are provided in Appendix A.

### 4.2.3 Qualitative Observations

Qualitative inspection of generated interpretations reveals several recurring patterns. Multimodal models often succeed in identifying the general mood or atmosphere of an artwork but struggle with subtle or conflicting emotional cues [7]. In some cases, emotionally expressive language is generated even when visual evidence is weak, suggesting that models may rely on learned associations rather than grounded visual reasoning. For example, models sometimes misidentify objects (e.g., calling a dog a horse), substituting visually present entities with statistically more frequent concepts learned during training.

Differences between the models are also evident in style. LLaVA tends to produce more narrative and expressive explanations, whereas BLIP-2 outputs are more restrained and closer to descriptive captioning. These stylistic differences interact with prompt type and dataset version, further emphasizing the importance of controlled evaluation.

## 4.3 Qualitative Case Study: *The Starry Night* (Van Gogh, 1889)

To complement the quantitative evaluation, there is also a qualitative case study on *The Starry Night* (Saint-Rémy, 1889) by Vincent van Gogh, one of the canonical examples used in the original ArtEmis paper. This artwork is characterized by expressive brushstrokes, swirling motion, and strong color contrasts, making it a challenging test case for emotional and context-aware interpretation. BLIP-2 is not included in this qualitative case study, as earlier results indicate that it produces shorter and more restrained outputs that are closer to descriptive captioning. While this behavior is suitable for factual interpretation, it limits the visibility of prompt induced variation in emotional expressiveness. LLaVA, by contrast, was found to generate longer and more narrative interpretations and to exhibit stronger sensitivity to prompt formulation, which makes it a more informative choice for illustrating qualitative differences in emotional framing and interpretive depth.



Figure 1: **Qualitative comparison on *The Starry Night* (1889)**. The dominant human emotion in ArtEmis for this painting is *awe*, with explanations emphasizing dreamlike atmosphere and emotional intensity. LLaVA generates an interpretation that captures the painting’s dynamic motion, luminous night sky, and affective tone, closely aligning with human emotional responses.

**Human (ArtEmis) explanation:**

*“The blue and white colors of this painting make me feel like I am looking at a dream.”*

**LLaVA (emotional prompt) output:**

*“The painting suggests a serene and peaceful atmosphere, with the blue sky and stars creating a sense of calmness. The presence of a church steeple and a house in the background adds to the tranquility of the scene.”*

**4.3.1 Analysis**

This example highlights a clear qualitative gap between human emotional explanations and current instruction-tuned multimodal model outputs. While LLaVA correctly identifies high-level visual elements and associates the scene with calmness and serenity, its interpretation remains generic and largely descriptive. In contrast, the human ArtEmis explanation emphasizes subjective experience and metaphor (“looking at a dream”), which conveys a stronger emotional grounding. Notably, even under an explicit emotional prompt, LLaVA tends to rely on neutral affective terms rather than expressive or metaphorical language. This suggests that, without task-specific fine-tuning or stronger affective supervision, multimodal models may struggle to reproduce the depth of emotional interpretation present in human-written ArtEmis explanations.

#### 4.4 Out-of-Distribution Qualitative Example: Astrophotography

To further explore the behavior of LMMs beyond the scope of curated art datasets, an additional qualitative example in the form of an astrophotography image is being used, captured by the author. The image depicts the Heart Nebula, an emission nebula characterized by abstract structure, low contrast, and the absence of recognizable everyday objects. Unlike artworks in ArtEmis, this image lies outside the distribution of training data typically used for art interpretation benchmarks and lacks associated human-written emotional annotations.

To assess whether this behavior generalizes beyond a single image, the same qualitative analysis was conducted on an additional set of 50 out of distribution astrophotography images, including nebulae, galaxies, and star forming regions. While these images are not used for quantitative evaluation, the resulting interpretations exhibit similar qualitative patterns, suggesting that the behavior observed for the Heart Nebula is representative rather than an isolated case.



Figure 2: **Out-of-distribution qualitative example: Heart Nebula (author photograph).** The image contains abstract luminous structures and diffuse color gradients without explicit objects, making it a challenging test case for emotionally grounded interpretation.

