Advancing Multimodal Emotion Recognition in Comics via Multi-disciplinary, Multi-Task, Multi-Lingual Transfer Learning

Umer Mushtaq[1], Jean-Christophe Burie[1], and Antoine Doucet[1,2]

[1] L3I, SAIL Joint Laboratory, La Rochelle Université, La Rochelle, France
[2] Faculty of Computer and Information Science, Večna pot 113, University of Ljubljana, SI-1000 Ljubljana, Slovenia
{umer.mushtaq,jcburie}@univ-lr.fr, Antoine.Doucet@fri.uni-lj.si

Abstract. Comics constitue a rich multimodal narrative medium in which visual composition, stylized illustration, and textual dialogue interact to convey nuanced emotional states. Automatic emotion recognition in this domain poses significant challenges due to the interplay of artistic abstraction, panel-level context dependencies, and cross-cultural visual conventions, which together require models capable of learning semantically grounded, stylistically robust, and context-aware representations. In this work, we propose a sequential transfer learning (STL) pipeline that progressively enriches a pre-trained vision–language model through multi-disciplinary, multi-task, and multi-lingual transfer. Building upon a baseline adaptation on EmoComics35, the model is incrementally refined through parameter-efficient LoRA stages governed by a merge-and-freeze strategy designed to preserve accumulated knowledge at each step. We first leverage EmoArt dataset to capture shared affective visual primitives across artistic domains, then introduce EMID dataset to promote emotion-aware reasoning through diverse task formulations, and finally adapt to multilingual and culturally distinct inputs using MangaVQA dataset. To assess the statistical robustness of our findings, all results are averaged over three independent runs.We evaluate the final model and each intermediate stage on EmoComics35 using a diverse set of state-of-the-art vision–language models, including LLaMA-Vision, Qwen-VL, LLaVA and Pixtral. Results show that staged transfer across complementary dimensions yields consistent improvements, enhancing emotion understanding in comics and more broadly across multimodal narrative media.

Keywords: Emotion Recognition · Comics · Transfer Learning · Vision Models · Multimodal · Multi-disciplinary · Multi-task · Multi-lingual

1 Introduction

Comics represent a distinctive multimodal narrative medium that combine visual imagery and written language to convey complex meaning, evolving story arcs, and richly drawn characters. Enjoyed by millions of readers across diverse

cultures and continents, comics function not only as entertainment but also as a medium for storytelling, cultural expression, and social commentary. Their inherently multimodal nature enables the simultaneous presentation of imagery, text, spatial layout, and symbolic cues, yielding narrative experiences that are both richer and more layered than those achievable through a single modality alone. As a result, comics occupy a unique position in contemporary media ecosystems, reflecting cultural values, aesthetic traditions, and communicative practices that transcend geographic boundaries while remaining adaptable to local contexts.

Emotion recognition in comics is fundamentally multi-disciplinary, as it benefits from methodological and representational insights developed in adjacent visual and audiovisual domains such as film, television, and fine art. These media share foundational expressive tools such as, facial depiction, body language and color symbolism, to encode emotional information in ways that can enable computational emotion modeling. Training AI models on heterogeneous data drawn from these related disciplines can promote the acquisition of more generalizable emotional representations, enabling systems to better account for the stylistic variation and artistic abstraction that characterize comics. Such cross-domain transfer leverages common perceptual and narrative conventions while mitigating the limitations of domain-specific datasets, ultimately leading to more robust and context-aware emotion recognition frameworks.

Emotion recognition in comics can be naturally extended to multi-lingual settings, particularly through adaptation to Japanese manga where linguistic diversity and culturally specific narrative conventions introduce additional representational challenges. Cross-lingual exposure encourages the development of language-agnostic visual–semantic abstractions and reduces reliance on language-specific lexical patterns, thereby promoting more robust multimodal grounding. In parallel, emotion recognition can be augmented with multi-task learning component, where auxiliary objectives, such as visual question answering (VQA) in a closely related domain, facilitate the acquisition of richer semantic and contextual representations. VQA requires joint reasoning over visual content and linguistic queries, fostering fine-grained alignment between scene elements, character expressions, and narrative structure. These competencies substantially overlap with the demands of emotion recognition, which likewise depends on interpreting visual cues in relation to contextual meaning.

Sequential transfer learning constitutes the backbone of our framework, organizing domain, task, and linguistic adaptation as a staged training pipeline rather than a single fine-tuning step. We apply this approach across multiple vision–language models, including Qwen2.5-VL, Qwen3-VL, LLaMA-3.2-Vision, Pixtral, and LLaVA, to assess generalization across architectures. Starting from a baseline adaptation on the target dataset, the model is progressively refined through parameter-efficient LoRA stages, each of which is trained, merged into the backbone, and frozen before the next transfer stage begins.This design systematically extends pre-trained multimodal capabilities, such as visual grounding, cross-modal alignment, and contextual reasoning, through structured supervision at each stage. Intermediate datasets introduce complementary signals

spanning visual diversity, task variation, and multilingual content, enabling incremental refinement of representations related to facial expression, compositional cues, and narrative context. This staged consolidation mitigates domain shift and limits overfitting to individual datasets. Overall, sequential transfer learning provides a robust mechanism for cumulative representation learning, yielding robust and generalizable multimodal emotion understanding across stylistically and culturally diverse multimodal domains. Our contributions are as follows:

