M$^3$L: Language-based Video Editing via Multi-Modal Multi-Level Transformer

Tsu-Jui Fu
Xin Wang
Scott Grafton
Miguel Eckstein
William Wang

UC Santa Barbara
Visual Editing using Natural Language

- Visual editing applications (Photoshop/Premiere) are widely used but difficult for novices
Visual Editing using Natural Language

- Visual editing applications (Photoshop/Premiere) are widely used but difficult for novices

- People can edit directly using language and improve accessibility

“a fire is on front feet of girl”

“move it to the lower right”
Language-based Video Editing (LBVE)

- Edit a source video $S$ into the target video $O$, guided by an instruction $X$
  - Scenario of $S$ is preserved, instead of completely different
  - Semantic of $O$ is presented differently, controlled by $X$

Source Video $S$

“waves down with his right hand” → Language-based Video Editing

Target Video $O$
Multi-Modal Multi-Level Transformer ($M^3L$)

- **Input:** Source $S = \{s_1, s_2, ..., s_N\}$, Instruction $X$
- **Output:** Target $O = \{o_1, o_2, ..., o_N\}$
Multi-Modal Multi-Level Transformer (M$^3$L)

- **Input:** Source $S = \{s_1, s_2, ..., s_N\}$, Instruction $X$
- **Output:** Target $O = \{o_1, o_2, ..., o_N\}$

- **Linguistic Feature:** $\{e_X, e_w\} = \text{RoBERTa}(X)$
- **Frame Feature:** $\{v_1, v_2, ..., v_N\} = \text{3D ResNet}(\{s_1, s_2, ..., s_N\})$
Multi-Modal Multi-Level Transformer (M\(^3\)L)

- **Input**: Source \(S = \{s_1, s_2, ..., s_N\}\), Instruction \(X\)
- **Output**: Target \(O = \{o_1, o_2, ..., o_N\}\)

- **Linguistic Feature**: \(\{e^X, e_w\} = \text{RoBERTa}(X)\)
- **Frame Feature**: \(\{v_1, v_2, ..., v_N\} = \text{3D ResNet}\{\{s_1, s_2, ..., s_N\}\}\)

- **M\(^3\)L**: \(d_i = T(\{o_1, ..., o_{i-1}\} \mid v^s, \{e^X, e_w\})\)
  - **Encoder**: \(f_i^s = \text{GF}(\text{LF}(v^s, e_w), e^X_i)\)
  - **Decoder**: \(f_i^o = \text{LF}(\text{GF}(v^o, e^X \mid f^s_i), e_w)\)
Multi-Modal Multi-Level Transformer (M$^3$L)

- **Input**: Source $S = \{s_1, s_2, ..., s_N\}$, Instruction $X$
- **Output**: Target $O = \{o_1, o_2, ..., o_N\}$

- **Linguistic Feature**: $\{e_X, e_w\} = \text{RoBERTa}(X)$
- **Frame Feature**: $\{v_1, v_2, ..., v_N\} = 3D \text{ ResNet}\{\{s_1, s_2, ..., s_N\}\}$

- **M$^3$L**: $d_i = T(\{o_1, ..., o_{i-1}\} \mid v^s, \{e_X, e_w\})$
  - Encoder: $f_i^s = \text{GF}(\text{LF}(v^s, e_w), e_X)$
  - Decoder: $f_i^o = \text{LF}(\text{GF}(v^o \mid f_i^s), e_w)$

- **Frame Generation**: $o_i = U(d_i)$
Multi-Level Fusion

- Both video and language are multi-level conveyed

- Follow multi-head attention (MHA)
  - Local-level Fusion (LF): single frame $\leftrightarrow$ word token
  - Global-level Fusion (GF): video sequence $\leftrightarrow$ whole instruction
Learning of M$^3$L

- Editing Loss $L_E$: $\text{MSE}(o_i', o_i)$

- Dual Discriminator ($D$)
  - Frame Quality: $\log(1- D_a(o_i'))$
  - Temporal Consistency: $\log(1- D_t(\{o_i', ..., o'_{i+K}\}))$

Initialize $T$, $U$, $D$
while TRAINING do
  $\{v_1, ..., v_N\} = 3D \text{ResNet}(S)$
  $e_x, \{e_{w1}, ..., e_{wN}\} = \text{RoBERTa}(X)$
  for $i \leftarrow 1$ to $N$ do
    $d_i \leftarrow T(\{o_1, ..., o_{i-1}\} \mid v, \{e_x, e_w\})$  
      \> Eq. 7
    $\hat{o}_i \leftarrow U(d_i)$
    $\mathcal{L}_E \leftarrow \text{visual difference loss with } O$
      \> Eq. 9
    $\mathcal{L}_G \leftarrow \text{video quality loss from } D$
      \> Eq. 10
  Update $T$ and $U$ by minimizing $\mathcal{L}_G + \mathcal{L}_E$
  $\mathcal{L}_D \leftarrow \text{discrimination loss for } D$
    \> Eq. 11
  Update $D$ by maximizing $\mathcal{L}_D$
end for
end while
"change the direction from lower left to upper right and the number from 5 to 0"

"move to the right front and change the large blue rubber into the small yellow metal"

"makes a cup gesture and turns his hand in a circle"
Experiments

- Collected Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Train / Test</th>
<th># Frame</th>
<th># Word</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-MNIST</td>
<td>11,070 / 738</td>
<td>354,240</td>
<td>16.0</td>
<td>64x64</td>
</tr>
<tr>
<td>M-CLEVR</td>
<td>10,133 / 729</td>
<td>217,240</td>
<td>13.4</td>
<td>128x128</td>
</tr>
<tr>
<td>E-JESTER</td>
<td>14,022 / 885</td>
<td>59,508</td>
<td>9.9</td>
<td>100x176</td>
</tr>
</tbody>
</table>
Experiments