## LLaVA (emotional prompt) output:

*“The image suggests a sense of love and passion, as the heart shape is formed by the glowing nebula in the sky. The red hues add warmth and emotional depth, creating a feeling that is both inviting and contemplative.”*

### 4.4.1 Analysis

This example highlights a notable strength of instruction-tuned multimodal models when applied to abstract and space driven imagery. In contrast to earlier results on canonical artworks, LLaVA produces an emotionally expressive and coherent interpretation that aligns naturally with common human affective associations, despite the absence of cultural context, recognizable objects, or prior emotional labels.

The improved quality of the interpretation can be attributed to several factors. First, the visual signal is dominated by global structure and color rather than fine-grained object details, reducing ambiguity and minimizing the likelihood of object hallucination. Second, the abstract and non-representational nature of the image aligns well with LLaVA’s tendency to reason at the level of atmosphere and affect rather than concrete semantics. Finally, the absence of dataset-specific priors encourages reliance on visual cues rather than learned associations from curated art.

Since no human emotion annotations are available for this image, it is not used for quantitative emotion evaluation. Instead, it serves as an exploratory qualitative analysis of model behavior on out of distribution imagery. While limited in scope, this example suggests that LMMs may be particularly well-suited for emotionally framed interpretation of abstract visual domains. A systematic evaluation of emotional alignment for such imagery would require collecting human emotion annotations and is left for future work.

## 4.5 Summary of Multimodal Results

The multimodal experiments demonstrate that combining vision and language leads to clear improvements over image-only models and offers complementary strengths compared to text-only baselines. However, model performance remains sensitive to dataset structure and prompt design. While ArtEmis v2 enables clearer quantitative gains, ArtEmis v1 exposes limitations in handling nuanced and mixed emotional interpretations. These results highlight both the potential and the current limitations of LMMs for emotional art interpretation.

## 5 Discussion

The results of the baseline experiments reveal important differences between ArtEmis v1 and v2 and provide insight into how emotional information is distributed across textual and visual modalities. These findings have direct implications for the evaluation of LMMs in subsequent experiments.

A key observation is that text only emotion recognition performs substantially better on ArtEmis v2 than on v1. This aligns with the design of the contrastive dataset, where emotional labels are sharper and less ambiguous. In contrast, the lower performance on ArtEmis v1 reflects the presence of mixed emotional reactions, which more closely resemble natural human responses to art but complicate automatic evaluation.

The image only baseline further highlights these differences. While visual models are able to capture rough emotional signals, their predictions show weaker alignment with human emotional distributions than text based models. This suggests that emotional interpretation of art cannot be reliably inferred from visual features alone and benefits strongly from linguistic explanation.

Importantly, the contrasting behaviors of v1 and v2 underscore a fundamental trade-off between realism and evaluability. ArtEmis v1 offers a more naturalistic representation of emotional responses but introduces ambiguity and positive emotion bias. ArtEmis v2 provides clearer emotional separation at the cost of realism. Evaluating models on only one of these datasets would therefore give an incomplete picture of their interpretive capabilities.

These findings inform the interpretation of the multimodal experiments conducted in this thesis. Improvements observed on ArtEmis v2 may reflect better alignment with dominant emotions, while performance on v1 offers insight into how models handle nuanced, mixed or ambiguous emotional content. Prompt sensitivity and stylistic variation are also expected to interact differently with each dataset version.

Several limitations should be noted. Emotional interpretation is inherently subjective and automatic metrics such as BLEU or JS-divergence cannot fully capture interpretive quality. Additionally, dataset biases and annotation choices influence both model training and evaluation. These limitations motivate the inclusion of qualitative analysis alongside quantitative metrics.

Overall, the baseline results provide a necessary foundation for understanding the strengths and limitations of LMMs in emotional art interpretation and emphasize the importance of careful dataset selection and evaluation design.