- We propose a sequential transfer learning (STL) framework for multimodal emotion recognition in comics. We evaluate the framework across five state-of-the-art vision–language models (LLaMA-Vision, Qwen-VL, LLaVA, Pixtral). The framework jointly models visual and textual signals and employs parameter-efficient LoRA with a merge-and-freeze strategy to enable staged knowledge transfer.
- *Multi-disciplinary transfer:* We introduce a cross-domain adaptation stage using EmoArt, encouraging the model to learn domain-invariant affective representations across diverse artistic styles, improving robustness to stylistic variation.
- *Multi-task transfer:* We incorporate EMID as an intermediate stage, introducing task diversity that promotes deeper semantic grounding and emotion-aware reasoning over visual–textual inputs.
- *Multi-lingual transfer:* We extend the pipeline with MangaVQA, enabling cross-lingual and cross-cultural alignment of emotion representations in comic-style narratives.

Collectively, this staged pipeline provides a unified mechanism for cumulative representation learning, yielding improved generalisation in multimodal emotion recognition. The implementation is available on GitHub: MANPU2026.

2 Related Works

Recent advances in multimodal emotion recognition for visual content have increasingly focused on combining textual and visual modalities to capture the nuanced affective narratives present in comic media. For instance, Kukreja et. al. [19] propose a multimodal emotion detection framework that integrates Vision Transformers (ViT) for visual feature extraction and a RoBERTa-based language model for dialogue sentiment analysis to jointly classify emotional states in comic panels. Their cross-attention–based fusion mechanism demonstrated substantial improvements over unimodal baselines on a custom dataset of over 10,000 annotated panels, highlighting the importance of balanced text–image integration for robust emotion inference in stylized contexts. In a complementary study, the same authors [18] introduce a comprehensive multimodal framework for comic emotion recognition that further incorporates background color cues alongside facial expression and textual sentiment analysis, employing fuzzy logic for adaptive fusion. Collectively, these works underscore the importance of exploiting

heterogeneous visual and textual cues, and directly motivate our decision to incorporate cross-domain artistic supervision through EmoArt, which extends color and compositional reasoning beyond the comics domain.

In the context of comics, multi-task learning has been explored to jointly model complementary prediction objectives. The EmoComicNet framework, proposed by Dutta et. al. [11], formulates comic emotion recognition as a multi-task problem in which separate image and text-specific classifiers are learned alongside a joint multimodal classifier. This prior work directly inspires our multi-task transfer stage using EMID, which provides VQA-based auxiliary supervision to enforce explicit visual–linguistic alignment, extending the multi-task paradigm to a sequential transfer setting.

For sequential visual understanding Wang et. al. [21] introduce StripCipher, a benchmark for evaluating reasoning over image sequences. The benchmark includes a human-annotated dataset and three subtasks: visual narrative comprehension, contextual frame prediction, and temporal narrative reordering. Evaluation of 16 state-of-the-art LMMs, including GPT-4o and Qwen2.5-VL, demonstrates a substantial performance gap compared to humans, particularly in temporal reordering, where GPT-4o achieves only 23.93% accuracy—over 56% below human performance. These findings are relevant to our work, as comic panels are inherently sequential; future extensions of our framework should incorporate temporal context modeling across panels.

Miyai et. al. [16] introduce Japanese Massive Multi-discipline Multimodal Understanding (JMMMU-Pro), an image-based Japanese multi-discipline multimodal understanding benchmark, together with a scalable construction methodology termed Vibe Benchmark Construction. Experimental results indicate that open-source Large Multimodal Models (LMMs) struggle significantly on JMMMU-Pro, highlighting persistent limitations in Japanese multimodal understanding.

3 Methods and Materials

We study multimodal emotion recognition in comics through a sequential transfer learning (STL) framework built on the central hypothesis that performance can be systematically improved by progressively enriching supervision along three complementary dimensions: multi-disciplinary, multi-task, and multi-lingual transfer. Rather than training independent models, we adopt a cumulative, stage-wise strategy, where each phase builds on the previously trained model using parameter-efficient LoRA with a merge-and-freeze scheme. This design enables controlled evaluation of knowledge transfer, retention, and incremental gains. The framework is depicted in Figure 1.

