Cap4Art: Improving Image Captioning Capabilities Through Multi-Task Learning

Willy Chan
Stanford University
willyc@stanford.edu

Sunny Yu
Stanford University
syu03@stanford.edu

Xiaofei Yan
Stanford University
fayeyan@stanford.edu

Abstract

Visual art communicates more than physical content — it evokes emotion, atmosphere, and cultural context. Yet existing image captioning models often reduce artworks to mere object lists, missing their expressive and stylistic depth. This limitation especially impacts blind and visually impaired users who rely on captions for interpretive access. Our project addresses this gap and improves the artistic expressivity of image captioning models by fine-tuning a BLIP-based captioning model based on multiple classification tasks (e.g. emotion, style, school, etc.), enabling it to generate descriptions that capture both what is shown in a painting and also how it "feels". First, we use vision transformer models to fine-tune classifiers to predict for subjective labels like emotion, style, and school, then add the loss to the encoder for the captioning model to learn such latent features. Results show that our training process improves the expressivity and captioning quality of the model; compared with both the frozen BLIP baseline and a caption-only fine-tune, our multi-head model produces markedly richer prose while retaining factual grounding. The trained variants improve BLEU by +8 points over caption-only fine-tuning and boosts METEOR to 11.1 while nearly doubling adjective density and Flesch–Kincaid grade: resulting in multiple models that vary in expressiveness and style. [1]

1. Introduction

Visual art conveys not only objects and scenes but also stylistic nuances and emotional resonance. However, state-of-the-art image captioning models, trained primarily on photographic datasets, tend to generate factual but expressionless descriptions when applied to artworks, limiting accessibility for blind and visually impaired users. To address this, we introduce Cap4Art, a multi-task learning framework that fine-tunes a BLIP-based vision-language model on art-specific data while jointly learning emotion classification and other tasks. Our model explicitly incorporates emotional classification into the captioning process through a multi-task learning approach. We fine-tune a BLIP-based vision–language model to jointly generate captions and classify artworks into one of nine affective categories, encouraging the image encoder to internalize emotionally salient features. Our model leverages ArtEmis emotion labels [1], WikiArt Emotion [10] annotations, and image-level information from SemArt [5] to train classification models to generate emotion, style, and school labels for unseen data. Then we leverage this information to encourage the encoder to capture affective and stylistic features. During training, we combine the standard captioning loss with auxiliary emotion classification loss and the classification loss for all the classification heads, guiding the network to internalize emotionally salient cues and other stylistic information.

Evaluated on ArtPedia [14], Cap4Art outperforms the original BLIP [7] baseline in most metrics that measure the linguistic richness and semantic diversity of the language. We also ablate on the weights for each image feature (emotion, sty;e, type, school) and a balanced one with equal weights to all features, and discuss the trade-off between factual accuracy and linguistic expressivity. We find that overall, the model trained based on equal weights for each feature has the best performance. In particular, its METEOR score increases to 11.1, indicating the generation of a richer and more complex language. Qualitative analysis further shows that our generated captions exhibit greater artistic expressiveness. These results demonstrate that multi-task training can bridge the gap between factual accuracy and emotional depth, improving the accessibility of fine art for non-sighted audiences.

2. Related Work

Recent work in artwork captioning emphasizes the need to move beyond object recognition to capture emotional and stylistic dimensions of visual art. [17] leverage CLIP-based multimodal models to generate affect-rich captions,

[1] The code can be found at https://github.com/sunnyych/Cap4Art

demonstrating the value of contrastive pretraining for aligning vision and language in artistic domains. [15] introduce ZeroCap, a zero-shot captioning framework that avoids paired training data by optimizing latent representations, showcasing the potential of flexible generative pipelines. Additionally, [6] propose a knowledge graph-enhanced captioning model that incorporates cultural and contextual knowledge, improving the richness of generated text. Our work builds on these foundations by integrating emotion classification and art-specific metadata into a BLIP-based captioning model to generate more expressive and accessible descriptions of visual art.

Other state-of-the-art methods include [12, 2], which present a tripartite approach for artistic outpainting that harmonizes image-to-text and text-to-image generation processes. Other methods, such as [3, 17], leverage CLIP-based models to further explore the complexities of transforming images into corresponding textual descriptions through an innovative ensemble framework based on contrast language image pretraining (CLIP). Additionally, data-efficient methods such as [9, 8] propose a novel training process aimed at achieving higher data efficiency in captioning fine art, utilizing a virtual-real semantic alignment training process to enhance the feature extraction process and support effective learning.

Given the inherently subjective nature of art, many work have sought to provide annotations for the subjective features. For example, [11, 4] offers a distinct approach by collecting a comprehensive dataset of artistic styles described through natural language captions. Moreover, [13] emphasizes the evolution of methodologies and technologies that enhance the generation of descriptions for visual art, focusing not only on technical accuracy but also on the emotional and narrative dimensions that contribute to a richer user experience.