## 6 Conclusion

The baseline experiments showed that textual explanations contain substantially more emotional information than visual features alone. Text-only models achieved stronger alignment with human emotional distributions than image-only models, while image-based predictions remained relatively coarse. These findings highlight the importance of language for expressing emotional interpretation, particularly when emotional responses are mixed or ambiguous. By explicitly accounting for secondary emotions rather than only dominant labels, this thesis provides a more nuanced evaluation of emotional understanding in multimodal models.

The multimodal experiments demonstrated that combining vision and language leads to clear improvements over image-only models and offers complementary strengths compared to text-only baselines. Both BLIP-2 and LLaVA were able to produce emotionally expressive interpretations, particularly when prompted explicitly to focus on emotional content. However, performance varied across datasets and prompt types.

A key result of this study is the contrasting behavior observed between ArtEmis v1 and ArtEmis v2. On the contrastive dataset (v2), multimodal models showed higher emotional alignment and more stable behavior, reflecting the clearer emotional structure of the data. In contrast, ArtEmis v1 revealed greater variability in model outputs, especially for artworks associated with mixed or ambiguous emotions. While this variability complicates quantitative evaluation, it also provides insight into how models handle more naturalistic emotional responses.

Differences between the multimodal models were primarily stylistic. LLaVA tended to generate longer, more narrative interpretations, often enriched with contextual details, whereas BLIP-2 produced more concise and restrained explanations. These stylistic differences did not consistently translate into superior emotional alignment, emphasizing that longer or more expressive text does not necessarily imply better emotional understanding.

In addition to differences across models and datasets, this thesis shows that prompt design itself plays a crucial role in shaping multimodal model behavior. Beyond prompt type, the length and level of detail of a prompt significantly influence the style and content of generated interpretations. Short and underspecified prompts tend to evoke generic and formulaic responses resembling standard image captioning, whereas longer and more detailed prompts encourage richer, more contextualized and more human-like interpretations that incorporate affective language and metaphor. This effect is particularly pronounced in LLaVA, where extended emotional prompts often yield more expressive atmospheric descriptions, while minimal prompts produce safer and more neutral outputs. These findings indicate that prompt sensitivity in multimodal models is not only a function of semantic intent, but also of prompt granularity and instruction richness.

Preliminary experiments on additional out-of-distribution images further suggest that these qualitative trends may generalize beyond curated art datasets. While these examples are not included in quantitative evaluation due to the absence of human emotion annotations, they indicate that LMMs may be particularly effective for emotionally framed interpretation of abstract or non-representational imagery. A larger-scale evaluation of such out-of-distribution data remains an

important direction for future work.

Overall, the findings suggest that current LMMs can approximate emotional interpretation of art to a limited extent, but their behavior remains sensitive to dataset design and prompt formulation. While ArtEmis v2 enables clearer quantitative gains, ArtEmis v1 exposes limitations in handling nuanced and mixed emotional content. Future work may explore task specific fine-tuning, interactive user feedback, and real-time deployment scenarios, bringing AI systems closer to functioning as reliable and engaging “gallery buddies” for art exploration. These results contribute to a clearer understanding of the gap between factual description and emotional interpretation in multimodal AI systems.

## References

- [1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapped language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Llava: Large language and vision assistant. *arXiv preprint arXiv:2304.08485*, 2023.
- [3] Panos Achlioptas, Maks Ovsjanikov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.
- [5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [6] Ziwei Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.
- [7] Yifan Zhang, Jianwei Yang, et al. Analyzing grounding and hallucination in large vision-language models. *arXiv preprint arXiv:2409.03521*, 2024.
- [8] Terry Barrett. *Interpreting Art: Reflecting, Wondering, and Responding*. McGraw-Hill, New York, 1991.
- [9] Kurt A. Bruder and Özüm Uçok-Sayrak. Interactive art interpretation: How viewers make sense of paintings in conversation. *Symbolic Interaction*, 23(4):337–358, 2000.
- [10] Sicheng Zhao, Guiguang Ding, and Jungong Han. Visual emotion analysis: A survey. *ACM Computing Surveys*, 54(5), 2021.
- [11] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025.
- [12] Shaohan Huang et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- [13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. On the reliability of multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- [14] Panos Achlioptas, Maks Ovsjanikov, Mohamed Elhoseiny, and Leonidas Guibas. Learning from contrast: Exploiting contrastive data for visual emotion recognition. *arXiv preprint arXiv:2204.07660*, 2022.

