Language-based Video Editing via Multi-Modal Multi-Level Transformer

Tsu-Jui Fu†, Xin Eric Wang‡, Scott T. Grafton†, Miguel P. Eckstein†, William Yang Wang‡

†UC Santa Barbara ‡UC Santa Cruz
{tsu-juifu, william}@cs.ucsb.edu, xwang366@ucsc.edu
{scott.grafton, miguel.eckstein}@psych.ucsb.edu

Abstract

Video editing tools are widely used nowadays for digital design. Although the demand for these tools is high, the prior knowledge required makes it difficult for novices to get started. Systems that could follow natural language instructions to perform automatic editing would significantly improve accessibility. This paper introduces the language-based video editing (LBVE) task, which allows the model to edit, guided by text instruction, a source video into a target video. LBVE contains two features: 1) the scenario of the source video is preserved instead of generating a completely different video; 2) the semantic is presented differently in the target video, and all changes are controlled by the given instruction. We propose a Multi-Modal Multi-Level Transformer (M3L-Transformer) to carry out LBVE. The M3L-Transformer dynamically learns the correspondence between video perception and language semantic at different levels, which benefits both the video understanding and video frame synthesis. We build three new datasets for evaluation, including two diagnostic and one from natural videos with human-labeled text. Extensive experimental results show that M3L-Transformer is effective for video editing and that LBVE can lead to a new field toward vision-and-language research.

1. Introduction

Video is one of the most direct ways to convey information, as people are used to interacting with this world via dynamic visual perception. Nowadays, video editing tools like Premiere and Final Cut are widely applied for digital design usages, such as film editing or video effects. However, those applications require prior knowledge and complex operations to utilize successfully, which makes it difficult for novices to get started. For humans, natural language is the most natural way of communication. If a system can follow the given language instructions and automatically perform related editing actions, it will significantly improve accessibility and meet the considerable demand.

In this paper, we introduce language-based video editing (LBVE), a general V2V task, where the target video is controllable directly by language instruction. LBVE treats a video and an instruction as the input, and the target video is edited from the textual description. As illustrated in Fig. 1, the same person performs different hand gestures guided by the instruction. Different from text-to-video (T2V) [33, 3, 44, 39], video editing enjoys two following feature: 1) the scenario (e.g., scene or humans) of the source video is preserved instead of generating all content from scratch; 2) the semantic (e.g., property of the object or moving action) is presented differently in the target video. The main challenge of LBVE is to link the video perception with language understanding and reflect what semantics should be manipulated during the video generation but under a similar scenario. People usually take further editing steps onto a base video rather than create all content from the beginning. We believe that our LBVE is more practical and corresponding to human daily usage.

To tackle the LBVE task, we propose a multi-modal multi-level transformer (M3L) to perform video editing conditioning on the guided text. As shown in Fig. 2, M3L contains a multi-modal multi-level Transformer where the encoder models the moving motion to understand the entire
video, and the decoder serves as a global planner to generate each frame of the target video. For better video perception to link with the given instruction, the incorporated multi-level fusion fuses between these two modalities. During encoding, the local-level fusion is applied with the text tokens for fine-grained visual understanding, and the global-level fusion extracts the key feature of the moving motion. Reversely, during decoding, we first adopt global-level fusion from whole instruction to give a high-level plan for the target video, and then the local-level fusion can further generate each frame in detail with the specific property. With multi-level fusion, $M^3L$ learns explicit vision-and-language perception between the video and given instruction, yielding better video synthesis.

For evaluation, we collect three datasets under the brand-new LBVE task. There are E-MNIST and E-CLEVR, where we build from hand-written number recognition MNIST [31] and compositional VQA CLEVR [26], respectively. Both E-MNIST and E-CLEVR are prepared for evaluating the content replacing (different numbers or shapes and colors) and semantic manipulation (different moving directions or related positions). To investigate the capability of LBVE for natural video with open text, E-JESTER is built upon the same person performing different hand gestures with human-labeled instruction.

Our experimental results show that the multi-modal multi-level transformer ($M^3L$) can carry out the LBVE task, and the multi-level fusion further helps between video perception and language understanding in both aspects of content replacing and semantic manipulation. In summary, our contributions are four-fold:

- We introduce the LBVE task to manipulate video content controlled by text instructions.
- We present $M^3L$ to perform LBVE, where the multi-level fusion further helps between video perception and language understanding.
- For evaluation under LBVE, we prepare three new datasets containing two diagnostic and one natural video with human-labeled text.
- Extensive ablation studies show that our $M^3L$ is adequate for video editing, and LBVE can lead to a new field toward vision-and-language research.