3. Data

We leverage several existing datasets with art-image-text pairs. [1] is a dataset with image-emotion pairs (where crowdsource workers annotated each image as one of the following emotions: amusement, anger, awe, contentment, disgust, excitment, fear, sadness, or something else) [2] . Other datasets used for training the classifiers: [5] is a multi-modal dataset for semantic art understanding that includes fine-art painting images with text attribute labels including school, type, time frame, technique, and so on. In our experimentation, we use the school and type labels to train classification models. Another dataset that we use is [10], a dataset of 4,105 pieces of art (mostly paintings) with human annotations for the emotions evoked. Specifically, there are twenty emotion categories, which provide fine-grained emotion categorizations. We use the labels to train classifiers for emotions in art. Full details of the training data used for each classifier is included in Tab 3.

Dataset Source	Size	Features
[1]	455K	Emotion, Style
[5]	21394	School, Type, Time Frame, Technique
[10]	4105	Emotions

Table 1. The size and features information for the training data used.

We use [14] for evaluation. The data set contains a collection of 2,930 painting images, each associated with a variable number of textual descriptions. We intend to use the human-annotated captions as ground truth reference [16] to evaluate our Cap4Art model.

4. Methods

4.1. Emotion Classification and Ablations

Model	**Training Accuracy**	**Test Accuracy**
`clip`	25.87%	25.60%
`vit`	49.48%	30.37%
`convnext`	49.58%	29.19%

Table 2. Training and test accuracy for three different models, respectively openai/clip-vit-base-patch32, google/vit-base-patch16-224-in21k, and convnext_base.fb_in22k (from top to bottom) on the emotion classification task using training data from [1] using the full dataset.

We first perform classification of the emotion labels using various models to examine the effects of training data quality, size, and models using data from [1].

Training Details. To provide emotion classification ground truths, we explore fine-tuning with three pretrained vision architectures on the ARTEMIS dataset for emotion classification: ViT [3], ConvNeXt [4], DINO [5], and CLIP [6]. Each model was trained to predict one of nine emotional

[2] The dataset only provides image names from WikiArt `https://www.wikiart.org/`. To access the images, we extracted the image urls from the Kaggle dataset `https://www.kaggle.com/datasets/antoinegruson/-wikiart-all-images-120k-link/data` by matching with the image file names in [1]

[3] `google/vit-base-patch16-224-in21k` architecture via the HuggingFace `ViTForImageClassification` class and employed the associated image processor for input normalization

[4] We adopted the `convnext_base.fb_in22k` architecture from the Timm library

[5] Adopted `vit_base_patch16_224_dino` from the Timm library

[6] leveraged the vision encoder from `openai/clip-vit-base-patch32`, augmenting it with a linear classification head while using the HuggingFace `CLIPProcessor` for image preprocessing

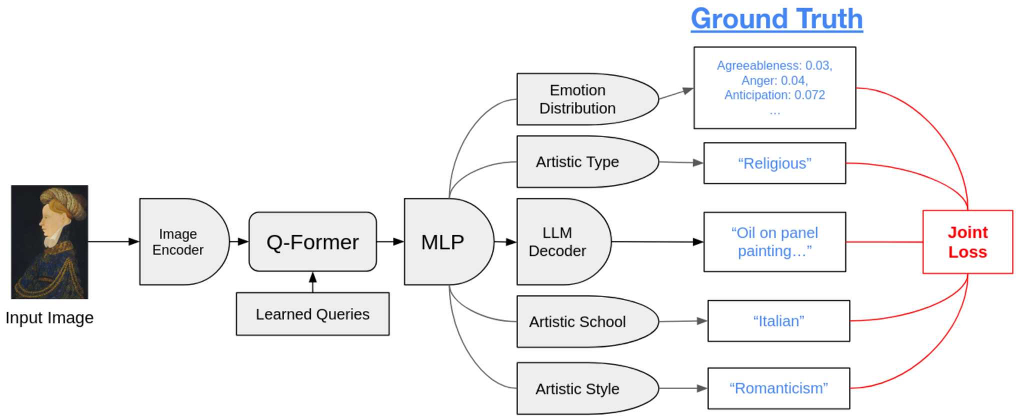


Figure 1. Computation graph of the modified BLIP-based captioning model. In addition to generating captions from visual features, the model incorporates a "mixture" of different artistic classification heads to help the image encoder capture the emotional essence of a painting. This dual-objective setup leverages BLIP's strong visual pretraining while fine-tuning on Artpedia's small but high-quality caption dataset. Losses from both tasks are combined and backpropagated through the network.

categories. For all experiments, we used a batch size of 32 and trained for 30 epochs using the AdamW optimizer with a learning rate of 3×10^{-4} and weight decay of 0.01. A cosine annealing learning rate scheduler was applied over the training duration. The dataset was split into training (72%), validation (18%), and test (10%) subsets with stratification over emotion labels. The results are reported in Table 2.