- [15] Kishore Papineni et al. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, 2002.
- [16] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004.
- [17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.
- [18] Yinhan Liu et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*, 2019.
- [20] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.
- [21] Christopher Burges, Tal Shaked, Erin Renshaw, et al. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.
- [22] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
- [23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

## A Prompt Sensitivity Examples

This appendix presents qualitative prompt sensitivity examples illustrating how different prompt formulations lead to systematically different model outputs for the same artwork. All examples are generated using LLaVA under controlled conditions and correspond to artworks from the ArtEmis dataset with known dominant emotion labels.

The qualitative examples focus on high-frequency ArtEmis emotions such as awe and sadness, as this category provides annotation density and visually grounded cues to isolate the effect of prompt formulation without unknown factors introduced by rare or semantically dependent emotions such as anger or disgust.



### Example A: Hugo Simberg, *Death Listens* (1897)

ArtEmis dominant emotion: Sadness

**Simplified prompt** *Describe the image.*

The image depicts a dark interior scene in which a skeletal figure is leaning forward, appearing to listen closely. The figure is positioned near a wall or surface, and the overall setting is dimly lit. The composition is simple, with limited background detail, emphasizing the central figure.

**Emotional prompt** *What emotion does this image evoke? Answer in 1–2 sentences.*

The image evokes a feeling of sadness and unease, as the presence of the skeletal figure creates a quiet and somber atmosphere.

**Extended emotional prompt** *Write a short paragraph about the emotional atmosphere and mood.*

The emotional atmosphere of the painting is subdued and melancholic. The skeletal figure, leaning in a posture of attentive stillness, suggests an intimate confrontation with mortality. The dark tones and sparse setting reinforce a sense of quiet inevitability, creating a reflective mood that invites contemplation rather than dramatic fear.



### **Example B: Paul Cézanne, *Sorrow* (1867)**

**ArtEmis dominant emotion: Sadness**

**Simplified prompt** *Describe the image.*

The image depicts a man lying on a bed, possibly sleeping or resting. He is wearing a white shirt and appears to be in a relaxed state. The bed is situated in a room with a chair nearby. The scene is painted in a darker tone, giving it a somewhat dramatic atmosphere.

**Emotional prompt** *What emotion does this image evoke? Answer in 1–2 sentences.*

The image evokes a sense of sadness and loneliness, as the man is seen sleeping alone in a dark room.

**Extended emotional prompt** *Write a short paragraph about the emotional atmosphere and mood.*

The emotional atmosphere and mood of the painting are quite somber and melancholic. The man is lying on a bed, possibly in a state of depression or exhaustion, with his head resting on his hand. The presence of a hanging light above the bed adds a sense of isolation and vulnerability to the scene. The overall mood of the painting is one of introspection and contemplation, as the viewer is left to ponder the emotional state of the man and the context of the image.



**Example C: Vasily Sadovnikov, *The Field Marshal's Hall of the Winter Palace* (1852)**

**ArtEmis dominant emotion: Awe**

**Simplified prompt** *Describe the image.*

The image is a drawing of a large, ornate room with a high ceiling. The room features a chandelier hanging from the ceiling, and there are several people walking around and interacting with each other. The people are scattered throughout the room, with some closer to the chandelier and others further away. The overall atmosphere of the room appears to be grand and elegant.

**Emotional prompt** *What emotion does this image evoke? Answer in 1-2 sentences.*

The image evokes a sense of grandeur and elegance, as it depicts a large, ornate room with chandeliers and a crowd of people dressed in fancy clothes.

**Extended emotional prompt** *Write a short paragraph about the emotional atmosphere and mood.*

The emotional atmosphere and mood of the scene can be described as grand and majestic. The large chandeliers hanging from the ceiling create a sense of grandeur, while the numerous people walking around the room contribute to the bustling and lively ambiance. The combination of these elements evokes a feeling of awe and admiration, as if one is witnessing a royal court or a grand event taking place in a historical building.