2. Related Work

Language-based Image Editing Different from text-to-image (T2I) [43, 48, 55], which generates an image that matches the given instruction, language-based image editing (LBIE) understands the visual difference and edits between two images based on the guided text description. Image Spirit [10] and PixelTone [30] first propose the LBIE framework but accept only rule-based instruction and pre-defined semantic labels, which limits the practicality of LBIE. Inspired by numerous GAN-based methods [47, 74, 71] in T2I, there are some previous works [9, 52] perform LBIE as image colorization by the conditional GAN. Since humans do not always finish editing all-at-once but will involve several different steps, iterative LBIE (ILBIE) [14, 15] is proposed to imitate the actual process by the multi-turn manipulation and modeling the instructed editing history. Similar to LBIE, language-based video editing (LBVE) is to edit the content in a video by the guided instruction. To perform LBVE, it is required to model the dynamic visual perception instead of just a still image and consider the temporal consistency of each frame during the generation to make a smooth result video.

Language-based Video Generation Generative video modeling [59, 4, 3, 45, 38, 53, 27, 66, 2, 13, 22, 50, 49, 60, 41, 56, 11, 23, 18, 19, 39] is a widely-discussed research topic that looks into the capability of a model to generate a video purely in pixel space. Built upon video generation, text-to-video (T2V) [33, 3, 44, 39] synthesizes a video by the guided text description, which makes the video output controllable by the natural language. In this paper, we investigate the video editing task, which replaces the specific object with different properties or changes the moving motion in the input video. Different from generating video from scratch, video editing requires extracting the dynamic visual perception of the source video and manipulating the semantic inside to generate the target video.

Video-to-Video Synthesis Video super-resolution [25, 1, 34, 36], segmentation video reconstruction [63, 62], video style transfer [8, 67, 12], or video inpainting [6, 28, 70] can be considered as the particular case of video-to-video synthesis (V2V). Since they all depend on the task themselves, the variability between source-target is still under the problem-specific constraint. Among them, video prediction [46, 61, 32, 17], which predicts future frames conditioning on the given video, is one of the most related to our present LBVE task. Both video prediction and LBVE should understand the hidden semantic of the given video first and then predict the target frames with different content inside. While for video prediction, there are many possibilities of appeared future events, which makes it not deterministic for real-world usage [39]. On the other hand, LBVE is controllable by the given instruction, which involves both content replacing (object changing) and semantic manipulation (moving action changing). With the guided text description, LBVE can perform V2V with content editing and lead to predictable target video.

3. Language-based Video Editing

3.1. Task Definition

We study the language-based video editing (LBVE) task to edit a source video $S$ into a target video $O$ by a given
Figure 2. An overview architecture of our multi-modal multi-level transformer (M$^3$L). M$^3$L contains the multi-modal multi-level transformer $T$ to encode the source video $S$ and decode for the target video frame $o$ by the multi-level fusion (MLF).

Instruction $X$, as shown in Fig. 1. Specifically, the source video $S$ contains $N$ frames as $\{s_1, s_2, ..., s_N\}$, and the instruction $X = \{w_1, w_2, ..., w_L\}$ where $L$ is the number of word token in $X$. The target video $O$ also includes $N$ frames as $\{o_1, o_2, ..., o_N\}$. For LBVE, the model should preserve the scenario from $S$ but change the related semantics in $O$ guided by $X$. Note that the editing process is at a pixel level where the model has to generate each pixel of each frame and then assemble them as the target video.

3.2. Overview

An overview of our multi-modal multi-level transformer (M$^3$L) for LBVE is illustrated in Fig. 2. M$^3$L first extracts the frame feature $v_i$ for the frame $s_i$ in the source video $S$; the sentence embedding $e_X$ and each word embedding $e_{w_i}$ for the instruction $X$. Then, the multi-modal multi-level transformer $T$ is proposed to model the sequential information of the source and the target video as the decoding feature $d_i$. In particular, the multi-level fusion (MLF) performs the cross-modal fusion between video $v$ and instruction $\{e_X, e_{w_i}\}$. The local-level fusion (LF) extracts which portion is perceived by token $e_{w_i}$ across all words in $X$. Besides, the global-level fusion (GF) models the interaction between the entire video perception and the semantic motion from the whole instruction $e_X$. Finally, with $d_i$, the generator $U$ generates the frame $o_i$ in the target video $O$. In addition, we apply the dual discriminator $D$, where the frame discriminator $D_a$ helps the quality of every single frame, and the temporal discriminator $D_t$ maintains the consistency as a smooth output video.

