CocoER: Aligning Multi-Level Feature by Competition and Coordination for Emotion Recognition

Supplementary Material

For convenience, the organization of the appendix is as follows:

- Technical details: Section 1, 2, 3, 4;
- Experimental details and more results: Section 5, 6, 7, 8, 9;
- Discussion of VLM: Section 10.

1. Optimization in Competition Process

The goal of the optimization in step one of competition process is to compute the descent direction for updating the linear mapping $g_{\boldsymbol{\theta}_m}$. Inspired by [8], we use parameter of the prediction layer to calculate the corresponding gradient, enabling us to leverage backpropagation and the chain rule to update all parameters in the linear mapping $g_{\boldsymbol{\theta}_m}$. Note that [8] aims to update in the training stage, while in this work we update parameters in the inference stage.

Denote the prediction layer of $g_{\boldsymbol{\theta}}$ as $\boldsymbol{\theta}^l \in \mathbb{R}^{d\times k}$, where d is the hidden dimension size and k refers to number of classes. Here, we flatten $\boldsymbol{\theta}^l \in \mathbb{R}^{d\cdot k}$ into a vector for clarity and computation convenience. We could also levarage Einstein summation convention in Pytorch, that is "torch.einsum", to obtain same output.

Recall the optimization problem in step one of competition process that is non-smooth:

$$\min_{\boldsymbol{\theta}^l} \max_{\mathbf{z}_m\in\mathbf{Z}} \mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l). \tag{1}$$

where $\mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l) = \{\mathcal{L}_{\text{hier}}(g_{\boldsymbol{\theta}^l_m}(\mathbf{z}_m), \mathbf{y_t}),\ \mathbf{z}_m \in \mathbf{Z}\}$. First we establish differentiability for $\max_{\mathbf{z}_m\in\mathbf{Z}}\{\mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l)\}$. Specifically, during the iteration of eliminating $n = \arg\max_{\mathbf{z}_m\in\mathbf{Z}}\{\mathcal{L}_{\text{hier}}(\mathcal{F}(g_{\boldsymbol{\theta}^l}(\mathbf{z}_m)), \mathbf{y_t})\}\}$, we smooth $\max()$ function by linearizing $\mathbf{f}_{\mathbf{z}_m}$ at $\boldsymbol{\theta}^l_n$ and obtain the convex approximation as follows:

$$\max_{\mathbf{z}_m\in\mathbf{Z}} \underbrace{\{\mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l_n) + \langle \nabla \mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l_n), \boldsymbol{\theta}^l - \boldsymbol{\theta}^l_n\rangle\}}_{\text{linearization term}}. \tag{2}$$

Next, in order to find a guaranteed and stable descent direction, a regularization term $\|\boldsymbol{\theta}^l - \boldsymbol{\theta}^l_n\|_2$ is added. Denote the descent direction as $\boldsymbol{\delta} = \boldsymbol{\theta}^l - \boldsymbol{\theta}^l_n$. The discrete min-max problem now is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\delta},\nu} \quad & \|\boldsymbol{\delta}\|_2 + \nu \\ \text{s.t.} \quad & \mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l_n) + \langle \nabla \mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l_n), \boldsymbol{\delta}\rangle \le \nu,\ \forall \mathbf{z}_m \in \mathbf{Z}. \end{aligned} \tag{3}$$

Problem (3) is a semi-definite quadratic programming (QP) since we choose ℓ_2 norm as the regularization term. Widely used QP algorithms, such as the active-set method, are time-consuming because their complexity depends on the dimension size d. Thus we turn to the dual problem. Consider the Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^q$ for problem (3), where $q = |\mathbf{Z}|$ refers to the number of elements in $\mathbf{Z}$:

$$\begin{aligned} L(\boldsymbol{\delta},\nu;\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\delta}\|^2 + \nu \\ + \sum_{\mathbf{z}_m\in\mathbf{Z}} \boldsymbol{\lambda}_{\mathbf{z}_m}\Big(\mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l_n) + \langle \nabla \mathbf{f}_{\mathbf{z}_m}(\boldsymbol{\theta}^l_n), \boldsymbol{\delta}\rangle - \nu\Big). \end{aligned} \tag{4}$$

According to the strong duality theorem, the minimum of the original problem is equal to the maximum of the dual problem under specific constraints:

$$\min_{\boldsymbol{\delta},\nu}\max_{\boldsymbol{\lambda}\ge 0} L(\boldsymbol{\delta},\nu;\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\ge 0}\min_{\boldsymbol{\delta},\nu} L(\boldsymbol{\delta},\nu;\boldsymbol{\lambda}). \tag{5}$$

Let $\mathbf{G} = \nabla_{\boldsymbol{\theta}^l_m}\mathbf{f}_{\mathbf{z}_m} \in \mathbb{R}^{q\times(d\cdot k)}$. By setting $\mathbf{e} = \mathbf{1}$, the above problem is equivalent to

$$\max_{\boldsymbol{\lambda}\ge 0}\min_{\boldsymbol{\delta},\nu}\big(\frac{1}{2}\|\boldsymbol{\delta}\|^2 + \nu + \boldsymbol{\lambda}^T(\mathbf{f}_{\mathbf{z}_m} + \mathbf{G}\boldsymbol{\delta} - \nu\mathbf{e})\big). \tag{6}$$

Note that

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{\delta}\|^2 + \nu + \boldsymbol{\lambda}^T(\mathbf{f}_{\mathbf{z}_m} + \mathbf{G}\boldsymbol{\delta} - \nu\mathbf{e}) \\ &\quad = \frac{1}{2}\|\boldsymbol{\delta}\|^2 + \boldsymbol{\lambda}^T(\mathbf{f}_{\mathbf{z}_m} + \mathbf{G}\boldsymbol{\delta}) + \nu(1 - \boldsymbol{\lambda}^T\mathbf{e}). \end{aligned} \tag{7}$$

If $1 - \boldsymbol{\lambda}^T\mathbf{e} \neq 0$, the objective function will be $-\infty$. Thus, we must have $1 - \boldsymbol{\lambda}^T\mathbf{e} = 0$ when the maximum is attained. The problem is converted to

$$\max_{\sum_{\mathbf{z}_m\in\mathbf{Z}}\boldsymbol{\lambda}_{\mathbf{z}_m}=1, \boldsymbol{\lambda}_{\mathbf{z}_m}\ge 0}\ \min_{\boldsymbol{\delta}} \frac{1}{2}\|\boldsymbol{\delta}\|^2 + \boldsymbol{\lambda}^T\mathbf{G}\boldsymbol{\delta} + \boldsymbol{\lambda}^T\mathbf{f}_{\mathbf{z}_m}. \tag{8}$$

Let the gradient of the inner minimization term to be zero, we have solution $\boldsymbol{\delta} = -\mathbf{G}^T\boldsymbol{\lambda} \in \mathbb{R}^{d\cdot k}$. By changing the sign of (8), the maximization term is reduced to

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & (\frac{1}{2}\boldsymbol{\lambda}^T\mathbf{G}\mathbf{G}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathbf{f}_{\mathbf{z}_m}) \\ \text{s.t.} \quad & \sum_{\mathbf{z}_m\in\mathbf{Z}}\boldsymbol{\lambda}_{\mathbf{z}_m} = 1, \boldsymbol{\lambda}_{\mathbf{z}_m} \ge 0. \end{aligned} \tag{9}$$

Suppose $\boldsymbol{\lambda}$ is the solution of the QP problem, then $\boldsymbol{\delta} = -\mathbf{G}^T\boldsymbol{\lambda} \in \mathbb{R}^{d\cdot k}$ is the solution of problem (3).