Ablations We hypothesize that the initial low test accuracy could be a result of the noisy data [7]. Therefore, we filtered the Artemis dataset by annotator numbers and only included images with more than 40 human annotations (where the classification is the majority vote), resulting in a final training set in Table 5. We fine-tuned vit and convnext using the smalldataset and report the results in Table 6. We show the full training and validation curves in Appendix A

4.2. Style Classification

The same dataset provides labels for the style of the painting, and we use the same base model to train a classifier for style, resulting in a test accuracy of 100%, showing that style classification is an easy task. We include the style labels as metadata for the CLIP model.

4.3. Type and School Classification

Besides emotions, the style and school of a painting are also important features that could provide helpful information about the subjective visual experience of the artwork. To predict for these features, we use ground-truth labels from [5] to train two classifiers, one for predicting the type, and one for predicting the school. There are ten categories for type, such as religious, portrait, landscape and mythological; and there are 26 categories for school, including Italian, Dutch, French, Flemish, German, Spanish, etc.

Training Details The training dataset size is 19244, the validation size is 1069, and the test size is 1069. Based on ablation studies in Sec 4.1, we find that the vision transformer and convnext both perform reasonably well on the classification task. Therefore, we chose to perform classification using convnext here. We explored training the model from scratch but the performance was suboptimal, so we instead fine-tune the pretrained convnext-base model on the SemArt dataset to classify paintings into one of 26 European art school categories (for school) or one of the ten painting types (for type). The training process uses cross-entropy loss with label smoothing and the AdamW optimizer, with a cosine annealing learning rate scheduler over 10 epochs. Images are resized to 224×224 and augmented during training with random horizontal flips, rotation, and color jitter. The model is partially frozen, with only the classification head being trainable. We used a batch size of 32.

Results The test accuracy for predicting the type for the painting is 37.89%, and the test accuracy for the school of the painting is 39.29%, and the training and test accuracy for the last epoch is reported in Tab 3.

4.4. Fine-Grained Emotion Classification

Finally, to provide even more fine-grained information about the subjective features of the paintings, we use emo-

[7] emotion tagging is inherently subjective and there could be disagreements among humans

Task	Train Accuracy	Validation
Type Classification	38.87%	39.85%
School Classification	41.63%	40.13%

Table 3. Training and validation accuracy for type and school classification tasks on the last epoch.

tion labels from [10] to predict the emotions evoked in a painting. The dataset provides annotation results from many annotators for the same image, including a distribution of each of the possible emotions (the proportion). With this information, we trained a classifier to model the distribution of the classes to match with the human annotation distributions. The other classifier we trained was based on the majority label.

Training Details We fine-tuned two variants of the ViT-Base model (vit-base-patch16-224-in21k) on the WikiArt dataset for emotion recognition. The first model framed the task as single-label classification using the majority emotion label per painting and was trained with a cross-entropy loss function. The second model treated emotion as a distribution prediction task, optimizing Kullback-Leibler (KL) divergence between predicted and target soft labels derived from annotator distributions. In both cases, only the classification head of the model was trainable, while the vision backbone was frozen. Models were trained using the AdamW optimizer (learning rate = 3e-4, weight decay = 0.01) and a cosine annealing learning rate scheduler. Training was performed for 10 epochs with early stopping based on validation accuracy.

Results The test accuracy for the classifier that predicts the majority label is 26.34%, and the test accuracy for the classifier that is trained based on KL divergence between the annotator distribution and the logits is 20.43% [8].

4.5. DINO

We experimented with the ArtEmis dataset (43k+ images) to learn self-supervised “art features” via DINO on two backbones: ViT-Base/16[3] and ConvNeXt-Base[4]. In each run, we generated six augmented views per image (two global 224×224 crops plus four local 96×96→224×224 crops) using random resized cropping, color jitter, horizontal flips, and Gaussian blur to encourage both global- and patch-level consistency. The DINO student–teacher pair consisted of a ConvNeXt (or ViT) trunk followed by a two-layer projection head (1024→2048→256 dimensions), where the teacher was updated by a 0.996 momentum from the student at each step. We warmed up the teacher temperature from 0.04 to 0.07 over 10 epochs while using a cosine-annealed learning rate (1e-5→3e-4→1e-5) and mixed-precision training. A running “center” of teacher outputs prevented collapse by normalizing logits before softmax. After only 15 epochs—roughly half of the 30 epochs commonly used in prior work, due to resource constraints—the backbones may not have fully converged; we then froze each backbone, appended a fresh two-layer MLP, and trained only that MLP on ArtEmis emotion labels. Under these conditions, ConvNeXt-Base (DINO) achieved 26.18 % validation accuracy (versus 21.75 % for ViT-Base/16)[9], but both results should be viewed as conservative estimates. Consequently, we chose not to use this DINO pretraining further for downstream models.

4.6. Captioning Model

For the caption generation task, we adopt the **BLIP** model introduced by Li et al [7], a vision-language architecture composed of a Vision Transformer (ViT) encoder and a Transformer-based language decoder. The ViT encodes an image into a sequence of visual tokens, including a dedicated [CLS] token that summarizes the global visual context. The decoder autoregressively generates text, cross-attending to these visual tokens to produce natural language captions. More details are included in Appendix B

Baseline As our primary baseline, we use the original BLIP model for the purposes of captioning the artwork (`Salesforce/blip-image-captioning-base`) without any fine-tuning. We also fine-tune BLIP with only ArtPedia data and no auxiliary heads to further isolate the impact of our extra feature classification information. More details are in Appendix B.