Frame and Linguistic Feature Extraction To perform the LBVE task, We first apply a 3D ResNet and RoBERTa language model [37] to extract the frame feature $v$ and linguistic feature $\{e_X, e_{w_i}\}$ for the two modalities independently:

\[
\begin{align*}
\{v_1, v_2, ..., v_N\} &= 3D \text{ResNet}\{s_1, s_2, ..., s_N\}, \\
e_X; \{e_{w_1}, e_{w_2}, ..., e_{w_L}\} &= \text{RoBERTa}(X),
\end{align*}
\]

where $e_{w_i}$ is the word embedding of each token $w_i$, $e_X$ is the entire sentence embedding of $X$, and $L$ represents the length of the instruction $X$. In detail, $v \in \mathbb{R}^{H \times W \times C_v}$ and each $e \in \mathbb{R}^{C_x}$, where $C_v$ and $C_x$ is the feature dimension of vision and language, respectively.

3.3. Multi-Modal Multi-Level Transformer

As illustrated in Fig. 2, with the frame feature $v$ and linguistic feature $\{e_X, e_{w_i}\}$ as the inputs, the multi-modal multi-level transformer $T$ contains an encoder to model the sequential information of the source video $S$ with the given instruction $X$, and a decoder to acquire the decoding feature $d_i$ for generating the target video frame $o_i$. Both the encoder and decoder are composed of multi-level fusion (MLF), which is applied to fuse between vision and language with aspects from different levels.
Multi-Level Fusion

Both video and language are multi-level conveyed, where video is composed of a series of image frames and language is a set of word tokens with a specific order. The multi-level fusion (MLF) consists of the local-level fusion (LF) to fuse between a single frame and each word token, and the global-level fusion (GF) models the entire video sequence with the whole instruction. The computation flow of MLF is illustrated in Fig. 3. Both LF and GF are computed with the multi-head attention (MHA) [58]. MHA acquires the weighted-sum of the value feature (V) by considering the correlation between the query feature (Q) and the key feature (K):

$$\text{MHA}(Q, K, V) = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) V. \tag{2}$$

For the local-level fusion (LF), it investigates which portion should be focused by each word $e_w$ in a single frame $v_i$. We provide the relative spatial information by concatenating a 8-D spatial coordinate feature $P$ [35] with $v_i$ as $p^l$. To fuse between vision and language, we apply the self-attention mechanism (SelfAtt) [73, 72] upon the concatenated feature $q_i^l$ to capture the correlation between word expression and visual context into $s_i^l$. Different from CMSA [72], which concatenates frame feature with all token embedding directly, our LF further considers the importance of each token. We adopt a 1-layer convolutional net (Conv) to extract the context-only visual feature $c_i^l$ along the channel of $v_i$; and the widely-used dot-product attention (DotAtt) [69, 7] for the word-focused visual feature $d_i^l$ with each word $e_{wi}$. Therefore, the correlation between $c_i^l$ and $d_i^l$ can be considered as the important portion of word $w_i$ for our LF. We treat the context-only visual feature $c_i^l$ as K, the word-focused visual feature $d_i^l$ as Q, and the cross-modal feature $s_i^l$ as V to perform LF through MHA. We also utilize the residual connection [20, 58] in LF:

$$\text{LF}(v_i^l) = v_i^l \oplus \text{MHA}(c_i^l, d_i^l, s_i^l), \tag{3}$$

where

$$p_i^l = [v_i^l, P], q_i^l = \{[v_i^l, P, e_{w_i}], ..., [v_i^l, P, e_{w_N}]\},$$
$$c_i^l = \text{Conv}^l(p_i^l),$$
$$d_i^l = \text{DotAtt}(p_i^l, e_{w_i}) = \sum_{(h, w)} \text{softmax}(p_i^l \cdot W_{h,w}^T \cdot e_{w_i}) \cdot p_i^l(h, w),$$
$$s_i^l = \text{SelfAtt}(q_i^l) = \sum_{(x, y)} \text{softmax}(q_i^l \cdot d_i^l(x, y)) \cdot q_i^l(x, y),$$

and $W_{h,w}^T$ is the learnable attention matrix between $p_i^l$ and $e_{w_i}$. In this way, our LF fuses between visual context and word expression from SelfAtt and take the important portion of each token from DotAtt into consideration.