Therefore, we have (9) that is the same as (14) in the main paper. Note that after we leverage Lagrange multiplier $\boldsymbol{\lambda}$, the the complexity depends on the q, which is much smaller than hidden dimension d, so that the calculation efficiency is also improved.

Algorithm 1 Competition and Coordination Process.

Input: The set of hierarchical emotion representation $\mathbf{Z} = \{\mathbf{z_c}, \mathbf{z_b}, \mathbf{z_h}\}$. The linear mappings $g_{\boldsymbol{\theta}_m}, m \in \{\mathbf{c}, \mathbf{b}, \mathbf{h}\}$, with flattened prediction layer $\boldsymbol{\theta}_m^l \in \mathbb{R}^{d \cdot k}$. The vocabulary informed pseudo label $\mathbf{y_t}$. The objective function of emotion classification $\mathcal{L}_{\text{hier}}$. Let $q = |\mathbf{Z}|$ refers to the number of elements in $\mathbf{Z}$.
Output: Refined multi-level feature $\{\hat{\mathbf{z}}_\mathbf{c}; \hat{\mathbf{z}}_\mathbf{b}; \hat{\mathbf{z}}_\mathbf{h}\}$.

- ***Competition Process***:

while $q > 1$ **do**
  Step 1:
    $n = \arg\max_{\mathbf{z}_m \in \mathbf{Z}} \{\mathcal{L}_{\text{hier}}(g_{\boldsymbol{\theta}_m}(\mathbf{z}_m), y_\mathbf{t})\}$
    $\mathbf{f}_{\mathbf{z}_m} = \{\mathcal{L}_{\text{hier}}(g_{\boldsymbol{\theta}_m}(\mathbf{z}_m), \mathbf{y_t}),\ \mathbf{z}_m \in \mathbf{Z}\}$
    $\mathbf{G} = \nabla_{\boldsymbol{\theta}^l} \mathbf{f}_{\mathbf{z}_m} \in \mathbb{R}^{q \times (d \cdot k)}$
    Solve Lagrange multiplier $\boldsymbol{\lambda}$:
    $\min\limits_{\boldsymbol{\lambda}} \left(\frac{1}{2}\boldsymbol{\lambda}^T \mathbf{G}\mathbf{G}^T \boldsymbol{\lambda} - \mathbf{f}^T \boldsymbol{\lambda}\right)$
    s.t. $\sum_{\mathbf{z}_m \in \mathbf{Z}} \boldsymbol{\lambda}_{\mathbf{z}_m} = 1, \boldsymbol{\lambda}_{\mathbf{z}_m} \geq 0$
    if $q = |\mathbf{Z}|$ **then**
      $\boldsymbol{\delta} = -\mathbf{G}^T \boldsymbol{\lambda}$
    else
      $\boldsymbol{\delta} = \boldsymbol{\delta} - \mathbf{G}^T \boldsymbol{\lambda}$
    end if
  Step 2:
    $\mathbf{Z} = \mathbf{Z}_{-n}$ # Eliminate $\mathbf{z}_n$ in $\mathbf{Z}$
    $q = q - 1$
end while

- ***Coordination Process***:

$\tilde{\boldsymbol{\theta}}_\mathbf{c} \leftarrow \boldsymbol{\theta}_\mathbf{c}^l + \boldsymbol{\delta};\ \tilde{\boldsymbol{\theta}}_\mathbf{b} \leftarrow \boldsymbol{\theta}_\mathbf{b}^l + \boldsymbol{\delta};\ \tilde{\boldsymbol{\theta}}_\mathbf{h} \leftarrow \boldsymbol{\theta}_\mathbf{h}^l + \boldsymbol{\delta}.$
$\hat{\mathbf{z}}_\mathbf{c} = g_{\tilde{\boldsymbol{\theta}}_\mathbf{c}}(\mathbf{z_c});\ \hat{\mathbf{z}}_\mathbf{b} = g_{\tilde{\boldsymbol{\theta}}_\mathbf{b}}(\mathbf{z_b});\ \hat{\mathbf{z}}_\mathbf{h} = g_{\tilde{\boldsymbol{\theta}}_\mathbf{h}}(\mathbf{z_h})$
return $\{\hat{\mathbf{z}}_\mathbf{c}; \hat{\mathbf{z}}_\mathbf{b}; \hat{\mathbf{z}}_\mathbf{h}\}$

2. Algorithm

We offer Algorithm 1 to conclude the emotion selection and emotion representation refinement within the workspace. We clarify the input and output at the beginning of algorithm so as to highlight the function of competition and coordination process.

3. Details of Different FFN

In our proposed framework, each module has its feed-forward network (FFN) block, and all the activation function is set to GELU, as depicted in Fig. 1. The network input dimension and structure vary according to the framework architecture:

- **FFN$_0$**: In hierarchical feature extraction module, the input dimension $N \cdot d_h$ of FFN$_0$ is the flattened embedding dimension of cross-level attention, where N refers to the patch number. The output dimension is d, which is same as the dimension of $\mathbf{z}_m$.
- **FFN$_1$**: In vocabulary-informed module, the input dimension size kd of FFN$_1$ is the flattened feature size of $\omega^T \hat{\mathbf{U}}_{\text{img}}$. The output dimension is equal to the number of emotion class k. Note that we extract the embedded feature $\mathbf{z_t}$ from FFN$_1$.
- **FFN$_2$**: The input dimension of FFN$_2$ in workspace module is $4d$, which corresponds to the dimension of the concatenated feature $\mathbf{z_G}$.