3.1 Datasets

In this subsection, we introduce the datasets we experiment with in our study to analyze our emotion recognition framework. The selected corpora are designed to support the four training configurations proposed in this study—multimodal,

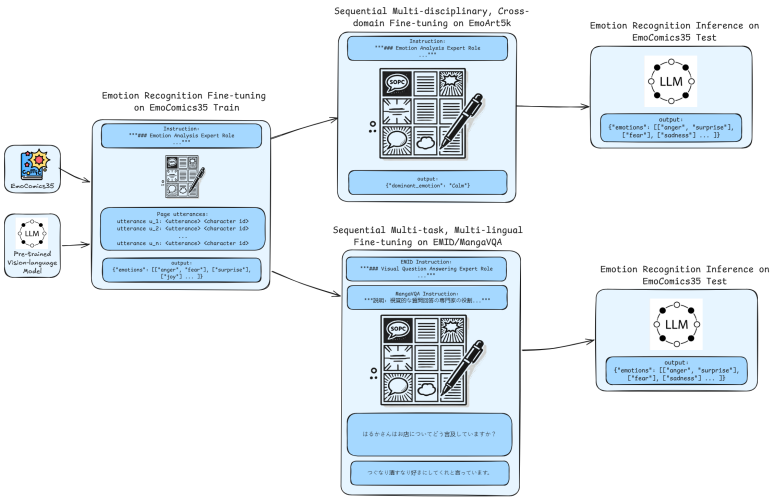

Fig. 1: Sequential transfer learning framework for multimodal emotion recognition in comics.

multi-disciplinary, multi-task, and multi-lingual learning—while maintaining coherence within a unified vision-language modeling paradigm.

EmoComic35 (target domain) We base our multi-stage study on EmoComics35, which is a newly constructed dataset designed to support research on emotion analysis in comic narratives. The dataset contains manually annotated dialogue segments extracted from 35 English-language comic books spanning multiple genres, thereby reflecting substantial diversity in narrative tone, character dynamics, and expressive styles. Emotional categories follow the discrete taxonomy proposed by Ekman, who classifies emotions as anger, fear, joy, sadness, surprise, disgust, and neutral [12]. Each dialogue instance is labeled with its corresponding emotion(s). By integrating rich textual content with visually grounded storytelling contexts, EmoComics35 enables systematic investigation of emotional expression and interaction patterns in multimodal comic media. EmoComics35 serves as the core benchmark for evaluating emotion classification performance across all experimental stages.

EmoArt (multi-disciplinary) To incorporate multi-disciplinary transfer learning in our framework, we utilize EmoArt, a recent dataset which contains artwork images annotated with emotion categories [23]. These artworks span diverse artistic styles and genres (such as Baroque, Surrealism, Abstract Art), emphasizing color, abstraction, composition, and symbolic representation. Each artwork is annotated with a dominant emotion label from twelve categories (such as Calm, Contented, Sad) as well as arousal and valence following the the Russell emotional model [20]. While EmoArt predominantly covers Western art traditions, its inclusion of abstract and symbolic art styles provides complementary visual

representations that differ significantly from comic panel aesthetics, motivating its use as a cross-domain bridge.

EMID and MangaVQA (multi-task, multi-lingual). For multi-task and multi-lingual supervision, we employ Emotionally paired Music and Image Dataset (EMID) and Manga Visual Question Answering (MangaVQA) respectively. In EMID dataset, each instance includes an audio sample, an image, a question, and an answer which represents the dominant emotion in the image. MangaVQA contains Japanese manga panels paired with question-answer annotations. EMID and MangaVQA datasets introduce multi-task and cross-lingual multimodal supervision, allowing the evaluation of language transfer effects and cultural variability in visual narrative interpretation. Collectively, these datasets provide complementary supervision signals that enable systematic investigation of sequential transfer learning across domains, tasks, and languages within multimodal emotion recognition. Statistical details on the datasets are presented in Subsection 4.1.

3.2 Multimodal Emotion Recognition

The baseline task is formulated as supervised multimodal emotion classification on EmoComics35 using vision–language models Each instance comprises a comic panel image, associated dialogue or narration, and a categorical emotion label from Ekman's taxonomy. The model jointly processes visual and textual inputs to predict the expressed emotional state within an instruction-tuning setup, where outputs are constrained to a predefined emotion taxonomy. This stage constitutes the initial adaptation of the pre-trained model to the target domain, implemented via LoRA fine-tuning.The resulting model serves as the baseline checkpoint for the sequential transfer learning (STL) pipeline. It captures in-domain multimodal representations without external supervision, providing a reference point for evaluating incremental gains from subsequent transfer stages.

3.3 Multi-disciplinary Transfer Learning

We extend the baseline model via cross-domain adaptation, motivated by the hypothesis that emotional representations share domain-invariant visual cues across artistic media, despite surface-level stylistic differences. Indeed, in EmoArt, emotion is conveyed through colour, texture, abstraction, and global composition rather than narrative structure. Starting from the baseline model trained on EmoComics35, we apply LoRA fine-tuning on EmoArt and subsequently merge the adapter into the backbone following the STL merge-and-freeze strategy. This intermediate adaptation encourages the model to capture domain-agnostic emotional features, reducing reliance on comic-specific artifacts and improving robustness to stylistic variation. The adapted model is then evaluated on EmoComics35 to quantify cross-domain transfer effects, isolating gains attributable to multi-disciplinary supervision.