For the global-level fusion (GF), it views the entire frame sequence $\{v_1, ..., v_N\}$ with the whole instruction $e_X$ to extract the global motion of the video. Similar to LF, we acquire the fused cross-modal feature $s_n^G$ from SelfAtt, the context-only visual feature $c_n^G$ from Conv$^G$, and the sentence-focused visual feature $d_n^G$ from DotAtt for $e_n^G$. To model the entire video, we follow [58], where the video-level feature of $v_i$ can be represented as the relative weighted-sum over all frame-level $v$ and add on the positional encoding $\phi$ to incorporate the sequential order. We treat $\{s_i^G\}$ as V, $\{c_i^G\}$ as Q and $\{d_i^G\}$ as K for the correlation between a frame pair, to perform GF through MHA:

$$\text{GF}(v^G) = v^G \oplus \text{MHA}(c^G \oplus \phi, d^G \oplus \phi, s^G \oplus \phi), \tag{4}$$

where

$$p_i^G = \{[v_i^G, P], ..., [v_N^G, P]\},$$
$$q_i^G = \{[v_i^G, P, e_X], ..., [v_N^G, P, e_X]\},$$
$$c_n^G = \text{Conv}^G(p_n^G),$$
$$d_n^G = \text{DotAtt}(p_n^G, e_X),$$
$$s_n^G = \text{SelfAtt}(q_n^G).$$

By considering the correlation between frame with respect to the whole instruction from DotAtt, our GF models the video sequence as fused cross-modal feature from SelfAtt.

Encoder and Decoder

The encoder (Enc) in the multi-modal multi-level transformer $T$ serves to model the source video sequence $S$ with the given instruction $X$. Enc first adopts the local-level fusion (LF) to extract important portion from each single frame $v^o$ with each word embedding $e_{w_i}$; then the global-level fusion (GF) extracts the entire video motion with the sentence embedding $e_X$ as the cross-modal feature $f^o_{v_i}$:

$$f^o_{v_i} = \text{GF}(\text{LF}(v^o, e_{w_i}), e_X). \tag{5}$$

During decoding, the decoder (Dec) also extracts the cross-modal feature $f^o_{v_i}$ as the same way from the previous generated frames $\{o_1, ..., o_{i-1}\}$. To acquire the decoding feature $d_i$ to generate the target frame, GF is first adopted to give the high-level concept of moving motion by the interaction between the cross-modal feature $f$ from source and target, where we treat $f^o$ as the fused feature (V). Then, LF

Figure 3. The computing flow of the multi-level fusion (MLF). The local-level fusion (LF) and the global-level fusion (GF)
is applied for detailed specific property provided from word tokens $e_w$:

$$f_i^v = LF(GF(\{ v_1^v, ..., v_n^v \}, e_X \mid f^v_i), e_w). \quad (6)$$

In summary, the multi-modal multi-level transformer $T$ models the source video frame $v^s$ and gives the instruction $\{ e_X, e_w \}$, and considers the previous generated target frames $\{ o_1, ..., o_{i-1} \}$ to acquire the decoding feature $d_i$:

$$d_i = T(\{ o_1, ..., o_{i-1} \} \mid v^s, \{ e_X, e_w \}). \quad (7)$$

### 3.4. Video Frame Generation

With the decoding feature $d_i$ from $T$, we adopt ResBlocks [42] into the generator $U$ to scale up $d_i$ and synthesize into $\hat{o}_i$:

$$\hat{o}_i = U(d_i), \quad \hat{O} = \{ \hat{o}_1, \hat{o}_2, ..., \hat{o}_N \}. \quad (8)$$

We calculate the editing loss $L_E$ as the average pixel difference using mean-square loss across each frame between $O$ and $\hat{O}$:

$$L_E = \frac{1}{N} \sum_{i=1}^{N} \text{MSELoss}(o_i, \hat{o}_i). \quad (9)$$

#### Dual Discriminator

Apart from the visual difference, we also consider the video quality of our generated $\hat{O}$. Similar to DVD-GAN [11], we apply the dual discriminator $D$, where the frame discriminator $D_a$ improves the single frame quality and the temporal discriminator $D_t$ constrains the temporal consistency for a smooth output video $\hat{O}$. We treat $D_a$ as a binary classifier, which discriminates a target video frame $o$ is from ground-truth $O$ or our synthesized $\hat{O}$.