Multi-task Learning: To further encourage stylistic and affective richness, we augment the BLIP model with four auxiliary feature classification heads, each predicting a discrete attribute of the artwork in addition to the caption. Inspired by common use of the ViT [CLS] token for classification tasks, we project the encoder’s final [CLS] embedding through multiple MLP layers to predict the emotion category:

- **Style**: Romanticism, Early Renaissance, Northern Renaissance, Impressionism, Post Impressionism, Symbolism, etc.
- **Art Type**: religious, portrait, mythological, historical, landscape, interior, genre, study, still-life
- **Art School**: Italian, Dutch, Portuguese, Swedish, Flemish, French, Spanish, Belgian, etc.

[8] This is calculated as taking the class with the maximum logits as the predicted answer and checking if it matches with the ground truth

[9] We also attempted to run DINO on a CLIP-ViT/32[6] backbone, but its self-supervised loss plateaued within the first few epochs, so we did not pursue it for downstream emotion classification.

- **Emotion Distribution**: $\mathbf{e} \in [0,1]^{K_{\text{emo}}}$ (twenty-way soft target from ARTEMIS)

Following the standard practice for ViT classification, we feed the encoder's global [CLS] embedding $\mathbf{v}_{[\text{CLS}]}$ into four independent linear heads:

$$\hat{\mathbf{s}} = \mathbf{W}_{\text{style}}\mathbf{v}_{[\text{CLS}]} + \mathbf{b}_{\text{style}}, \qquad \hat{\mathbf{t}} = \mathbf{W}_{\text{type}}\mathbf{v}_{[\text{CLS}]} + \mathbf{b}_{\text{type}},$$
$$\hat{\mathbf{c}} = \mathbf{W}_{\text{school}}\mathbf{v}_{[\text{CLS}]} + \mathbf{b}_{\text{school}}, \qquad \hat{\mathbf{e}} = \mathbf{W}_{\text{emo}}\mathbf{v}_{[\text{CLS}]} + \mathbf{b}_{\text{emo}},$$

where the first three heads are softmax classifiers and the last is a K_{emo}-dimensional sigmoid output. The total objective is the weighted sum of the captioning loss and four auxiliary losses:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{cap}} + \alpha_s\,\mathcal{L}_{\text{CE}}(s,\hat{\mathbf{s}}) + \alpha_t\,\mathcal{L}_{\text{CE}}(t,\hat{\mathbf{t}}) \\ +\alpha_c\,\mathcal{L}_{\text{CE}}(c,\hat{\mathbf{c}}) + \alpha_e\,\mathcal{L}_{\text{BCE}}(\mathbf{e},\hat{\mathbf{e}})\end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{cap}}$ is the standard cross-entropy over decoder tokens and we set $\alpha_s = \alpha_t = \alpha_c = \alpha_e = 0.1$ as an example weighting. The auxiliary heads encourage the encoder to encode stylistic features with the aim of encouraging the language model decoder to be more expressive in captions. A picture of the computation graph is shown in Fig. 1.

Training details We fine-tune the BLIP base checkpoint for five epochs on a subset of 1,004 image–caption pairs from the ARTPEDIA dataset. We use the AdamW optimizer ($\eta = 3 \times 10^{-5}$) with mixed-precision training, and run all experiments on a single NVIDIA T4 GPU with 16GB of memory.

5. Experiments

To isolate the impact of each of our auxiliary heads, we fine-tuned BLIP under four configurations (note that all models share identical hyperparameters outlined in 4.6):

1. **Caption-only.** BLIP is fine-tuned on ARTPEDIA captions with no auxiliary heads. This serves as our single-task baseline.

2. **Multi-head (equal).** All four heads—*style, art-type, art-school, emotion-dist*—are enabled with a uniform weight of $\alpha = 0.1$ each (Eq. 1).

3. **Head-dominant (style).** We bias training toward *style* by setting $\alpha_{\text{style}} = 0.9$ and the other weights to 0.1.

4. **Head-dominant (type / school / emotion).** Analogous runs where *art-type*, *art-school*, or *emotion* receives the 0.9 weight while the remaining heads keep 0.1. These three additional runs let us probe how prioritizing a single attribute reshapes caption content.

Quantitatively, we report BLEU, ROUGE-L, METEOR, and BERTSCORE on a 10 % held-out test split (see App. C for metric details). For stylistic analysis we compute adjective density, TTR and FK-grade.

Quantitative Results The results can be seen in Table 4. As expected, the frozen BLIP baseline scores highest on surface-overlap metrics (BLEU 28.7, chrF 38.3), but the captions are more descriptive rather than evocative (e.g. only 2.6% adjective rate, FK 3.3, etc.). When finetuning on captions alone, BLEU and chrF fall sharply while other stylistic metrics show modest gains or small change (e.g. METEOR, adjective usage, and BERTScore nudge upward).