3.4 Multi-task and Multi-lingual Adaptation with VQA

Visual Question Answering (VQA) is a multimodal task in which a model jointly processes visual and linguistic inputs to generate semantically grounded responses. This requires tight alignment between low-level visual perception, spatial reasoning, and contextual language understanding — competencies that directly overlap with those demanded by fine-grained emotion inference in comics, where affective meaning frequently emerges from implicit narrative cues and character interactions rather than from explicit visual signals alone.

Within our sequential transfer learning framework, we leverage VQA-based auxiliary supervision as a structured mechanism for extending the baseline model's representational capacity. Concretely, starting from the EmoComics35 checkpoint, we apply two successive fine-tuning stages using EMID and MangaVQA. EMID reinforces cross-modal emotional alignment by grounding visual content in affective semantic representations, while MangaVQA extends this alignment to Japanese-language multimodal reasoning over manga narratives, thereby fostering cross-lingual generalisation and culturally situated understanding.

Both stages are implemented via parameter-efficient LoRA, with each adapter merged into the backbone and frozen before the next stage begins, ensuring that knowledge accumulated at each step is preserved rather than overwritten. This multi-task and multi-lingual supervision jointly promotes structured visual–language reasoning, deepens semantic grounding, and improves cross-cultural robustness. The model resulting from these two stages is subsequently re-evaluated on EmoComics35 to quantify the contribution of VQA-based auxiliary learning to emotion recognition performance. The complete pipeline is formalised in Algorithm 1.

4 Results

We present a systematic evaluation of our sequential transfer learning framework for multimodal emotion recognition in comics. We analyse performance across the baseline, multi-disciplinary, multi-task, and multi-lingual stages to quantify incremental gains and assess model stability, generalisation, and transfer effectiveness. Cross-stage comparisons provide a rigorous assessment of each transfer stage's contribution to improved multimodal emotion understanding.

4.1 Experimental Setup

We now present our experimental setup in detail.

Data: **EmoComics35** comprises 35 comic albums, 874 comic pages, 3,180 comic panels and 7,129 annotated utterances. Of the 35 comic albums, 26 are set aside for the train and validation sets, whereas the remaining 9 are set aside for the test set. For utterance-level, the train and validation sets together consist of 5,803 utterances and the test set consists of 1,326 utterances. At page-level,

Algorithm 1 Sequential Transfer Learning for Emotion Recognition in Comics

Input: Pre-trained VLM W_{base}
Datasets: EmoComics35, EmoArt, EMID, MangaVQA
Output: Final model W_3 ▷ Baseline
Train $LoRA_E$ on W_{base} using EmoComics35
$W_{baseline} \leftarrow W_{base} + LoRA_E$
Evaluate $W_{baseline}$ on EmoComics35
▷ Stage 1 — Multi-discipline
Freeze $W_{baseline}$
Train $LoRA_{MD}$ on $W_{baseline}$ using EmoArt
$W_1 \leftarrow W_{baseline} + LoRA_{MD}$
Evaluate W_1 on EmoComics35
▷ Stage 2 — Multi-task
Freeze W_1
Train $LoRA_{MT}$ on W_1 using EMID
$W_2 \leftarrow W_1 + LoRA_{MT}$
Evaluate W_2 on EmoComics35 ▷ Stage 3 — Multi-lingual
Freeze W_2
Train $LoRA_{ML}$ on W_2 using MangaVQA
$W_3 \leftarrow W_2 + LoRA_{ML}$
Evaluate W_3 on EmoComics35

the train and validations sets consist of 718 pages and the test set consists of 156 pages. In EmoComics35, emotion labels as distributed as: *Anger* (2382), *Surprise* (1840), *Sadness* (1792), *Fear* (1729), *Joy* (1687), *Neutral* (438), and *Disgust* (327). **EmoArt** dataset consists of more than 132,000 diverse curated artworks in 56 distinct art styles. Each artwork is annotated with arousal and valence analysis, emotion recognition and attribute analysis. For our experiments, we use a subset of EmoArt, called EmoArt5k, which consists of 5532 annotated artworks [2,3]. In EmoArt5k, emotion labels as distributed as: *Calm* (2872), *Excited* (1240), *Contentment* (692), *Sad* (216), *Alarmed* (207), *Frustrated* (157), *Happy* (50), *Annoyed* (28), *Bored* (25), *Aroused* (24), *Tired* (19), and *Glad* (2). **EMID** (Emotionally paired Music and Image Dataset) consists of diversely sourced 10,738 multimodal music-image pairings categorized into distinct emotional labels [4]. For our framework, we utilize a randomly selected subset of 3,000 annotated samples from EMID dataset. In EMID, emotion labels as distributed as: *Sadness* (767), *Fear* (496), *Contentment* (486), *Anger* (434), *Excitement* (350), *Awe* (235), and *Amusement* (232). **MangaVQA** (Manga Visual Question Answering) consists of 526 manually created Japanese question-answer pairs based on images selected from Manga 109 dataset [5,7]. The question-answer pairs include both direct extraction and descriptive analysis in Japanese.