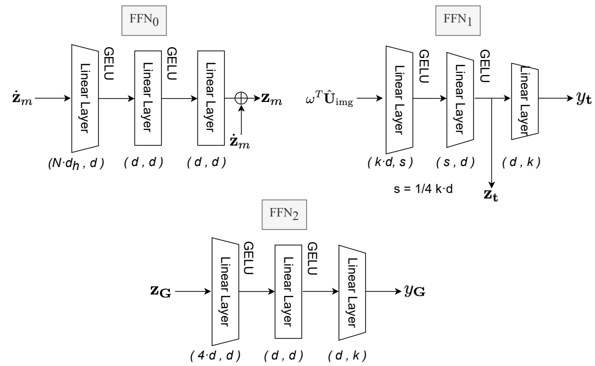


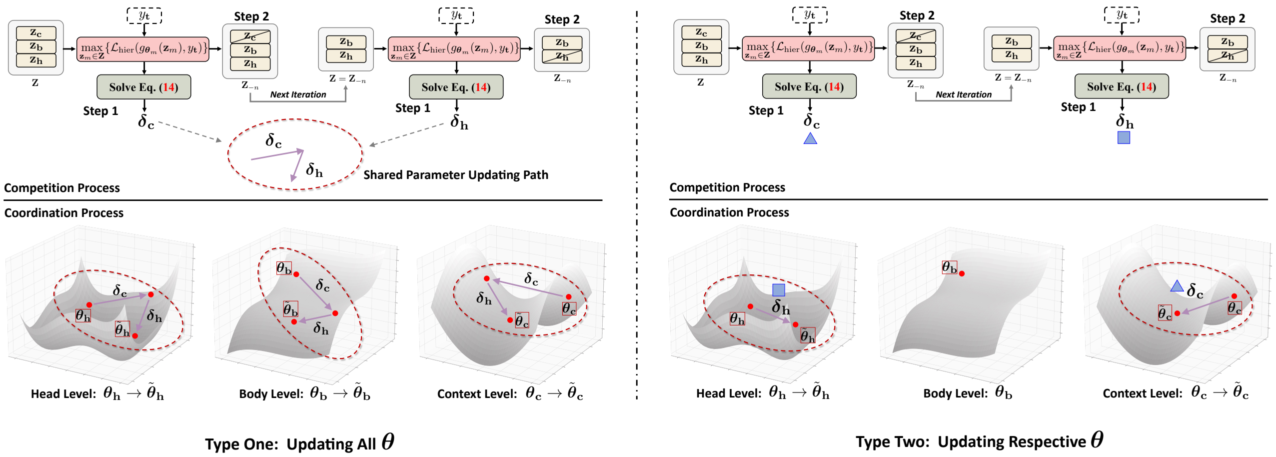
Figure 1. The FFN structures in our proposed framework.

4. Variations of Parameter Updating

In the competition and coordination process within the workspace, we provide two comparable options for the proposed inference-stage parameter update method, as shown in Fig 2. Type one is the proposed method that updates all the parameters for the linear mapping $g_{\boldsymbol{\theta}_m}$. Type two is the variation which involves updating only the parameters of the respective $g_{\boldsymbol{\theta}}$. For example, if the competition process eliminates the context and head levels, we update the parameters of the context linear mapping $g_{\boldsymbol{\theta}_\mathbf{c}}$ and head linear mapping $g_{\boldsymbol{\theta}_\mathbf{h}}$. Based on the performance results in Fig. 2, we observe that the performance of type one is better than that of type two. This is because one of the feature levels is not refined, causing the concatenated feature to be less semantic than when we update all the linear mappings to obtain all the refined features. Nevertheless, the performance of type two is higher than that of other baseline methods on both datasets.

5. Implementation Details

In the hierarchical feature extraction (HFE) module, we follow the settings in [10] to segment windows for head and body and ensured that all images contain head and body windows. We utilize ResNet-50 initialized with ImageNet pre-trained weights as the backbone to extract features for all the above-mentioned windows. All inputs are resized to $(224, 224)$. The feature map size is $(7, 7, 2048)$. We set the patch size p to 2 and use zero-padding for feature maps to ensure they are divisible by p. The hidden dimension d_h of



Datasets	Type One	Type Two
EMOTIC (mAP)	**37.32**	35.14
CAER-S (Acc.)	**91.86**	88.82

Figure 2. Comparison with two type of parameter updating in competition and coordination process.

feature map projection $\mathbf{E}$ and dimension d of linear mapping $g_{\boldsymbol{\theta}_m}$ are set to $(1024, 256)$. We use multi-head cross-level attention for all $\mathcal{T}_{\mathbf{h,c}}, \mathcal{T}_{\mathbf{b,c}}, \mathcal{T}_{\mathbf{c,c}}$, and head number is set to 3 for both datasets.

Adam is used as the optimizer with weight decay 0.0005. For EMOTIC, we use a batch size of 64 to train the model of total 9 epochs with an initial learning rate 0.00006 and a reduction factor 0.01 every 3 epochs. For CAER-S which is twice larger than EMOTIC, we train the model with a batch size of 96 for a total of 70 epochs. The initial learning rate is 0.00008, and gradually decays by a factor of 0.5 every 20 epochs. Our method is implemented on Pytorch and trained on Nivida-A100. The framework comprises approximately 352.14 million parameters. The computational cost for inference on a single sample is 0.25 ± 0.012 seconds. At last, we set random seeds to assess the performance of all the baselines. The standard deviation of recognition performance is reported in Table 1 and Table 2.

Methods	Accuracy (%)
BLIP2-6.7b [2] (*2022*)	14.21 $\pm$ 0.42
LLaVa-1.6-7b [4] (*2023*)	28.59 $\pm$ 0.22
GPT-4o [6] (*2024*)	29.33 $\pm$ 0.16
Li et al. [3] (*2021*)	32.41 $\pm$ 0.51
Hoang et al. [1] (*2021*)	35.16 $\pm$ 0.24
Mittal et al. [5] (*2020*)	35.48 $\pm$ 0.37
Wu et al. [9] (*2024*)	26.68 $\pm$ 0.28
Ours	**37.32 $\pm$ 0.05**

Table 1. Performance comparisons on EMOTIC *test* split.

Methods	Accuracy (%)
CLIP-ViT-L [7] (*2021*)	21.57 $\pm$ 0.57
BLIP2-6.7b [2] (*2022*)	14.21 $\pm$ 0.44
LLaVa-1.6-7b [4] (*2023*)	28.59 $\pm$ 0.19
GPT-4o [6] (*2024*)	29.33 $\pm$ 0.20
EfficientFace [12] (*2021*)	85.87 $\pm$ 0.24
MA-Net [11] (*2021*)	88.42 $\pm$ 0.28
Li et al. [3] (*2021*)	84.42 $\pm$ 0.34
Wu et al. [9] (*2024*)	90.83 $\pm$ 0.17
Ours	**91.86$\pm$ 0.11**

Table 2. Performance comparisons on CAER-S *test* split.

6. More Results of Ablations Studies

6.1. Image Feature Extractor

In Tab. 3, we present the ablation studies of image feature extractor (FE) for hierarchical feature extraction (in Sec 3.1 of the main paper) and vocabulary-informed module (in Sec 3.2 of the main paper). For hierarchical feature extractor, our comparison reveals that the model's representation capability is not sensitive to the number of model parameters. Changing the architecture of the extraction backbones also does not affect recognition performance. Within the vocabulary-informed module, we find that the CLIP image encoder CLIP-RN50 outperforms the transformer-based structure. Overall, the feature extractors in both modules have a relatively minor impact on the overall performance of emotion recognition in a hierarchical manner.

