DialMAT: Dialogue-Enabled Transformer with Moment-Based Adversarial Training

Kanta Kaneda⋆, Ryosuke Korekata⋆, Yuiga Wada⋆, Shunya Nagashima⋆, Motonari Kambara, Yui Iioka, Haruka Matsuo, Yuto Imai, Takayuki Nishimura, and Komei Sugiura
Keio University
{k.kaneda, rkorekata, yuiga, ng.sh, motonari.k714, kmngrd1805, haruka.matsuo-25, ytim8812, t-nishimura, komei.sugiura}@keio.jp

Abstract
This paper focuses on the DialFRED task, which is the task of embodied instruction following in a setting where an agent can actively ask questions about the task. To address this task, we propose DialMAT. DialMAT introduces Moment-based Adversarial Training, which incorporates adversarial perturbations into the latent space of language, image, and action. Additionally, it introduces a crossmodal parallel feature extraction mechanism that applies foundation models to both language and image. We evaluated our model using a dataset constructed from the DialFRED dataset and demonstrated superior performance compared to the baseline method in terms of success rate and path weighted success rate. The model secured the top position in the DialFRED Challenge, which took place at the CVPR 2023 Embodied AI workshop.

1. Introduction
In this paper, we focus on the task of embodied instruction following in a setting where an agent can ask questions to the human and utilize the information provided in the response to enhance its ability to complete the task effectively.

The main contributions of this paper are as follows:
• We introduce Moment-based Adversarial Training (MAT) [4] to incorporate adversarial perturbations into the latent space of language, image, and action.
• We introduce a crossmodal parallel feature extraction mechanism to both language and image using foundation models [3, 6].

2. Problem Statement
This study focuses on the DialFRED task [1], which involves embodied instruction following. In this task, an agent has the ability to actively ask questions to the human user and utilize the information provided in the response to enhance its effectiveness in completing the task. DialFRED is based on the standard benchmark ALFRED [7].

*Equal contribution
manipulated object involved. In addition, $D^{(k)}$ refers to the set of question-answer statements in the $k$th subgoal, which is composed of pairs of question statements $q_i^{(k)}$ and corresponding response statements $s_i^{(k)}$.

First, the Questioner determines which question needs to be asked at time $t$. This module has the same structure as the Questioner proposed in [1], and is composed of an LSTM with an attention mechanism. The input in Questioner is $l^{(k)}$, and LSTM encoder and decoder perform multi-level classification. That is, at time $t$, it determines which of the Location, Appearance, and Direction questions to ask, and obtains the response $s^{(k)}$ for each question.

Second, the MAPer takes $x_t$ as input and outputs the robot’s behavior at time $t$. First, it computes the respective embedded representations $h_{txt}$ and $h_{deb}$ from $l^{(k)}$ using CLIP [6] and DeBERTa v3 [3]. Then, following the MAT approach, we obtain the features $h_{txt}$ by adding a learnable perturbation $\delta_{txt}$ as $h_{txt} = [h_{txt}; h_{deb}] + \delta_{txt}$. Note that $\delta_{txt}$ is updated based on the following steps. First, we compute the gradient of the loss function $E$ with respect to the perturbation $\nabla_{\delta} E$. Next, we introduce two types of moving averages using $\nabla_{\delta} E$ as follows:

$$m_t = \rho_1 m_{t-1} + (1 - \rho_1) \nabla_{\delta} E(\delta_t),$$

$$v_t = \rho_2 v_{t-1} + (1 - \rho_2) (\nabla_{\delta} E(\delta_t))^2,$$

where $t$ denotes the current perturbation update step and $\rho_1, \rho_2$ denote the smoothing coefficients for each moving average. Finally, using the above $m_t, v_t$, the update width of the perturbation $\Delta \delta_t$ is calculated as follows:

$$\delta_{t+1} = \Pi_{\|\delta\| \leq \varepsilon}(\delta_t + \frac{\Delta \delta_t}{\|\Delta \delta_t\|_F}),$$

where $\Pi_{\|\delta\| \leq \varepsilon}$ and $\|\cdot\|_F$ denote the learning rate of the MAT, minute value that prevents zero division, the projection onto the $\varepsilon$-sphere and the Frobenius norm, respectively. Next, we compute the respective embedded representations $h_{cimg}$ and $h_{res}$ from CLIP and ResNet [2]. Then, we apply MAT to obtain the feature $h_{img}$ by adding the perturbation $\delta_{img}$ as $h_{img} = [h_{cimg}; h_{res}] + \delta_{img}$. Similarly, for $s^{(k)}$ and $a_{t-1}$, MAT is applied to obtain the respective latent representations $h_{ans}$ and $h_{act}$. The following embedded representation $h^{1}$ is then input to the $N$-layer Transformer to obtain $h^{(N)}$:

$$h^{1} = [h_{txt}; h_{ans}; h_{img}; h_{act}] + E_{pos},$$

$$h^{i} = \text{Transformer}(h^{i-1}),$$

where $E_{pos}$ refers to positional encoding and embeds the positional information of each token in $[h_{txt}; h_{ans}; h_{img}; h_{act}]$. Finally, we obtain the predicted action $\hat{a}_{t}$ by applying the fully connected layer to $h^{(N)}$. We define the loss function as the cross-entropy between $\hat{a}_{t}$ and expert action $a_{t}$.

---

### Table 1. Quantitative comparison. The best scores are in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pseudo Test</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [1]</td>
<td>0.31</td>
<td>0.19</td>
</tr>
<tr>
<td>Ours (w/o MAT [4])</td>
<td>0.34</td>
<td>0.20</td>
</tr>
<tr>
<td>Ours (w/ CLIP text encoder [6])</td>
<td>0.35</td>
<td>0.22</td>
</tr>
<tr>
<td>Ours (MAT for action)</td>
<td>0.36</td>
<td>0.21</td>
</tr>
<tr>
<td>Ours (DialMAT)</td>
<td><strong>0.39</strong></td>
<td><strong>0.23</strong></td>
</tr>
</tbody>
</table>

---

![Figure 2](image.png)  
Figure 2. A successful subgoal prediction. In this case, the instruction was “Move to the floor lamp, power on the floor lamp.”

### 4. Experimental Setup

In this study, we employ the same experimental settings as DialFRED [1]. In this study, we divided the DialFRED dataset as described in [1]. As the test set of DialFRED is not publicly accessible, we further divided the valid/unseen set to pseudo_valid and pseudo_test set, comprising 685 and 678 tasks, respectively. We used the training set to train the MAPer and the valid_seen set for reinforcement learning of the Questioner.

### 5. Experimental Results

Table 1 shows the quantitative results of the baseline and the proposed methods. We evaluated the models by success rate (SR) and path weighted success rate (PWSR). Table 1 shows that the SR on the pseudo_test set for the baseline and the proposed methods were 0.31 and 0.39, respectively. Therefore, the proposed method outperformed the baseline by 0.08 points in SR. Similarly, the proposed method also outperformed the baseline method in PWSR.

Fig. 2 shows the qualitative results. In this sample, the instruction was “Move to the floor lamp, power on the floor lamp.” In this case, the robot was required to move to the floor lamp and then turn it on. The proposed method was able to appropriately navigate to the floor lamp and successfully turn it on.

### 6. Conclusions

In summary, the contributions of this work are twofold:

- We introduced MAT to incorporate adversarial perturbations into the latent spaces.
- We introduced crossmodal parallel feature extraction mechanisms to both language and image using foundation models.

**Acknowledgements** This work was partially supported by JSPS KAKENHI Grant Number JP23H03478, JST Moonshot, and NEDO.
References