Models: We evaluate a range of state-of-the-art large multimodal models (LMMs) for comic emotion recognition. These include Qwen2.5-VL and Qwen3-VL, vision–language models with scalable parameter sizes and optimized multimodal attention mechanisms [8]; LLaMA-3.2-Vision, which extends LLaMA-3.2 with

a frozen visual encoder for joint image–text processing [1]; Pixtral, a model designed for robust visual reasoning and cross-modal alignment [6]; and LLaVA-NeXT and LLaVA-1.5, instruction-tuned vision–language models with large-scale image–text pretraining [15]. Together, these models cover diverse architectural designs, parameter scales, and training paradigms, enabling comprehensive benchmarking. For inference, all models are prompted to output structured JSON emotion labels. Predictions are post-processed to resolve formatting inconsistencies and map synonymous labels to a fixed taxonomy (anger, fear, joy, sadness, surprise, disgust, and neutral), ensuring compatibility with standard metrics such as per-class and weighted F1 scores.

Implementation details: Experiments were conducted on NVIDIA RTX A6000 (48GB) and NVIDIA H100 NVL (94GB) GPUs. Fine-tuning was implemented using LLaMA-Factory [24], with Unsloth for memory-efficient and accelerated training via optimized kernels and quantized loading [9]. All models were adapted using parameter-efficient fine-tuning (PEFT), specifically LoRA applied to transformer attention layers (and optionally MLP projections) [13,10]. For fairness, base architectures and optimization hyperparameters were kept constant across all experiments. Only training data composition and objectives varied across stages of the sequential transfer pipeline. Each result is reported as the average over three runs, and evaluation is performed on a fixed held-out EmoComics35 test split.

Hyperparameter	**Value**
Batch Size (per GPU)	2
Gradient Accumulation Steps	4
Learning Rate	[2e-4, 5e-5, 1e-6]
LR Scheduler	[Linear, Cosine]
Warmup Ratio	[0.1]
Optimizer	AdamW
Weight Decay	[0.01 – 0.05]
Temperature	[0.0 – 0.2]
LoRA Rank (r)	[16–32]
LoRA Alpha	[16–32]
LoRA Dropout	[0.0 – 0.2]
Number of Epochs	[1, 3, 5]

Table 1: Training hyperparameters

4.2 Baseline Results

Table 2 reports baseline emotion recognition performance on EmoComics35, where models are trained exclusively for multimodal emotion classification. This setting establishes the reference point for evaluating subsequent transfer stages. We compare LLaMA-Vision, Qwen-VL, LLaVA, and Pixtral under identical fine-tuning and optimization conditions. The results reflect each model's intrinsic

ability to integrate visual and textual cues for emotion recognition in comics, without external domain, task, or linguistic supervision. These baselines form the foundation for assessing gains from the proposed sequential transfer learning framework.

Model	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	F1
LLaMA-3.2-11B-Vision	0.55	0.38	0.37	0.46	0.23	0.41	0.49	0.60
LLaMA-3.2-90B-Vision	0.56	0.38	0.43	0.50	0.25	0.39	0.49	0.59
Qwen2.5-VL-32B	0.56	0.35	0.30	0.44	0.29	0.34	0.43	0.57
Qwen2.5-VL-72B	0.56	0.16	0.30	0.45	0.35	0.40	0.34	0.58
Qwen3-VL-8B	0.55	0.21	0.47	0.50	0.17	0.37	0.47	0.55
LLaVA-NeXT	0.53	0.21	0.42	0.39	0.08	0.39	0.43	0.57
Pixtral	0.57	0.19	0.43	0.48	0.43	0.43	0.48	0.59
LLaVA-1.5	0.54	0.24	0.30	0.43	0.10	0.41	0.36	0.55

Table 2: Baseline multimodal emotion recognition performance on EmoComics35 using different vision–language models under identical fine-tuning settings.

4.3 Sequential Transfer Learning Results

This section reports results for the multi-disciplinary, multi-task, and multilingual transfer stages. In each case, the model is evaluated on the EmoComics35 test set after completing the corresponding STL stage.

Table 3 reports emotion recognition performance after multi-disciplinary transfer using EmoComics35 and EmoArt5k. Following initial fine-tuning on EmoComics35, the model undergoes a second adaptation stage on EmoArt5k before being re-evaluated on the EmoComics35 test set. This sequential protocol deliberately exposes the model to stylistically heterogeneous artistic representations of emotion (spanning Baroque, Surrealism, and Abstract Art) after it has already specialised on the comic domain, with the aim of broadening its affective visual vocabulary without sacrificing target-domain performance. The central question addressed by this experiment is whether cross-domain visual supervision, drawn from fine art, can enrich the model's emotion representations in a way that transfers back to the stylized and narrative-driven comic setting.