The pre-trained BLIP model was optimised on 129 M web pairs, so it naturally achieves the highest n-gram overlap (BLEU/chrF). Fine-tuning solely on ARTPEDIA shrinks that web vocabulary to a few thousand art captions; as a result surface metrics plunge, yet METEOR and BERTScore rise slightly, indicating that the model is learning task-specific synonyms while losing verbatim overlap. The jump in adjective density confirms that even a small amount of domain data encourages richer modifiers.

Adding a single *emotion–distribution* head increases stylistic richness: METEOR climbs to 10.6, adjective density reaches its maximum, FK grade more than doubles, and the fall in TTR points to greater lexical variety. The cost is lower BLEU/chrF, confirming that **heightened affective language reduces n-gram overlap with human references**.

Overweighting any one attribute head noticeably steers caption style (see Sec. 5). For instance, *style-dominant* captions favor niche terms; *type-dominant* runs preserve surface fidelity (highest BLEU among single-heads), while *emotion-dominant* captions are the most evocative.

Overall, the **equally-weighted configuration** ($\alpha = 0.1$ for all four heads) achieves the best compromise: BLEU rebounds to 24.6, chrF and ROUGE attain or approach their peaks, and all stylistic metrics remain well above the caption-only baseline: showing that a gentle multi-task signal enriches prose without sacrificing factual grounding.

Key Takeaways

- **Finetuning with captions alone is brittle:** BLEU/chrF collapse when auxiliary control signals are absent which indicates overfitting to the small artistic corpus.
- **Balanced multi-task architecture wins:** equal α (0.1) per head yields the best trade-off between surface accuracy and stylistic depth.
- **Head dominance is a stylistic dial:** boosting a single head meaningfully amplifies its trait but can degrade other orthogonal quality metrics.

Qualitative Results To investigate how each auxiliary signal shapes generation, we ran four "dominant–head" models in which a single head receives $\alpha = 0.9$ while the remaining heads keep $\alpha = 0.1$. Figure 2 highlights how each auxiliary head shapes the narrative tone of the captions.

For the portrait in the first row, the baseline offers a terse description ("a man in a suit and hat"), whereas the equally-weighted configuration adds concrete, verifiable details (beard, black hat, paintbrush) without straying from the image. When the emotion head is dominant, the prose becomes even more vivid, inventing a "white suit and white coat" that does not exist, revealing the head's tendency to sacrifice factual precision for affective flourish (also reflected in the lower BLEU score). The type dominant caption, by contrast, only appears to "see" material features in the foreground (e.g. "a man with a beard and mustache", but dropping the paintbrush, illustrating how the type head looks more at the main subject rather than specific details).

The cubist female portrait in the second row shows a different pattern with respect to figure pose: the baseline casts it as a woman "sitting in front of a mirror," while the equally-weighted model corrects the pose, calling it simply a "woman's face in the foreground." The emotion-dominant variant appears to embellish posture ("hands on her knees, feet on the ground"), again trading true exact accuracy for vividness and exact minute features.

Finally, in the rainy street scene (third row) the equally-weighted model disambiguates plurality ("three people walking in the rain"), whereas the emotion-dominant caption spins a short narrative about "a couple walking through the streets of Paris," importing setting and a specific **emotional** relationship that are not guaranteed by the pixels alone. The school-dominant run shifts focus to curatorial language, noting the "right half-length of the painting" and a "couple in the foreground," mirroring catalog descriptions more than everyday speech.

Across all three examples, a consistent theme is that balanced heads enrich captions, while overweighting a single head pushes the text toward specific behaviors reflective of the emphasized feature, often at the expense of other forms of fidelity.

Connection Between Quantitative and Qualitative Results The quantitative trends seen in Table 4 appear to be consistent with the linguistic "pressure" that each training configuration seems to observe, as discussed in 5.

For instance, when we overweight the *style* head, which maps to coarse historical time periods, the model inserts specialist terms that are rarely present verbatim in the references, lowering BLEU/chrF, yet the added jargon lengthens sentences and lifts FK grade.

Emphasizing the *art-type* head (categories like *religious*, *landscape*, or *still-life*) has the opposite effect: those medium descriptors appear frequently in museum captions, so n-gram metrics stay high, but METEOR only inches upward because these nouns add little syntactic variety.

A strong *art-school* signal (e.g., *Italian*, *Flemish*, *French*) appears to inject more long proper-name modifiers that inflate ROUGE (token overlap) while hurting chrF (character mismatch), explaining the divergent scores in the school-dominant run.

Finally, emphasizing the twenty-way *emotion distribution* head forces the decoder to output affective adjectives drawn from ArtEmis, which seems to drive METEOR and FK sharply upward, but it also introduces novel color/mood-related vocabulary that diverges from ground-truth phrasing, hence the drop in BLEU and BERTScore.