Simultaneously, $D_t$ judges that if $K$ consecutive frames are smooth and consistent enough to be a real video fragment as the binary discriminator. The video quality loss $L_G$ is computed for both frame quality and temporal consistency:

$$L_a = \frac{1}{N} \sum_{i=1}^{N} \log(1 - D_a(\hat{o}_i)),$$

$$L_t = \frac{1}{M} \sum_{i=1}^{M} \log(1 - D_t(\{ \hat{o}_i, ..., \hat{o}_{i+K-1} \})) \quad (10)$$

$$L_G = L_a + L_t,$$

where $M = N - K + 1$.

On the other hand, the dual discriminator $D$ is training to distinguish between $O$ and $\hat{O}$ by the following:

$$L_a = \frac{1}{N} \sum_{i=1}^{N} (\log(1 - D_a(o_i)) + \log(D_a(\hat{o}_i))),$$

$$L_t = \frac{1}{M} \sum_{i=1}^{M} (\log(1 - D_t(\{ \hat{o}_i, ..., \hat{o}_{i+K-1} \})) + \log(D_t(\{ o_i, ..., o_{i+K-1} \}))),$$

$$L_D = L_a + L_t.$$

Therefore, they are optimized through an alternating min-max game:

$$\min_{G} \max_{D} L_G + L_D. \quad (12)$$

### 3.5. Learning of M$^3$L

Algo. 1 presents the learning process of the proposed multi-modal multi-level transformer (M$^3$L) for LBVE. Since LBVE is also a sequential generation process, we apply the widely used teacher-forcing training trick, where we feed in the ground-truth target frame $o_{i-1}$ instead of the predicted $\hat{o}_{i-1}$ from the previous timestamp to make the training more robust. We adopt the multi-modal multi-level transformer $T$ to model the source video and input instruction, and the frame generator $U$ to generate the target video frame. During training, we minimize the video quality loss $L_G$ with the visual difference $L_E$ to optimize M$^3$L. We also update the dual discriminator $D$, including the frame discriminator $D_a$ and the temporal discriminator $D_t$, by maximizing $L_D$. Therefore, the entire optimization object can be summarized as:

$$\min_{G,E} \max_{D} L_G + L_E + L_D. \quad (13)$$

### 4. Datasets

To the best of our knowledge, there is no dataset that supports video editing with the guided text. Therefore, we build three new datasets specially designed for LBVE, including two diagnostic datasets (E-MNIST and E-CLEVR) and one human gesture dataset (E-JESTER) for the language-based video editing (LBVE) task. An overview of our built
datasets is shown in Table 1, and examples of these three datasets are illustrated in Fig. 4.

**E-MNIST** Extended from Moving MNIST [31, 53], the new E-MNIST dataset contains the instruction to describe the difference between two video clips. Hand-written numbers are moving along a specific direction and will reverse its direction if bumps into a boundary. The instructions include two kinds of editing actions: *content replacing* is to replace the specific number with the given one, and *semantic manipulation* changes the starting direction for different moving motion. We prepare two levels of E-MNIST, S-MNIST and D-MNIST. S-MNIST is an easier one and includes only a single number, so the model only needs to replace the number or change the moving direction at a time. There are two numbers in the advanced D-MNIST, where the model is required to perceive which number should be replaced and which starting direction should be changed simultaneously. For both S-MNIST and D-MNIST, there are 11,808 pairs of source-target video.

**E-CLEVR** Following CATER [16], we create each frame and combine them as the video in our E-CLEVR upon the original CLEVR dataset [26]. Each example consists of a pair of source-target videos with an instruction described the semantic altering. The editing action includes changing the property of the specific object and placing the moving object into a particular given final position. E-CLEVR contains plentiful object properties (e.g., color, shape, size, ...) and different relative positions of the final target. To highlight the importance of visual perception, not all aspects of the property will change; only the mentioned properties like the color and shape should be changed but keeps others the same. Besides, we will not always alter the moving object but sometimes the still one, where all clues are provided in the guided text. We generate 10,862 examples for the E-CLEVR dataset.

**E-JESTER** Toward human action understanding, 20BN-JESTER [40] builds a large gesture recognition dataset. Each actor performs different kinds of gesture moving in front of the camera, which brings out 27 classes in total. This setting is appropriate to the video editing task where the source-target videos are under the same scenario (same person in the same environment) but with different semantics (different hand gestures). To support our LBVE task, we prepare pairs of clips from the same person as the source-target videos and collect the human-labeled instruction by Amazon Mechanical Turk (AMT)\(^1\). In this way, we can have the natural video whose scenario is preserved, but semantic is changing with natural guided text for our E-JESTER dataset, which can be a sufficient first step for LBVE. There are 14,907 pairs in E-JESTER.