Model	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	F1
LLaMA-3.2-11B-Vision	0.58	0.42	0.35	0.44	0.28	0.40	0.46	0.62
LLaMA-3.2-90B-Vision	0.53	0.37	0.47	0.52	0.24	0.36	0.47	0.60
Qwen2.5-VL-32B	0.51	0.39	0.38	0.48	0.28	0.39	0.41	0.61
Qwen2.5-VL-72B	0.55	0.19	0.34	0.49	0.31	0.43	0.38	0.63
Qwen3-VL-8B	0.51	0.26	0.41	0.58	0.19	0.35	0.49	0.64
LLaVA-NeXT	0.51	0.24	0.47	0.36	0.18	0.36	0.42	0.57
Pixtral	0.59	0.22	0.46	0.49	0.48	0.46	0.49	0.61
LLaVA-1.5	0.55	0.28	0.33	0.49	0.15	0.45	0.39	0.60

Table 3: Results of multi-disciplinary transfer learning using EmoComics35 and EmoArt5k, evaluated on the EmoComics35 test set.

Table 4 presents results for the multi-task transfer stage, in which emotion recognition on EmoComics35 is augmented with VQA-based supervision drawn from EMID. Following adaptation on EMID, the model is re-evaluated on the EmoComics35 test set to measure the impact of this auxiliary task on emotion

recognition performance. By requiring the model to answer structured questions grounded in visual content, VQA supervision enforces explicit visual–linguistic alignment and encourages the model to associate localised visual evidence, such as facial expressions, body posture, and scene context, with semantically precise linguistic outputs. This form of task-level supervision is hypothesised to yield richer and more transferable representations than emotion-only training, as it compels the model to engage in more fine-grained cross-modal reasoning rather than learning a direct mapping from panel appearance to emotion label.

Model	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	F1
LLaMA-3.2-11B-Vision	0.59	0.31	0.39	0.45	0.28	0.40	0.52	0.68
LLaMA-3.2-90B-Vision	0.59	0.39	0.44	0.52	0.26	0.40	0.51	0.66
Qwen2.5-VL-32B	0.57	0.38	0.38	0.47	0.32	0.39	0.44	0.65
Qwen2.5-VL-72B	0.59	0.19	0.35	0.47	0.38	0.41	0.38	0.67
Qwen3-VL-8B	0.55	0.24	0.49	0.55	0.22	0.39	0.48	0.69
LLaVA-NeXT	0.58	0.26	0.43	0.37	0.18	0.42	0.47	0.58
Pixtral	0.56	0.18	0.46	0.45	0.49	0.47	0.49	0.61
LLaVA-1.5	0.56	0.27	0.34	0.47	0.18	0.47	0.39	0.63

Table 4: Results of multi-task sequential transfer learning combining emotion recognition (EmoComics35) with VQA-based supervision (EMID), evaluated on EmoComics35.

Table 5 reports results for the multi-lingual adaptation stage, in which the model is exposed to Japanese multimodal supervision from MangaVQA following its prior training on English data from EmoComics35. After adaptation on MangaVQA, the model is re-evaluated on the EmoComics35 test set to assess whether cross-lingual exposure yields measurable gains on the English target domain. The hypothesis for this stage is that pairing manga-style visuals with Japanese-language queries encourages the model to ground emotional meaning in visual features rather than language-specific lexical patterns, thereby fostering more language-agnostic visual–semantic representations. To the extent that such representations are less tied to English surface forms, they should exhibit greater robustness to lexical variation and improved generalisation across linguistically and culturally diverse comic contexts. It should be noted, however, that given the limited scale of MangaVQA (526 question-answer pairs), the contribution of this stage is best interpreted as a cross-lingual regularisation signal rather than a comprehensive multilingual adaptation.

Model	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	F1
LLaMA-3.2-11B-Vision	0.51	0.39	0.37	0.44	0.29	0.46	0.50	0.67
LLaMA-3.2-90B-Vision	0.59	0.39	0.48	0.51	0.29	0.38	0.51	0.68
Qwen2.5-VL-32B	0.55	0.39	0.32	0.45	0.32	0.35	0.48	0.72
Qwen2.5-VL-72B	0.59	0.19	0.34	0.49	0.32	0.41	0.39	0.70
Qwen3-VL-8B	0.54	0.25	0.49	0.55	0.19	0.39	0.48	0.68
LLaVA-NeXT	0.59	0.24	0.48	0.42	0.15	0.43	0.47	0.69
Pixtral	0.59	0.22	0.48	0.50	0.49	0.41	0.50	0.66
LLaVA-1.5	0.51	0.29	0.38	0.45	0.18	0.48	0.39	0.68

Table 5: Results of multi-lingual sequential transfer learning using MangaVQA and EmoComics35, evaluated on the EmoComics35 test set.

5 Analysis

5.1 Overall Performance Trends

Overall, performance improvements are consistent across the enhanced training settings, indicating that the proposed transfer strategies yield incremental and systematic gains. Under the multimodal baseline (Table 2), LLaMA-3.2-11B-Vision achieves the strongest performance (weighted F1 = 0.60), closely followed by LLaMA-3.2-90B-Vision and Pixtral (weighted F1 = 0.59). With multi-disciplinary transfer (Table 3), peak performance increases to 0.64 weighted F1 score (Qwen3-VL-8B), suggesting that cross-domain exposure improves emotion generalization. The multi-task setting (Table 4) further improves results, with Qwen3-VL-8B reaching 0.69, a +0.09 gain over its baseline. Finally, multi-lingual adaptation (Table 5) yields the strongest overall performance, with Qwen2.5-VL-32B achieving a weighted F1 of 0.72.