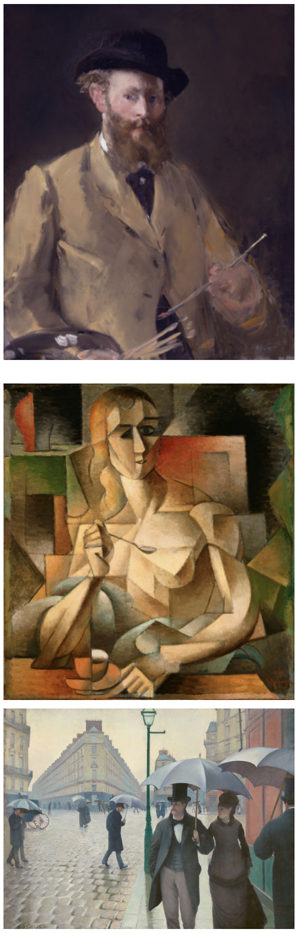
The equal-weight model balances these competing forces: stylistic heads enrich prose, factual heads anchor surface overlap, and the result is the best joint performance across BLEU, ROUGE and the stylistic indicators. **However ultimately the "best" caption model is task-dependent; our head-weight experiments demonstrate that by simply tuning the head weights we can steer the model toward either higher factual overlap or richer stylistic vocabulary, letting practitioners trade precision for expressiveness according to their needs.**

6. Conclusion

Across all of our experiments—whether using purely supervised training or leveraging self-supervised DINO pretraining—we used the ConvNeXt-Base (FB-IN22K) backbone, and it consistently outperforms other pretrained architectures in extracting both stylistic and semantic cues from art images. it produces richer "art features" that capture not only what objects appear in a painting but also how they are rendered in a particular style. When fine-tuned on art-specific tasks—such as style classification, emotion prediction, or caption generation—ConvNeXt-Base (FB-IN22K)

Model Variant	BLEU	chrF	ROUGE	METEOR	BERT	Adj.%	TTR	FK
Baseline (pre-trained)	**28.7**	**38.3**	13.5	6.98	15.4	2.6	80.4	3.3
Caption-only fine-tune	16.9	28.1	15.4	8.05	**17.8**	3.9	82.0	3.1
Single-head								
Emotion ($\alpha_{\text{emo}} = 0.9$)	22.9	33.9	15.8	10.6	9.1	**5.1**	**72.1**	**6.4**
Style ($\alpha_{\text{style}} = 0.9$)	20.5	26.8	15.6	10.2	11.8	3.6	76.7	5.2
Type ($\alpha_{\text{type}} = 0.9$)	25.2	34.2	14.5	8.6	13.5	4.1	77.2	4.7
School ($\alpha_{\text{school}} = 0.9$)	16.5	14.7	16.2	10.3	12.2	4.5	77.4	5.0
Multi (all heads, 0.1 each)	24.6	38.1	**16.4**	**11.1**	11.9	4.5	77.1	5.3

Table 4. **Quantitative comparison** of baseline, caption-only fine-tune, and multi-task variants. Higher is better for BLEU, chrF, ROUGE, METEOR, BERTScore; higher adjective density (Adj.%), FK grade indicate richer language, while lower TTR indicates more lexical diversity. The best performing model for each metric is highlighted.



Baseline: a painting of a man in a suit and hat
Equally-Weighted: a painting of a man with a beard, wearing a black hat, and holding a paintbrush
Emotion-Dominant: the painting depicts a man in a white suit and a white coat, with a black hat, is shown in the foreground.
Type-Dominant: a painting of a man with a beard and mustache

Baseline: a painting of a woman sitting in front of a mirror.
Equally-Weighted: a painting of a woman's face is shown in the foreground
Emotion-Dominant: the painting depicts a woman in a white dress, with her hands on her knees and feet on the ground, as if she is looking at the viewer.

Baseline: a painting of **people** walking down a street with umbrellas
Equally-Weighted: a painting of **three people** walking in the rain
Emotion-Dominant: the painting depicts a **couple** walking through the streets of paris.
School-Dominant: The right half-length of the painting depicts a couple in the foreground

Figure 2. **Qualitative examples of generated captions.** For each artwork we show the original image, the baseline caption, and the caption from our fine-tuned model.

delivers more nuanced and accurate representations than any competing pretrained backbone, making it our recommended choice for art-focused training pipelines.

In the future, we can transfer the DINO self-supervised pretraining on ArtEmis or other art image datasets using ConvNeXt-base backbone to the encoder so that it has already internalized rich, emotion-aware art features, it adapts much more quickly and accurately to art image annotations—whether the task is style classification, attribute tagging, or caption generation.

Overall, we clearly demonstrate that augmenting BLIP with a small suite of auxiliary heads yields captions that are demonstrably richer than those produced by either the frozen or caption-only models. A uniform weight across style, type, school and emotion provides the strongest all-round performance—recovering n-gram fidelity while al-most doubling descriptive markers such as adjective density and FK grade. Head-dominant variants confirm that the framework is controllable: raising a single weight tips the prose toward that attribute, be it art-historical jargon, labels, or ”emotional” adjectives. These findings show that multi-head BLIP not only boosts caption quality but also gives curators and assistive-tech designers a simple dial for balancing factual accuracy against stylistic flare.