### Table 1. The statistics of our collected datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#Test</th>
<th>#Frame Avg.</th>
<th>#Word Avg.</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-MNIST</td>
<td>11,070</td>
<td>738</td>
<td>354,240</td>
<td>5.5</td>
<td>64x64</td>
</tr>
<tr>
<td>D-MNIST</td>
<td>11,070</td>
<td>738</td>
<td>354,240</td>
<td>16.0</td>
<td>64x64</td>
</tr>
<tr>
<td>E-CLEVR</td>
<td>9,233</td>
<td>1,629</td>
<td>21,7240</td>
<td>13.4</td>
<td>128x128</td>
</tr>
<tr>
<td>E-JESTER</td>
<td>14,022</td>
<td>885</td>
<td>59,508</td>
<td>9.9</td>
<td>100x176</td>
</tr>
</tbody>
</table>

\(^1\)Amazon Mechanical Turk: [https://www.mturk.com/](https://www.mturk.com/)

---

Figure 4. The sampled source videos, the ground-truth target videos, and the generated LBVE videos on all three datasets. More examples can be found on [https://lbvce2515.github.io/](https://lbvce2515.github.io/).
frame pairs as the video SSIM.

• FAD: Similar to IS [51] and FID [21], frame activation distance (FAD) is computed by the mean L2 distance of the activations from the Inception V3 [54] feature. We treat the ground-truth target video as the reference. Because of being a distance metric, a lower FAD has a higher similarity between the two videos.

• VAD: Inspired by FVD [57], we apply the 3D CNN for more accurate video feature, and compute the video activation distance (VAD) following the same equation as FAD. Specifically, ResNeXt [68] is adopted for the diagnostic E-MNIST and E-CLEVR dataset. Besides, we utilize I3D [5] to extract the action video feature for E-JESTER. Similar to FAD, a lower VAD means that videos are more related to each other.

• GA: Apart from the visual-based evaluation, we report the gesture accuracy (GA) for E-JESTER. GA is calculated as the gesture classification accuracy of the edited video by MFFs² [29]. Although the generated video may not be totally the same as the ground truth, higher GA represents that the model is able to follow the guided text and generate the corresponding type of gesture.

### Baselines
Since our LBVE is a brand new task, there is no existing baselines. We consider following methods conditioning on an instruction, by concatenating the linguistic feature with the hidden visual feature, to carry out LBVE as the compared baselines.

- **pix2pix** [24]: pix2pix is a supervised image-to-image translation approach. For the sake of video synthesis, we process the source video frame-by-frame to perform pix2pix.
- **vid2vid** [63]: vid2vid applies the temporal discriminator for better video-to-video synthesis, which considers several previous frames to model the translation.
- **E3D-LSTM** [64]: E3D-LSTM incorporates 3D CNN into LSTM for video prediction. We treat the source video as the given video and predict the remaining part as the target video.

### 5.2. Quantitative Results
Table 2 shows the overall testing results compared between the baselines and ours M³L. pix2pix only adopts image-to-image translation, resulting in insufficient output video (e.g., 63.7 SSIM under D-MNIST and 2.84 VAD under E-CLEVR). Even if vid2vid and E3D-LSTM consider temporal consistency, the lack of explicit cross-modal fusion still makes them difficult to perform LBVE. While, our M³L, which incorporates the multi-level fusion (MLF), can fuse between vision-and-language with different levels and surpass all baselines. In particular, M³L achieves the best results across all metrics under all diagnostic datasets (e.g., 93.2 SSIM under S-MNIST, 1.24 FAD under D-MNIST, and 1.96 VAD under E-CLEVR).

Similar trends can be found on the natural E-JESTER dataset. pix2pix only has 8.6% GA, which shows that it cannot produce a video with the correct target gesture. Although vid2vid and E3D-LSTM may have similar visual measurement scores to our approach, M³L achieves the highest 89.3% GA. The significant improvement of GA demonstrates that the proposed MLF benefits not only the visual quality but also the semantic of the predicted video and makes it more corresponding to the given instruction.