6.2. Hyperparameters

The ablation results of hyperparameters are presented in Tab. 4. We have the following observations:

- Tab. 4(a) provides hidden dimension d_h of linear projections for feature maps and encoding dimension d for all linear mapping and embedded feature. It can be observed that using dimension of $(1024, 256)$ significantly achieves better performance than $(512, 128)$, but there are no significant improvements when the dimension exceeds $(1024, 256)$.
- In Tab. 4(b), we demonstrate the effectiveness of hyperparameters α of the loss function in Section 3.4. It is evident that without applying regularization $\alpha\|\boldsymbol{\delta}\|_2^2$ for loss function, that is $\alpha = 0$, recognition performance shows subpar results.
- In Tab. 4(c), we clarify the weight of learning rate for $\mathcal{L}_{\text{emo}}$, denoted as $\{\beta_{\mathbf{c}}, \beta_{\mathbf{b}}, \beta_{\mathbf{h}}, \beta_{\mathbf{t}}, \beta_{\mathbf{G}}\}$ with respect to each component of $\mathcal{L}_{\text{emo}}$. We provide several different weighting strategies. From top to bottom in Tab 4(c), they are "balanced," "emphasizing V-I and Workspace," "emphasizing V-I," "emphasizing Workspace," and "emphasizing HFE." As seen in the first row, when we use the balanced weights, the model's performance is the best. From the other rows in the table, we can conclude that as the value of $\beta_{\mathbf{G}}$ increases, the model's performance improves, indicating that the overall performance of the model relies on the integrated features in the workspace. Additionally, we conducted an assessment of the presence of each loss component at each level by individually setting β of each level to a constant value while setting the remaining levels to zero. Upon examining the last three rows of the table, it is evident that omitting any of the loss components at a given level results in a significant decline in performance.
- In Tab. 4(d), we show that a smaller patch size leads to the best performance for cross-level attention. This indicates that constructing correlations between two image levels necessitates smaller image granularity.

6.3. Combined Loss Function

Since our method depends on multi-level input, with each level having its own recognition results and loss, we want to compare the effectiveness of different loss combinations that can enhance model performance. In Tab. 4(e), we conduct ablation study for combined loss $\mathcal{L}_{\text{emo}}$. First, we use only the loss from the concatenated feature ($\mathcal{L}_{\text{G}}$), and the performance turned out to be lower than that of $\mathcal{L}_{\text{emo}}$. Nevetheless, our framework can still be trained without the supervision of the hierarchical feature extraction (HFE) module. Next, we combine the concatenated loss with the vocabulary informed loss, that is $\mathcal{L}_{\mathbf{t}} + \mathcal{L}_{\mathbf{G}}$, but the performance was even worse than merely applying $\mathcal{L}_{\mathbf{G}}$. It indicates that when HFE predictions do not influence $\mathcal{L}_{\text{emo}}$, the vocabulary-informed loss $\mathcal{L}_{\mathbf{t}}$ deteriorates the emotion representation of the backbones in the HFE module.

Hierarchical Image FE (Sec 3.1)	Vocabulary-Informed Image FE (Sec 3.2)	mAP
VGG-16	CLIP-RN50	35.38
DenseNet-161		35.91
ResNet-50		**37.32**
ResNet-101		36.59
ResNet-152		36.33
ResNet-50	CLIP-RN50	**37.32**
	CLIP-RN50x16	36.21
	CLIP-RN50x64	36.05
	CLIP-B/16	35.76
	CLIP-B/32	35.52

Table 3. Ablation studies of image feature extractor (FE) on EMOTIC dataset.

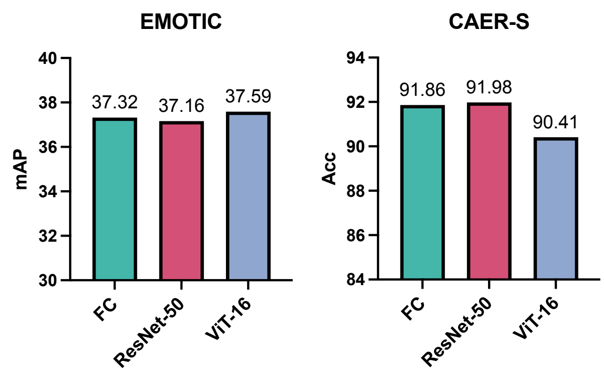


Figure 3. Comparison with different structure of $g_{\boldsymbol{\theta}}$.

6.4. Structure of the Mappings

In our framework, $g_{\boldsymbol{\theta}}$ is set as simple linear mapping for the multi-level feature. Here, we discuss the effectiveness of different network structure for $g_{\boldsymbol{\theta}}$, shown in Fig. 3. We compare the linear mapping, denoted as FC, with ResNet-50 and ViT-16. The parameter updating in workspace for ResNet-50 and ViT-16 also follows the chain rule and back-propagation. We empirically find that a linear mapping yields competitive recognition results compared to the other two larger networks across both datasets. Additionally, ResNet-50 shows relatively higher performance than ViT-16 on CAER-S, while it displays the lowest performance on EMOTIC. Thus, this indicates that emotional feature refinement does not depend on the representational ability of $\boldsymbol{\theta}$ in our proposed framework, allowing us to directly choose the linear mapping for inference efficiency using a lightweight model.

d_h, d	mAP
$512, 128$	35.05
$1024, 256$	**37.32**
$2048, 512$	36.89
$3072, 768$	36.44
$4096, 1024$	36.03

(a) Varying d for encoding dimension of g_θ.

α	mAP
0.0	34.81
0.1	**37.32**
0.4	36.70
0.8	36.16
1.0	35.87

(b) Varying α for controlling the dependency of $\boldsymbol{\delta}$.

$\{\beta_\mathbf{c}, \beta_\mathbf{b}, \beta_\mathbf{h}, \beta_\mathbf{t}, \beta_\mathbf{G}\}$	mAP
$\{0.2, 0.2, 0.2, 0.2, 0.2\}$	**37.32**
$\{0.1, 0.1, 0.1, 0.35, 0.35\}$	36.57
$\{0.2, 0.2, 0.2, 0.3, 0.1\}$	34.93
$\{0.2, 0.2, 0.2, 0.1, 0.3\}$	37.05
$\{0.3, 0.3, 0.3, 0.05, 0.05\}$	35.19
$\{0.4, 0.0, 0.0, 0.3, 0.3\}$	35.77
$\{0.0, 0.4, 0.0, 0.3, 0.3\}$	34.30
$\{0.0, 0.0, 0.4, 0.3, 0.3\}$	35.43

(c) Varying β for controlling the weight of respective loss in $\mathcal{L}_{\text{emo}}$.

p	mAP
2	**37.32**
3	35.93
4	36.77

(d) Varying patch size p.

$\mathcal{L}_{\text{emo}}$	mAP
$\mathcal{L}_{\text{emo}} = \mathcal{L}_\mathbf{G}$	36.46
$\mathcal{L}_{\text{emo}} = \mathcal{L}_\mathbf{t} + \mathcal{L}_\mathbf{G}$	35.09
$\mathcal{L}_{\text{emo}}$ (ours)	**37.32**

(e) Varying combined loss $\mathcal{L}_{\text{emo}}$.