The progression from multimodal baseline to multi-disciplinary, multi-task, and multi-lingual training reflects a structured process of incremental representational refinement. The baseline learns joint visual–textual embeddings for emotion classification but remains constrained to the target domain. Multi-disciplinary transfer expands the visual distribution, encouraging more style-invariant affect representations. Multi-task supervision via VQA introduces explicit cross-modal grounding, reshaping representations toward more compositional and relational encoding of emotion. Multi-lingual adaptation with Japanese manga further extends alignment to typologically distinct linguistic structures, including mixed scripts, ellipsis, and honorific forms, promoting language-agnostic semantic abstractions and robustness to lexical variation. However, unlike multi-task supervision, it primarily increases linguistic diversity rather than introducing new reasoning constraints, making its contribution complementary rather than structurally transformative.

5.2 Class-wise performance

Across all configurations, certain emotions remain consistently challenging. Disgust shows high variance and low F1 (down to 0.16–0.31 for some Qwen variants), indicating difficulty in capturing subtle or stylised negative expressions in comics. Neutral is also unstable, particularly for LLaVA-based models, with F1 sometimes below 0.20. We attribute this instability partly to class imbalance: Disgust (327 instances) and Neutral (438 instances) are severely underrepresented relative to Anger (2382). In contrast, Anger and Joy are consistently strongest (0.50–0.59), suggesting more transferable and visually salient cues. Fear and Surprise benefit notably from multi-disciplinary and multi-task training, indicating improved sensitivity to high-arousal and context-dependent emotions. Pixtral shows comparatively strong performance on Neutral (up to 0.49), suggesting better calibration for non-expressive states. Overall, multi-disciplinary training improves robustness across stylistic variation, particularly for Fear and Surprise, while multi-task learning yields the most consistent macro-F1 gains, especially

for Anger and Surprise through enhanced cross-modal reasoning. Multi-lingual training provides smaller but stable improvements, mainly in Sadness and Fear, acting as a regularising signal rather than a major representational shift.

5.3 SOTA Comparison

We further extended our analysis by comparing our results with state-of-the-art (SOTA) performance on three benchmark emotion recognition datasets: EmoryNLP, MELD, and DailyDialog [14,17,22]. We stress that this comparison is indicative and directional rather than definitive: EmoComics35, EmoryNLP, MELD, and DailyDialog differ substantially in domain, modality, annotation scheme, and intrinsic difficulty. A higher absolute F1 on EmoComics35 does not imply that our model is superior across all emotion recognition settings. The comparison is intended solely to situate our results within the broader literature. State-of-the-art results on EmoryNLP (0.42) and DailyDialog (0.55) are well below our EmoComics35 score (0.68), highlighting a clear advantage of our approach. Although MELD reports the highest score (0.70), our model remains closely competitive at 0.68. Overall, this shows that our framework matches or surpasses existing SOTA methods, despite tackling the more complex multimodal and stylized domain of comics.

Model	**Joyful**	**Mad**	**Neutral**	**Peaceful**	**Powerful**	**Sad**	**Scared**	**F1**
EmoryNLP SOTA	0.49	0.32	0.52	0.17	0.13	0.24	0.33	**0.42**
MELD SOTA	0.51	0.37	0.54	0.04	0.13	0.28	0.32	**0.70**
DailyDialog SOTA	0.54	0.44	0.55	0.10	0.11	0.31	0.33	**0.55**
EmoComics35 (our)	0.59	0.31	0.39	0.45	0.28	0.40	0.52	**0.68**

Table 6: SOTA Comparison

Taken together, these results demonstrate that performance gains in multimodal emotion recognition are driven primarily by the structure of the training pipeline rather than by model scale. The multimodal baseline establishes a functional cross-modal fusion capability, but remains constrained by domain-specific biases and the limited diversity of its training signal. Multi-disciplinary transfer partially alleviates these constraints by promoting stylistic invariance and broadening the model's emotional feature coverage across artistic domains. The most substantial gains, however, arise from multi-task supervision via VQA, which enforces explicit visual–linguistic alignment and encourages compositional reasoning over scene elements — competencies that transfer directly to the nuanced demands of comic emotion inference.

The sequential organisation of these stages is itself a key contributor to the observed improvements. By merging and freezing each adapter before proceeding to the next stage, the pipeline enables cumulative knowledge acquisition while limiting representational interference between successive training objectives. This allows features learned at each stage to be progressively refined rather than overwritten, preserving the complementary contributions of each transfer dimension.

Multi-lingual adaptation via MangaVQA further consolidates these gains by encouraging language-agnostic visual–semantic grounding, though its effect remains complementary rather than structurally transformative, consistent with the limited scale of the dataset and the regularisation-oriented nature of this stage. Overall, the findings support a hierarchical transfer paradigm in which domain, task, and language adaptations act synergistically: cross-domain exposure builds stylistic robustness, task-level supervision drives the primary representational shift, and cross-lingual regularisation provides a final layer of generalisation, with multi-task sequential transfer emerging as the central mechanism underlying robust multimodal emotion understanding.