### 5.3. Ablation Study
**Ablation Results** Table 3 presents the testing results of the ablation setting under E-JESTER. If without the given instruction, the model lacks the specific editing target and results in poor 78.8 SSIM and 4.7% GA. The performance comprehensively improves when incorporating our proposed multi-level fusion (MLF) (e.g., VAD from 1.50 down to 1.44 and GA from 85.4% up to 89.3%). The multi-level modeling from MLF benefits not only the understanding be-

### Table 2. The overall testing results of the baselines and our M³L under the E-MMIST, E-CLEVR, and E-JESTER datasets.

<table>
<thead>
<tr>
<th></th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
<th>GA ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>pix2pix [24]</td>
<td>85.0</td>
<td>1.31</td>
<td>2.06</td>
<td>63.7</td>
<td>2.13</td>
<td>3.05</td>
<td>80.1</td>
<td>2.05</td>
<td>2.84</td>
<td>74.6</td>
<td>1.62</td>
<td>2.00</td>
<td>8.6%</td>
</tr>
<tr>
<td>vid2vid [63]</td>
<td>90.0</td>
<td>0.96</td>
<td>1.30</td>
<td>82.0</td>
<td>1.51</td>
<td>2.30</td>
<td>94.0</td>
<td>1.45</td>
<td>2.21</td>
<td>82.0</td>
<td>1.50</td>
<td>1.62</td>
<td>82.0%</td>
</tr>
<tr>
<td>E3D-LSTM [64]</td>
<td>91.8</td>
<td>0.80</td>
<td>1.29</td>
<td>84.0</td>
<td>1.41</td>
<td>2.10</td>
<td>94.2</td>
<td>1.44</td>
<td>2.11</td>
<td>88.6</td>
<td>1.10</td>
<td>1.55</td>
<td>83.6%</td>
</tr>
<tr>
<td>M³L (Ours)</td>
<td>93.2</td>
<td>0.80</td>
<td>1.28</td>
<td>87.9</td>
<td>1.24</td>
<td>1.90</td>
<td>96.8</td>
<td>1.32</td>
<td>1.96</td>
<td>91.7</td>
<td>1.04</td>
<td>1.44</td>
<td>89.3%</td>
</tr>
</tbody>
</table>

### Table 3. The ablation results when without the instruction or MLF.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>MLF</th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
<th>GA ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
<td>78.8</td>
<td>1.49</td>
<td>1.99</td>
<td>4.7%</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>89.6</td>
<td>1.16</td>
<td>1.50</td>
<td>85.4%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>91.7</td>
<td>1.04</td>
<td>1.44</td>
<td>89.3%</td>
</tr>
</tbody>
</table>

### Table 4. The testing results of the zero-shot generalization.

<table>
<thead>
<tr>
<th></th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
<th>SSIM ↑</th>
<th>FAD ↓</th>
<th>VAD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>pix2pix [24]</td>
<td>72.89</td>
<td>1.91</td>
<td>2.64</td>
<td>86.61</td>
<td>1.70</td>
<td>2.32</td>
</tr>
<tr>
<td>vid2vid [63]</td>
<td>77.84</td>
<td>1.75</td>
<td>2.35</td>
<td>90.07</td>
<td>1.51</td>
<td>2.29</td>
</tr>
<tr>
<td>E3D-LSTM [64]</td>
<td>91.7</td>
<td>1.04</td>
<td>1.44</td>
<td>91.7</td>
<td>1.04</td>
<td>1.44</td>
</tr>
</tbody>
</table>
Table 5. Human evaluation on E-JESTER. Human judges evaluate video quality, video-instruction alignment, and similarity to GT video of the generated video w/ or w/o multi-level fusion (MLF).

tween video and instruction, but also leads to accurate frame generation. The above ablation results show that the instruction is essential under the video editing task, and our MLF further helps to perform LBVE.

**Zero-Shot Generalization** To further investigate the generalizability of M^3L, we conduct a zero-shot experiment for both the D-MNIST and E-CLEVR datasets. In D-MNIST, there are 40 different object-semantic combinations^3. We remove out 10 of them in the training set (e.g., number 1 with upper left or number 3 with lower down) and evaluate under the complete testing set. For E-CLEVR, we filter out 12 kinds (e.g., small gray metal sphere or large purple rubber cube) from the total 96 possibilities^4. The results are shown in Table 4. Due to the lack of object properties or moving semantics, the model has a significant performance drop under the zero-shot settings. While, our proposed MLF helps the property and moving motion for both video perception and generation by multi-modal multi-level fusion. Therefore, MLF still improves the generalizability (e.g., VAD from 2.35 down to 2.29 under D-MNIST and SSIM from 86.6 up to 90.1 under E-CLEVR) even if training with the zero-shot examples.