Table 4. Additional ablation experiments on EMOTIC.

7. Saliency Maps of Multi-Level Feature

To illustrate the refinement of multi-level feature by updating $g_{\boldsymbol{\theta}}$ in inference-stage, we employed the widely-acknowledged Grad-CAM to show how the workspace module refines the multi-level features, as depicted in Fig. 4. By examining the saliency maps related to the head, body, and context, our workspace module adjusts parameter weights to focus on more semantic regions. For example, in the body level shown in the left part of Fig. 4, the saliency maps of $\tilde{\boldsymbol{\theta}}$ identify regions that are more semantically relevant than those identified by $\boldsymbol{\theta}$, displaying a binge-eating action. For the head level in all three examples, the visualization indicates that $\tilde{\boldsymbol{\theta}}$ leads to more focused attention on facial organ, whereas $\boldsymbol{\theta}$ deviates towards facial contours and the background. These visual outcomes demonstrate that our proposed method enhances multi-level semantic representation through the inference-stage parameter updating method.

8. More Results On EMOTIC Datasets

We present additional qualitative studies on the EMOTIC Dataset in Fig. 5. It can be observed that, under the guidance of Vocabulary Informed (V-I) Alignment, the workspace module has eliminated conflicting emotional information at three different image levels, thereby achieving consistent global emotional recognition. For example, in the first row of Fig. 5, the context and body level information leads to emotional disturbances (specifically, “Disconnection” in body level and “Excitement” in context level) for the child. Under the emotion selection mechanism of the global workspace, one can identify the head as the primary level that has the most significant impact on emotions, thus optimizing the final emotional output. In the last row, we observe that using only V-I for emotional recognition is insufficient because the emotion “anticipation” does not show in the V-I result but exist in the final prediction. It indicates that, through the regularization of the selection process, the final emotional recognition results are not limited by the V-I recognition outcomes.

9. More Results On CAER-S Datasets

We display additional case studies on CAER-S dataset in Fig. 6. In the task of single-label emotion recognition, the proposed method also display effectiveness. When different emotion recognition results appear at all image levels, such as in the first three rows of Fig. 6, we are able to leverage the V-I pseudo labels to refine visual selection methods. Additionally, we observe that body-level features tend to be eliminated in the first iteration, and this result is consistent with Tab. 5, which shows that the percentage of eliminated body-level features in the first iteration is 88.15%. Our proposed method adaptively selects the most consistent level of features and rectifies the final recognition result on single-label emotion recognition task.

10. Discussion on Emotion Recognition for VLM

10.1. Failure Cases Caused by Multi-Level Inputs

The recognition results from the VLM are surprisingly lower than those of other baseline models. We observe

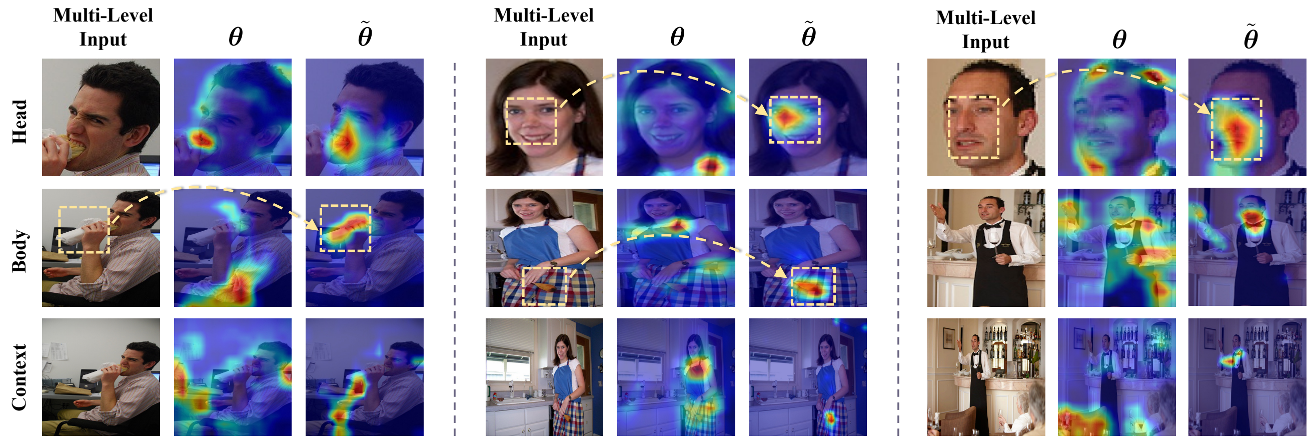


Figure 4. We display saliency map comparison of multi-level image feature. $\tilde{\theta}$ refers to the updated parameters of linear mappings after coordination process. Semantic regions, highlighted in yellow boxes, are successfully inferred using the updated linear mappings.

Dataset	Context	Body	Head
EMOTIC	10.73	25.50	**63.77**
CAER-S	06.21	**88.15**	05.64

Table 5. We report the percentage (%) of the level that is **first excluded** by our framework on both datasets. It shows that different datasets have different levels of image, which conflicts with the final emotion recognition results.

that the performance of the VLM is potentially deteriorated by inconsistent multi-level predictions. In Fig. 7, we conduct case studies using whole image inputs (left column) and multi-level inputs (right column) for GPT-4o, employing respective prompts to display multi-level results. For example, in the middle row of Fig. 7, one incorrect label, "Annoyance", appears in the original result, and the body-level input (Image 2) also produces the same incorrect label. It indicates that the VLM combines results from multi-level inputs but struggles to filter out inconsistent emotion predictions. Although prevailing VLMs consider multi-scale effects on tokenized image representation, it is also necessary to incorporate multi-level feature alignment to improve emotion recognition performance.

10.2. Improve VLM via Our Framework

In the main paper, we use modified prompts or multi-level inputs to improve emotion recognition performance by eliminating inconsistent level κ. Here, we provide detailed descriptions of the modified prompts and multi-level inputs for the emotion recognition results of VLM (using GPT-4o as example). From the results of two datasets, EMOTIC and CAER-S, we display the emotionally consistent prediction with ground-truth label (in **blue**) has been inferred through the modified multi-level inputs or prompt.

- Multi-Level$^{\kappa}$: We compare the generated results from single image input and multi-level inputs on EMOTIC in Fig. 8. We display the same input image, specifically the one in the first row of Fig 7 and Fig. 8. By eliminating the body level, which contains the label "Doubt/Confusion", our modified multi-level image successfully removes this confusing label and predicts "Fatigue" which is present in the ground truth (GT) labels.
- Prompt$^{\kappa}$: In Fig. 9, we add "Please avoid potentially inconsistent recognition result from $\{\kappa\}$" to the prompt on CAER-S dataset. We demonstrate that the results are successfully rectified by our modified prompt.