6 Conclusions

This work presents a sequential transfer learning framework for multimodal emotion recognition in comics, organising domain, task, and linguistic adaptation as a cumulative pipeline rather than independent fine-tuning stages. Starting from EmoComics35, the framework progressively incorporates cross-domain supervision from EmoArt, VQA-based multi-task learning via EMID, and cross-lingual regularisation using MangaVQA, reflecting the hypothesis that robust comic emotion understanding requires exposure to heterogeneous visual styles, structured cross-modal reasoning, and linguistic diversity simultaneously. Empirical results consistently support this design. Multi-task supervision via VQA emerges as the primary performance driver, while cross-domain adaptation contributes complementary stylistic robustness and multi-lingual regularisation provides a final generalisation gain. Future work will focus on incorporating inter-panel sequential context modeling, expanding annotation for underrepresented emotion categories, exploring continual learning strategies and larger cross-cultural datasets to further improve emotion understanding in visual narratives.

References

1. Llama 3.2 vision. `meta-llama/Llama-3.2-11B-Vision-Instruct` (2024), accessed: 23 Feb 2026
2. Emoart5k. `https://huggingface.co/datasets/printblue/EmoArt-5k` (2025), accessed: 23 Feb 2026
3. Emotion-aware artistic generation. `https://zhiliangzhang.github.io/EmoArt-130k` (2025), accessed: 23 Feb 2026
4. Emotionally paired music and image dataset (emid). `https://huggingface.co/datasets/orrzohar/EMID-Emotion-Matching` (2025), accessed: 23 Feb 2026
5. Manga visual question answering (vqa). `https://huggingface.co/datasets/hal-utokyo/MangaVQA` (2025), accessed: 23 Feb 2026
6. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., et al.: Pixtral 12b (2024), `https://arxiv.org/abs/2410.07073`
7. Aizawa, K., Fujimoto, A., Otsubo, A., Ogawa, T., Matsui, Y., Tsubota, K., Ikuta, H.: Building a manga dataset "manga109" with annotations for multimedia applications. IEEE MultiMedia **27**(2), 8–18 (Apr 2020). `https://doi.org/10.1109/mmul.2020.2987895`

8. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
9. Daniel Han, M.H., team, U.: Unsloth (2023), https://github.com/unslothai/unsloth
10. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms (2023), https://arxiv.org/abs/2305.14314
11. Dutta, A., Biswas, S., Das, A.K.: Emocomicnet: A multi-task model for comic emotion recognition. Pattern Recognition **150**, 110261 (2024), https://www.sciencedirect.com/science/article/pii/S0031320324000128
12. Ekman, P.: Facial expressions of emotion: New findings, new questions. Psychological Science **3**, 34 – 38 (1992), https://api.semanticscholar.org/CorpusID:9274447
13. Hu, E.J., Shen, Y., Wallis, P., et al.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
14. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset (2017), https://arxiv.org/abs/1710.03957
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), https://arxiv.org/abs/2304.08485
16. Miyai, A., Onohara, S., Baek, J., Aizawa, K.: Jmmmu-pro: Image-based japanese multi-discipline multimodal understanding benchmark via vibe benchmark construction (2025), https://arxiv.org/abs/2512.14620
17. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: 57th Annual Meeting of the ACL. pp. 527–536. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://aclanthology.org/P19-1050/
18. Rishu, Kukreja, V.: Leveraging multimodal fusion for emotion detection in comics: integrating text and visual cues with roberta and vision transformers. International Journal on Document Analysis and Recognition (July 2025). https://doi.org/10.1007/s10032-025-00546-6
19. Rishu, Kukreja, V.: A novel multimodal framework for emotion recognition in comics: Integrating background color, facial expressions, and text sentiment using fuzzy logic. Expert Systems with Applications **289**, 128267 (2025)
20. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**, 1161–1178 (1980), https://api.semanticscholar.org/CorpusID:145278842
21. Wang, X., Xia, H., Song, J., Guan, L., et al.: Beyond single frames: Can lmms comprehend temporal and contextual narratives in image sequences? (2025), https://arxiv.org/abs/2502.13925
22. Zahiri, S.M., Choi, J.D.: Emotion detection on tv show transcripts with sequence-based convolutional neural networks (2017), https://arxiv.org/abs/1708.04299
23. Zhang, C., Xie, H., Wen, B., Zuo, S., Zhang, R., Cheng, W.H.: Emoart: A multidimensional dataset for emotion-aware artistic generation. In: 33rd ACM International Conference on Multimedia. p. 12644–12650. Association for Computing Machinery, New York, NY, USA (2025), https://doi.org/10.1145/3746027.3758201
24. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: LLaMA-Factory: Unified efficient fine-tuning of 100+ language models. In: 62nd Annual Meeting of the ACL (Vol 3). Thailand (2024), http://arxiv.org/abs/2403.13372