References

[1] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas. Artemis: Affective language for visual art, 2021. 1, 2, 10

[2] Z. Bai, Y. Nakashima, and N. García. Explain me the painting: Multi-topic knowledgeable art description generation, 2021. 2

[3] C. Che, Q. Lin, X. Zhao, J. Huang, and L. Yu. Enhancing multimodal understanding with clip-based image-to-text transformation. In *Proceedings of the 2023 6th International Conference on Big Data Technologies*, ICBDT ’23, page 414–418, New York, NY, USA, 2023. Association for Computing Machinery. 2

[4] A. Fekete, M. Pelowski, E. Specker, D. Brieber, R. Rosenberg, and H. Leder. The vienna art picture system (vaps): A data set of 999 paintings and subjective ratings for art and aesthetics research. *Psychology of Aesthetics, Creativity, and the Arts*, 17(5):660–671, 2023. 2

[5] N. García and G. Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Lecture Notes in Computer Science (LNCS), Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 676–691. Springer, 2019. 1, 2, 3, 10

[6] Y. Jiang, K. A. Ehinger, and J. H. Lau. Kale: An artwork image captioning system augmented with heterogeneous graph, 2024. 2

[7] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 4

[8] Z. Li, Q. Tran, L. Mai, Z. Lin, and A. Yuille. Context-aware group captioning via self-attention and contrastive features. https://doi.org/10.48550/arxiv.2004.03708, 2020. 2

[9] Y. Lu, C. Guo, X. Dai, and F.-Y. Wang. Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training. *Neurocomputing*, 490:163–180, 2022. 2

[10] S. M. Mohammad and S. Kiritchenko. An annotated dataset of emotions evoked by art. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018. 1, 2, 4, 10

[11] D. Ruta, A. Gilbert, P. Aggarwal, N. Marri, A. Kale, J. Briggs, C. Speed, H. Jin, B. Faieta, A. Filipkowski, Z. Lin, and J. Collomosse. Stylebabel: Artistic style tagging and captioning. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 219–236, Cham, 2022. Springer Nature Switzerland. 2

[12] I. N. Sari, R. Sugahara, and W. Du. Artistic outpainting through adaptive image-to-text and text-to-image generation. In *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, ICCAI ’24, page 20–25, New York, NY, USA, 2024. Association for Computing Machinery. 2

[13] H. Singh, A. Sharma, and M. Pant. Pixels to prose: Understanding the art of image captioning, 2024. 2

[14] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, and R. Cucchiara. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*, pages 729–740. Springer, 2019. 1, 2

[15] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic, 2022. 2

[16] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[17] Z. Zhou, Z. Zhai, X. Gao, and J. Zhu. Improved iec performance via emotional stimuli-aware captioning. https://doi.org/10.21203/rs.3.rs-6231128/v1, 2025. Preprint, Research Square. 1, 2

A. Classification Details

Split	Number of Samples	Percentage
Training	8,576	72%
Validation	2,144	18%
Test	1,192	10%

Table 5. ArtEmis Dataset split for Classification Model

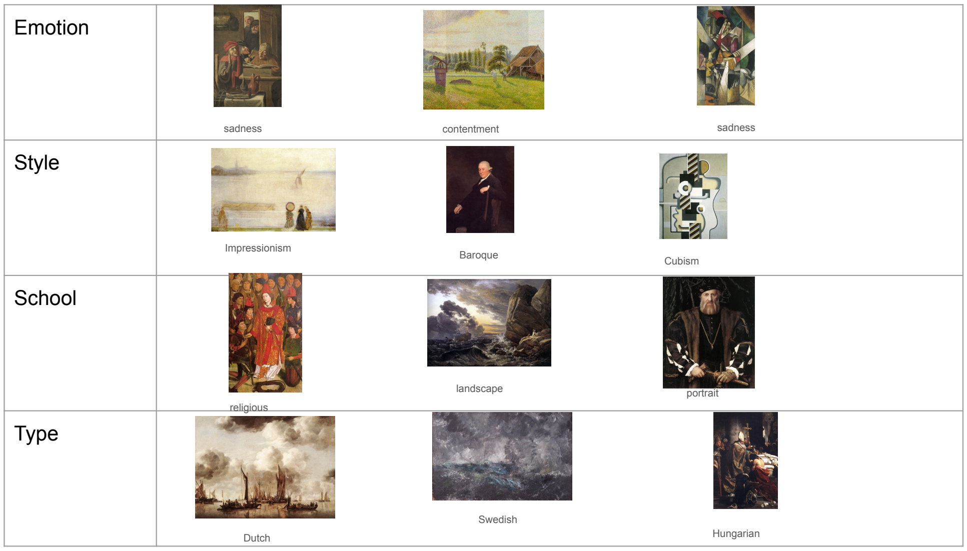


Figure 3. Examples of paintings and their corresponding labels for each classification task.