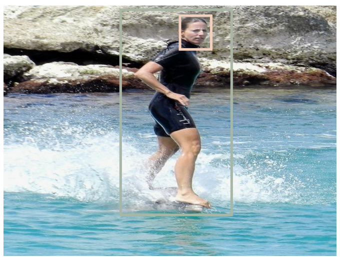
10.3. Limitation and Future Work

Limitation. Due to the efficacy of person-centric multi-level alignment, our method is well-suited in complex scenes. One potential limitation is that if individuals in different crowds interact with each other in different emotions, we should extend multi-level feature extraction to a multi-crowd manner.

Future work. Aligning multi-level inconsistent predictions is a potential way to enhance VLM performance. In the future, beyond just inputs and prompts, we will dive into the VLM model in terms of training and fine-tuning, leading to a more fundamental improvement of emotion recognition capability. However, to the best of our knowledge, there are no previous works that focus on multi-level feature alignment that is able to generalize in a multi-task paradigm within VLM architecture. We leave this challenge for future research.

GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Affection Confidence Doubt / Confusion Engagement Esteem Excitement Happiness Pleasure Surprise	Affection Anticipation Engagement Happiness Peace Pleasure	Affection Engagement Peace Pleasure	~~Annoyance~~ ~~Confidence~~ ~~Disconnection~~ ~~Doubt / Confusion~~ ~~Esteem~~ Happiness ~~Surprise~~	Anticipation Engagement ~~Excitement~~ Happiness Pleasure	Affection Engagement Happiness Peace Pleasure Yearning
		$\mathcal{L}_{\text{hier}} = \mathbf{1.01}$	$\mathcal{L}_{\text{hier}} = 2.06$	$\mathcal{L}_{\text{hier}} = 1.72$	

Exclusion Sequence: (1) Body → (2) Context → (3) Head



GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Anticipation Confidence Engagement Pleasure	Anticipation Confidence Engagement Excitement	~~Affection~~ ~~Aversion~~ ~~Sadness~~ ~~Sensitivity~~ ~~Yearning~~	~~Affection~~ ~~Disapproval~~	Anticipation Confidence Excitement ~~Pain~~ ~~Sensitivity~~ ~~Sympathy~~	Anticipation Confidence Excitement Happiness Surprise
		$\mathcal{L}_{\text{hier}} = 2.45$	$\mathcal{L}_{\text{hier}} = 2.45$	$\mathcal{L}_{\text{hier}} = \mathbf{1.51}$	

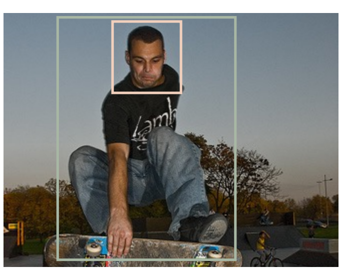
Exclusion Sequence: (1) Body → (2) Head → (3) Context

GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Confidence Excitement Happiness Pleasure Surprise	Anticipation Happiness Pleasure Surprise	~~Peace~~	Confidence ~~Fear~~ ~~Sensitivity~~	Anticipation ~~Peace~~ Pleasure Surprise ~~Sympathy~~	Anticipation Confidence Pleasure Surprise
		$\mathcal{L}_{\text{hier}} = 2.43$	$\mathcal{L}_{\text{hier}} = 1.70$	$\mathcal{L}_{\text{hier}} = \mathbf{1.49}$	

Exclusion Sequence: (1) Head → (2) Body → (3) Context

GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Affection Confidence Engagement Excitement Happiness Pleasure Surprise	Affection Anticipation Engagement Excitement Happiness Pleasure	~~Anger~~ ~~Disconnection~~ ~~Embarrassment~~ ~~Fatigue~~ ~~Surprise~~	Anticipation ~~Embarrassment~~ Engagement Pleasure ~~Sensitivity~~ ~~Suffering~~	Affection Engagement Excitement Happiness ~~Sensitivity~~	Affection Anticipation Excitement Happiness Peace
		$\mathcal{L}_{\text{hier}} = 2.43$	$\mathcal{L}_{\text{hier}} = 1.66$	$\mathcal{L}_{\text{hier}} = \mathbf{1.25}$	

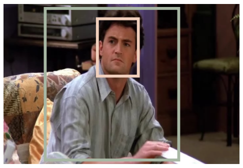
Exclusion Sequence: (1) Head → (2) Body → (3) Context



GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Anticipation Aversion Confidence Engagement Excitement Fear Happiness Sadness	Anticipation Confidence Engagement Excitement Happiness	~~Aversion~~ ~~Sadness~~	~~Anger~~ Anticipation Excitement ~~Pain~~	Anticipation Confidence ~~Fear~~ Excitement ~~Happiness~~	Confidence Engagement Excitement
		$\mathcal{L}_{\text{hier}} = 2.39$	$\mathcal{L}_{\text{hier}} = 1.68$	$\mathcal{L}_{\text{hier}} = \mathbf{1.33}$	

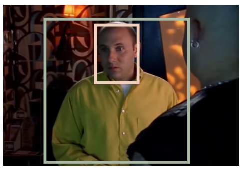
Exclusion Sequence: (1) Head → (2) Body → (3) Context

Figure 5. Case study of randomly sampled results on EMOTIC. The V-I column shows the pseudo labels of vocabulary informed module. We illustrate the elimination of inconsistent recognition results in orange, which are not shown in either the V-I column or the prediction column.



GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Anger	Anger	Anger $\mathcal{L}_{\text{hier}} = \mathbf{0.02}$	~~Happy~~ $\mathcal{L}_{\text{hier}} = 0.36$	~~Neutral~~ $\mathcal{L}_{\text{hier}} = 0.33$	Anger

Exclusion Sequence: (1) Body → (2) Context → (3) Head

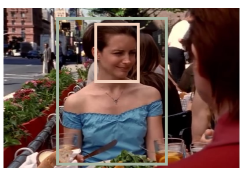


GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Fear	Fear	~~Disgust~~ $\mathcal{L}_{\text{hier}} = 0.36$	~~Surprise~~ $\mathcal{L}_{\text{hier}} = 0.31$	Fear $\mathcal{L}_{\text{hier}} = \mathbf{0.02}$	Fear

Exclusion Sequence: (1) Head → (2) Body → (3) Context

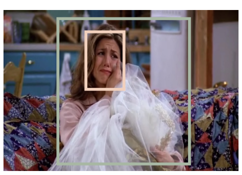
GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Neutral	Neutral	~~Fear~~ $\mathcal{L}_{\text{hier}} = 0.33$	~~Happy~~ $\mathcal{L}_{\text{hier}} = 0.31$	Neutral $\mathcal{L}_{\text{hier}} = \mathbf{0.03}$	Neutral

Exclusion Sequence: (1) Head → (2) Body → (3) Context



GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Disgust	Disgust	Disgust $\mathcal{L}_{\text{hier}} = 0.03$	~~Sad~~ $\mathcal{L}_{\text{hier}} = 0.36$	Disgust $\mathcal{L}_{\text{hier}} = \mathbf{0.02}$	Disgust