- Collected Dataset

- Baselines: concatenate linguistic feature with visual feature for LBVE
  - pix2pix: frame-by-frame video translation
  - vid2vid: video-to-video synthesis with temporal discriminator
  - E3D-LSTM: CNN-LSTM for video prediction

pix2pix: [CVPR’17] Image-to-Image Translation with Conditional Adversarial Networks
vid2vid: [NeurIPS’18] Video-to-Video Synthesis
E3D-LSTM: [ICLR’19] Eidetic 3D LSTM: A Model for Video Prediction and Beyond
Experiments

- Collected Dataset
- Baselines

Evaluation Metrics
- **VAD**: video feature distance with ground-truth \( O \)
- **OA**: object accuracy in generated \( O' \)
- **mIoU**: mean intersection over union between \( O \) and \( O' \)
- **GA**: gesture accuracy of generated E-JESTER \( O' \)
Experiments

- Quantitative Results
  - pix2pix: insufficient video temporal
  - vid2vid & E3D-LSTM: lack of explicit cross-modal modeling
  - M^3L: incorporate multi-level fusion to achieve the best performance

<table>
<thead>
<tr>
<th>Method</th>
<th>M-MNIST</th>
<th>M-CLEVR</th>
<th>E-JESTER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VAD ↓</td>
<td>OA ↑</td>
<td>mIoU ↑</td>
</tr>
<tr>
<td>pix2pix</td>
<td>3.05</td>
<td>87.7</td>
<td>64.1</td>
</tr>
<tr>
<td>vid2vid</td>
<td>2.30</td>
<td>87.5</td>
<td>77.9</td>
</tr>
<tr>
<td>E3D-LSTM</td>
<td>2.10</td>
<td>90.4</td>
<td>81.3</td>
</tr>
<tr>
<td>M^3L</td>
<td>1.90</td>
<td>93.2</td>
<td>84.7</td>
</tr>
</tbody>
</table>
Experiments

- Ablation Study
  - **Instruction is necessary** for controllable video editing
  - **Multi-level Fusion (MLF) further benefits** cross-model modeling

<table>
<thead>
<tr>
<th>Ablation Settings</th>
<th>E-JESTER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VAD ↓</td>
</tr>
<tr>
<td>Instruction</td>
<td>MLF</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Experiments

- **Ablation Study**

- **Zero-shot Generalization:** blue square + red circle $\rightarrow$ blue circle
  - Filter $\frac{10}{40}$ number-direction combinations for M-MNIST
  - Filter $\frac{12}{96}$ size-color-material-shape combinations for M-CLEVR
  - **MLF helps generalization even training with zero-shot examples**

<table>
<thead>
<tr>
<th>Ablation</th>
<th>M-MNIST</th>
<th>M-CLEVR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VAD ↓</td>
<td>OA ↑</td>
</tr>
<tr>
<td>MLF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>2.35</td>
<td>87.5</td>
</tr>
<tr>
<td>✗</td>
<td>2.64</td>
<td>82.6</td>
</tr>
</tbody>
</table>
Experiments

- Ablation Study
- Zero-shot Generalization
- Human Evaluation

<table>
<thead>
<tr>
<th></th>
<th>w/ MLF</th>
<th>w/o MLF</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Quality</td>
<td>67.1%</td>
<td>27.1%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Video-Instruction Align.</td>
<td>53.3%</td>
<td>35.1%</td>
<td>11.6%</td>
</tr>
<tr>
<td>Simil. to GT Video</td>
<td>59.6%</td>
<td>28.9%</td>
<td>11.6%</td>
</tr>
</tbody>
</table>
### Qualitative Examples

<table>
<thead>
<tr>
<th>Source</th>
<th>Ground Truth</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="source_1" alt="Images" /></td>
<td><img src="ground_truth_1" alt="Images" /></td>
<td><img src="ours_1" alt="Images" /></td>
</tr>
<tr>
<td>&quot;change the <strong>number to 2</strong>&quot;</td>
<td>&quot;move to the <strong>front</strong> and change the small cyan metal sphere into the <strong>large yellow rubber cube</strong>&quot;</td>
<td>&quot;<strong>uses two fingers to raise a line with his right hand</strong>&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Ground Truth</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="source_2" alt="Images" /></td>
<td><img src="ground_truth_2" alt="Images" /></td>
<td><img src="ours_2" alt="Images" /></td>
</tr>
<tr>
<td>&quot;change the <strong>direction from upper right to lower right and the number from 1 to 8</strong>&quot;</td>
<td>&quot;change the brown metal sphere into the <strong>blue rubber cube and move it to the left</strong>&quot;</td>
<td>&quot;<strong>motions her right hand from left to right while showing two fingers</strong>&quot;</td>
</tr>
</tbody>
</table>
Qualitative Examples

Source 7 7 7 7 7 3 3 3 3 3 3
Ground Truth 3 3 3 3 3 3 3 3 3 3 3
Ours 3 3 3 3 3 3 3 3 3 3 3

“change the number to 3”

“move to the left front and change the large yellow cylinder into the small purple cube”

“rotates and swipes her right hand from left to right”

Source 4 4 4 4 4 4 4 4 4 4 4
Ground Truth 4 4 4 4 4 4 4 4 4 4 4
Ours 4 4 4 4 4 4 4 4 4 4 4

“change the number from 1 to 2 and the direction from upper left to upper right”

“move to the left front and change the large purple into the small gray”

“raising and opening the index and thumb fingers”