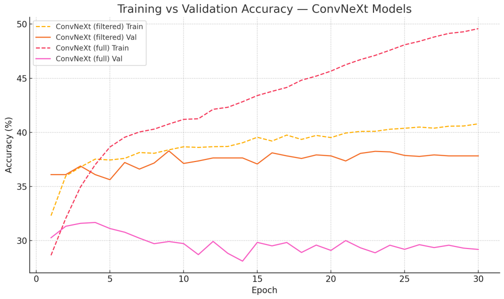


Figure 4. The training and validation curves for the Convnext model for the full and filtered training dataset sizes. The results show that the model trained on the full dataset seems to overfit while the training accuracy and the validation accuracy are closer for the model trained on the higher-quality dataset.

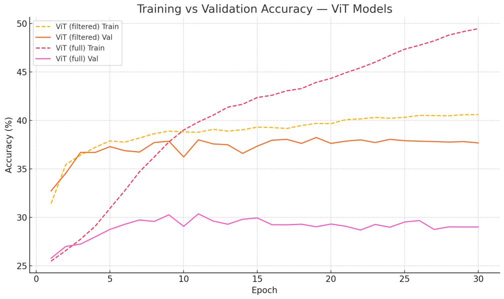


Figure 5. The training and validation curves for the vit model for the full and filtered training dataset sizes. The results show that the model trained on the full dataset seems to overfit while the training accuracy and the validation accuracy are closer for the model trained on the higher-quality dataset.

Model	Training Accuracy	Test Accuracy
vit	40.75%	36.74%
convnext	40.64%	37.00%
DINO	38.47%	35.82%

Table 6. Training and test accuracy for google/vit-base-patch16-224-in21k and convnext_base.fb_in22k (from top to bottom) on the emotion classification task using the smaller dataset with only images with more than 40 human annotations.

B. Captioning Model Details

We chose BLIP because (1) it is relatively lightweight (∼200M parameters), enabling us to rapidly experiment and fine-tune within our compute constraints; (2) it is fully open source and easy to modify end-to-end; and (3) it was pretrained on 129 million noisy image–text pairs, providing strong zero-shot performance on artwork even before fine-tuning. This makes it well suited for our setting, where the dataset of ”artistic” captions is limited in size.

This pretrained checkpoint was trained on 129 million noisy image–text pairs from the web and serves as a strong general-purpose captioning model. By comparing our fine-tuned version of BLIP, which incorporates supervision from the ArtPedia dataset as well as an auxiliary emotion classification objective, we aim to evaluate whether this additional training improves performance on metrics related to stylistic and artistic description.

Features	All Possible Labels
Emotion [1]	amusement, anger, awe, contentment, disgust, excitement, fear, sadness, something else
Style [1]	Impressionism, Northern Renaissance, Post Impressionism, Expressionism, Abstract Expressionism, Romanticism, Symbolism, Naive Art Primitivism, Cubism, Realism, Minimalism, Baroque, Art Nouveau Modern, Pop Art, Rococo, Early Renaissance, Contemporary Realism, Color Field Painting, Ukiyoe, Mannerism Late Renaissance, High Renaissance, New Realism, Fauvism, Action painting, Synthetic Cubism, Analytical Cubism
School [5]	Italian, Dutch, French, Flemish, German, Spanish, English, Netherlandish, Austrian, Hungarian, American, Danish, Swiss, Russian, Scottish, Belgian, Greek, Catalan, Bohemian, Swedish, Other, Irish, Norwegian, Polish, Finnish, Portuguese
Type [5]	religious, portrait, landscape, mythological, genre, still-life, historical, other, interior, study
Emotion [10]	agreeableness, anger, anticipation, arrogance, disagreeableness, disgust, fear, gratitude, happiness, humility, love, optimism, pessimism, regret, sadness, shame, shyness, surprise, trust, neutral

Table 7. The full list of all possible labels for each classification task.

Dataset Split	Number of Images
Training	1004
Test	87

Table 8. Preliminary training/testing data size for Captioning Model

C. Evaluation Details

For each test-set image, which is paired with ground-truth reference captions, we compare the outputs of both the baseline and fine-tuned models against these references.

- **Overlap-Based Metrics:** BLEU, CHRF, and ROUGE are traditional metrics that compare how much the generated captions overlap with reference captions. BLEU focuses on matching word sequences (n-grams), chrF works at the character level, and ROUGE-L looks for the longest shared phrases. While widely used, they don't always reflect how fluent or expressive a caption is.

- **Semantics-Aware Metrics:** METEOR goes beyond exact word matches by considering word stems, synonyms, and paraphrases, making it better at capturing meaning. BERTSCORE measures how similar the generated and reference captions are by comparing their meanings using embeddings from a pretrained language model. This makes it particularly valuable for evaluating emotional and contextual alignment.

- **Style and Expressiveness Metrics: Adjective Density (Adj.%)** shows how many words in a caption are adjectives — a higher percentage often means the caption is more descriptive or stylistic. **Type–Token Ratio (TTR)** measures vocabulary variety by comparing the number of unique words to the total word count. **Flesch–Kincaid Grade Level (FK)** estimates how complex a caption is, with higher scores indicating longer and more intricate sentence structures. These metrics help capture the stylistic richness and expressive tone important in sentiment-aware captioning.