Exclusion Sequence: (1) Body → (2) Head → (3) Context



GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Sad	Sad	Sad $\mathcal{L}_{\text{hier}} = \mathbf{0.03}$	~~Anger~~ $\mathcal{L}_{\text{hier}} = 0.31$	Sad $\mathcal{L}_{\text{hier}} = \mathbf{0.03}$	Sad

Exclusion Sequence: (1) Body → (2) Context → (3) Head

GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Surprise	Surprise	Surprise $\mathcal{L}_{\text{hier}} = \mathbf{0.28}$	~~Neutral~~ $\mathcal{L}_{\text{hier}} = 0.29$	Surprise $\mathcal{L}_{\text{hier}} = \mathbf{0.28}$	Anger

Exclusion Sequence: (1) Body → (2) Context → (3) Head

GT	Prediction	Head	Body	Context	V-I (Pseudo label)
Happy	Happy	Happy $\mathcal{L}_{\text{hier}} = 0.29$	~~Fear~~ $\mathcal{L}_{\text{hier}} = 0.32$	Neutral $\mathcal{L}_{\text{hier}} = \mathbf{0.02}$	Neutral

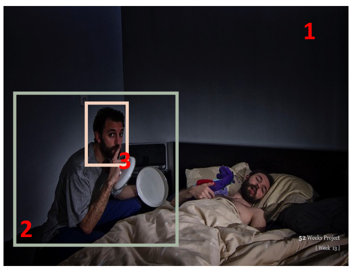
Exclusion Sequence: (1) Body → (2) Head → (3) Context

Figure 6. Case study of randomly sampled results on CAER-S. The V-I column shows the pseudo labels of vocabulary informed module. We illustrate the elimination of inconsistent recognition results in orange, which are not shown in either the V-I column or the prediction column.

Ground-truth Label

Input Image : 1 (w/o multi-level inputs)
Prompt: Whats the emotion in this picture. Choose multiple emotions from ['affection', 'anger', 'annoyance', 'anticipation', 'aversion', 'confidence', 'disapproval', 'disconnection', 'disquietment', 'doubt/confusion', 'embarrassment', 'engagement', 'esteem', 'excitement', 'fatigue', 'fear', 'happiness', 'pain', 'peace', 'pleasure', 'sadness', 'sensitivity', 'suffering', 'surprise', 'sympathy', 'yearning'].

Input Image : 1,2,3 (w/ multi-level inputs)
Prompt: Whats the emotion in this picture. Choose multiple emotions from ['affection', 'anger', 'annoyance', 'anticipation', 'aversion', 'confidence', 'disapproval', 'disconnection', 'disquietment', 'doubt/confusion', 'embarrassment', 'engagement', 'esteem', 'excitement', 'fatigue', 'fear', 'happiness', 'pain', 'peace', 'pleasure', 'sadness', 'sensitivity', 'suffering', 'surprise', 'sympathy', 'yearning']. ***Please recognize three images separately and print results.***



Ground-truth Label:
Anticipation
Confidence
Disquietment
Engagement
Esteem
Excitement
Fatigue

w/o multi-level inputs:

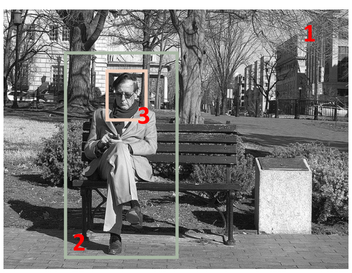
The emotions depicted in this picture appear to be:

Surprise
Doubt/Confusion
Disquietment
Engagement
The person who is awake appears to be in the middle of an unusual or humorous action, indicating a sense of surprise and engagement in the moment. The person who is asleep or just waking up might experience doubt or confusion and disquietment upon realizing what is happening.

w/ multi-level inputs:

In this picture, several emotions can be interpreted Based on the emotions depicted in the images, here are the interpretations for each:

Image 1:
Anticipation, Disconnection, Disquietment
Image 2:
Doubt/Confusion, Fatigue, Annoyance
Image 3:
Engagement, **Fatigue**, **Peace**



Ground-truth Label:
Anticipation
Confidence
Disconnection
Engagement
Pleasure

w/o multi-level inputs:

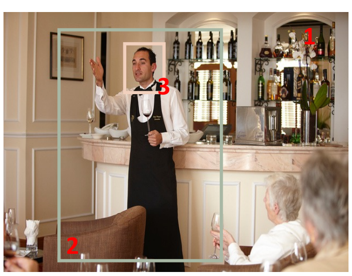
In the image, the emotions that can be interpreted include:

Disconnection: The individual appears to be seated alone, potentially lost in thought or detached from the surroundings.
Disquietment: The solitary setting, along with the posture, may suggest an underlying unease or discomfort.
Annoyance: The focused gesture of holding a device could imply frustration or irritation, possibly with the outside world.
Fatigue: The relaxed yet tense posture might hint at weariness or a need for rest.

w/ multi-level inputs:

Based on the provided images, here are the emotional responses associated with each one:

Image 1:
Emotions: **Disconnection**, **Fatigue**, **Anticipation**
Image 2:
Emotions: **Doubt/Confusion**, **Annoyance**, **Disquietment**
Image 3:
Emotions: **Sympathy**, **Sadness**, **Embarrassment**



Ground-truth Label:
Anticipation
Confidence
Engagement
Excitement
Happiness

w/o multi-level inputs:

Based on the images, the emotions that can be inferred include:

Engagement - The individual appears to be speaking or presenting, suggesting active involvement with the audience.
Confidence - The posture and presentation style indicate a sense of assurance in their role.
Doubt/Confusion - The individual appears to be questioned by audience.
Anticipation - The audience seems to be awaiting information or interaction, creating a sense of eagerness.

w/ multi-level inputs:

For the three images, I would suggest the following emotions:

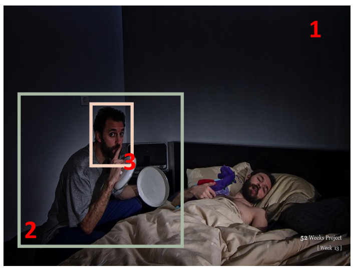
Image 1: **Anticipation**, **Engagement**, **Confidence**.
Image 2: **Anticipation**, **Engagement**, **Focus**.
Image 3: **Focus**, **Engagement**, **Doubt/Confusion**.

Figure 7. We provide failure cases of GPT-4o to illustrate that VLM also suffers from conflicting results of multi-level predictions for emotion recognition, by comparing w/ and w/o multi-level inputs. We use orange color to highlight the wrong predictions that both appear in GPT-4o results, but are not contained in ground-truth label. It indicates that **if there are inconsistent predictions at a certain image level, it can also confuse the final recognition results for VLM**. Therefore, VLM frameworks also need to eliminate multi-level pollution.