**Human Evaluation** Apart from the quantitative results, we also investigate the quality of the generated video from the human aspect. Table 5 demonstrates the comparison between without and with MLF. We randomly sample 75 examples and ask three following questions: (1) Which video has better quality; (2) Which video corresponds more to the given instruction; (3) Which video is more similar to the ground-truth target video. Each example is assigned to 3 different MTurkers to avoid evaluation bias. Firstly, about 67% think that generated videos from MLF have better quality. Moreover, more than 50% of MTurkers denote that the target videos produced from MLF correspond more to the instruction and are also more similar to the ground truth. The results of the human evaluation indicate that our MLF not only helps improve the generating quality but also makes the target video more related to the guided text.

**Multi-Level Fusion** Fig. 5 illustrates the attention map from our multi-level fusion (MLF) under D-MNIST during frame decoding. The instruction tells to replacing the number 8 and change the moving direction of the number 5.

<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 Alignment</td>
<td>53.3%</td>
<td>35.1%</td>
<td>11.6%</td>
</tr>
<tr>
<td>Siml. to GT Video</td>
<td>59.6%</td>
<td>28.9%</td>
<td>11.6%</td>
</tr>
</tbody>
</table>

Table 5. Human evaluation on E-JESTER. Human judges evaluate video quality, video-instruction alignment, and similarity to GT video of the generated video w/ or w/o multi-level fusion (MLF).

During decoding, the global-level fusion (GF) first gives the high-level video motion, and the attention maps reflect the moving directions for each output frame. Then, the local-level fusion (LF) helps extract perceived feature with respect to each word token. For instance of the 3rd frame, the word ‘direction’ and ‘lower’ are described for the number 5 of its moving direction. While, the tokens ‘number’ and ‘6’ are responsible for replacing, thus they are attended on the position of the replaced number. With the cooperation of GF and LF, our MLF gives a concrete concept and brings out better target videos during the decoding.

**5.4. Qualitative Results**

Fig. 4 shows the keyframes of the generated examples of LBVE on all three datasets. For E-MNIST, we have to recognize which number should be replaced and which one will change the moving semantic. Note that the instruction only tells the replacing number, but without the style, thus our model replaces with another kind of number 2 under SMNIST. Under the advanced D-MNIST dataset, our model can replace with the number 8 and move the number 5 along the lower right with multi-level fusion. The challenge of E-CLEVR is to transform object properties and move to the different target positions related to the fixed object. The visualization examples show that our model can understand the linguistic to change the specific object into the correct properties. Also, it has the spatial concept that can perceive the final related position and maintain the moving motion. The E-JESTER dataset, which contains nature video and human-labeled instruction, requires the link of the complex natural language with the human gesture action. The presented video indicates that our model can not only preserve a similar scenario (the background and the person) in the target video but also generate the visual motion of the corresponding gesture.

**6. Conclusion**

We introduce language-based video editing (LBVE), a novel task that allows the model to edit, guided by a natural text, a source video into a target video. We present

---

^3D-MNIST: 10 different numbers and 4 different directions
^4E-CLEVR: 3 shapes, 8 colors, 2 materials, and 2 shapes
multi-modal multi-level transformer (M$^3$L) to dynamically fuse video perception and language understanding at multiple levels. For the evaluation, we release three new datasets containing two diagnostic and one natural video with human-labeled text. Experimental results show that our M$^3$L is adequate for video editing, and LBVE can bring out a new field toward vision-and-language research.
References


[19] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic Video Generation using Holistic Attribute Control. In ECCV, 2018. 2


[22] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to Decompose and Disentangle Representations for Video Prediction. In NeurIPS, 2018. 2


[31] Yann LeCun, Corrinna Cortes, and CJ Burges. MNIST Handwritten Digit Database. 2010. 2, 6


[33] Yitong Li, Martin Renqiang Min, Dinhuan Shen, David Carlson, and Lawrence Carin. Video Generation from Text. In AAAI, 2018. 1, 2

[34] Renjie Liao, Xiu Tan, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video Super-Resolution via Deep Draft-Ensemble Learning. In ICCV, 2015. 2


[73] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-Attention Generative Adversarial Networks. In PMLR, 2019. 4

[74] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In ICCV, 2017. 2