Ground-truth Label

Prompt: What's the emotion in this picture. Choose multiple emotions from ['affection', 'anger', 'annoyance', 'anticipation', 'aversion', 'confidence', 'disapproval', 'disconnection', 'disquietment', 'doubt/confusion', 'embarrassment', 'engagement', 'esteem', 'excitement', 'fatigue', 'fear', 'happiness', 'pain', 'peace', 'pleasure', 'sadness', 'sensitivity', 'suffering', 'surprise', 'sympathy', 'yearning'].

Prompt: <u>*What's the emotion in these two pictures*</u>. Choose multiple emotions from ['affection', 'anger', 'annoyance', 'anticipation', 'aversion', 'confidence', 'disapproval', 'disconnection','disquietment', 'doubt/confusion', 'embarrassment', 'engagement', 'esteem', 'excitement', 'fatigue', 'fear', 'happiness', 'pain', 'peace', 'pleasure', 'sadness', 'sensitivity', 'suffering', 'surprise', 'sympathy', 'yearning'].



Anticipation
Confidence
Disquietment
Engagement
Esteem
Excitement
Fatigue

Input Image : 1

The emotions depicted in this picture appear to be:

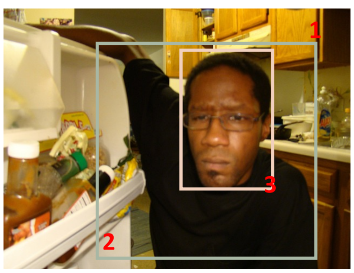
Surprise
Doubt/Confusion
Disquietment
Engagement

The person who is awake appears to be in the middle of an unusual or humorous action, indicating a sense of surprise and engagement in the moment. The person who is asleep or just waking up might experience doubt or confusion and disquietment upon realizing what is happening.

Input Image : 1,3

Based on the two images provided, the emotions that can be interpreted include:

Engagement - There is a sense of interaction and involvement in the situations depicted.
Affection - The presence of close proximity and shared moments suggests emotional warmth.
Disquietment - The overall ambiance could imply some tension or unresolved feelings.
Fatigue - The setting, especially the bed scene, may allude to tiredness or weariness.
These emotions reflect the nuances in the depicted scenarios.



Anticipation
Aversion
Disapproval
Disconnection
Doubt / Confusion
Embarrassment

Input Image : 1

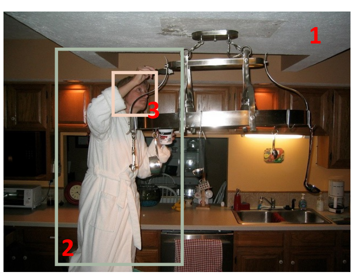
The emotions depicted in this picture appear to be:

Doubt/Confusion
Annoyance
Disquietment
The person is looking into a refrigerator with a puzzled and somewhat annoyed expression, indicating a sense of doubt or confusion about what they are seeing, as well as a hint of annoyance and disquietment.

Input Image : 1,2

Based on the images, the emotions that might be conveyed are:

Disconnection: The person's posture and position suggest a sense of separation or withdrawal from the surrounding environment.
Annoyance: The way they seem to engage with the refrigerator could indicate mild irritation or frustration.
Doubt/Confusion: The overall demeanor and body language may suggest uncertainty of something in the refrigerator.
These interpretations are subjective and could vary depending on additional context.



Anticipation
Confidence
Doubt / Confusion
Embarrassment
Engagement
Excitement
Happiness
Peace

Input Image : 1

Based on the description of the image, possible emotions could include:
Anticipation - The person appears to be engaged in a task, suggesting they are focused on something upcoming.
Engagement - The individual seems involved in what they're doing, which reflects interest and participation.
Disquietment - If the expression suggests uncertainty or discomfort in the situation, this might also apply.
These interpretations can vary based on specific visual details and context not visible in the description alone.

Input Image : 1,2

In the images, the emotions conveyed could include:
Engagement - The person appears focused on a task, indicating involvement.
Anticipation - There may be a sense of expecting results from the action being performed.
Doubt/Confusion - The actions and context suggest there might be uncertainty about what is being done or if it's correct.
These emotions together reflect a mix of being present in the moment and possibly questioning the outcome of the task.

Figure 8. Case study of enhancing answer of GPT-4o via our proposed framework. We modify the input multi-level image (in **red**) by eliminating first-excluded level that may potentially produce inconsistent recognition result. We emphasize the inferred emotion label (in **blue**) from enhanced multi-level inputs, which does not exist in the answer of single image input.

	Ground-truth Label	**Input Image : Single Image** **Prompt**: What's the emotion of this image? Choose from ['Anger','Disgust','Fear','Happy','Neutral','Sad','Surprise'].	**Input Image : Single Image** **Prompt**: What's the emotion of this image? Choose from ['Anger','Disgust','Fear','Happy','Neutral','Sad','Surprise'].
	**Anger**	The emotion of the image appears to be **Happy**.	Add prompt: *Please eliminate potentially inconsistent recognition result from* **body.** The emotion of the person in the striped sweater appears to be **Anger.**
	**Anger**	The emotion of the image appears to be **Happy**.	Add prompt: *Please eliminate potentially inconsistent recognition result from* **head.** The emotion of the person in the image appears to be **Anger**.
	**Sad**	The emotion in this image appears to be **Neutral**.	Add prompt: *Please eliminate potentially inconsistent recognition result from* **body.** The emotion in this image appears to be **Sad**. The person in the center seems to have a downcast expression, and the body language of the others suggests concern or sympathy.
	**Surprise**	The emotion of the image appears to be **Neutral**.	Add prompt: *Please eliminate potentially inconsistent recognition result from* **head.** The emotion of the person in the image appears to be **Surprise**.
	**Surprise**	**Neutral**.	Add prompt: *Please eliminate potentially inconsistent recognition result from* **body.** The emotion in this image appears to be **Surprise**.

Figure 9. Case study of enhancing answer of GPT-4o via our findings in workspace selection. Given the single image input, we add description to prompt that emphasizes the first-excluded level (in **red**) for avoiding potentially inconsistent recognition results on CAER-S. We emphasize the inferred emotion label (in **blue**) from the enhanced prompt, but the original prompt can not produce correct answer.

References

[1] Manh-Hung Hoang, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9: 90465–90474, 2021. 3

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[3] Weixin Li, Xuan Dong, and Yunhong Wang. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing*, 2021. 3

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[5] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020. 3

[6] OpenAI. GPT-4o API, 2024. Available at: `https://openai.com/index/hello-gpt-4o`, last accessed on 2024-11-01. 3

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[8] Xuli Shen, Xiaomei Wang, Qing Xu, Weifeng Ge, and Xiangyang Xue. Towards scalable and fast distributionally robust optimization for data-driven deep learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 448–457, 2022. 1

[9] Shichao Wu, Lei Zhou, Zhengxi Hu, and Jingtai Liu. Hierarchical context-based emotion recognition with scene graphs. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3725–3739, 2024. 3

[10] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context de-confounded emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 2

[11] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 3

[12] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3510–3519, 2021. 3