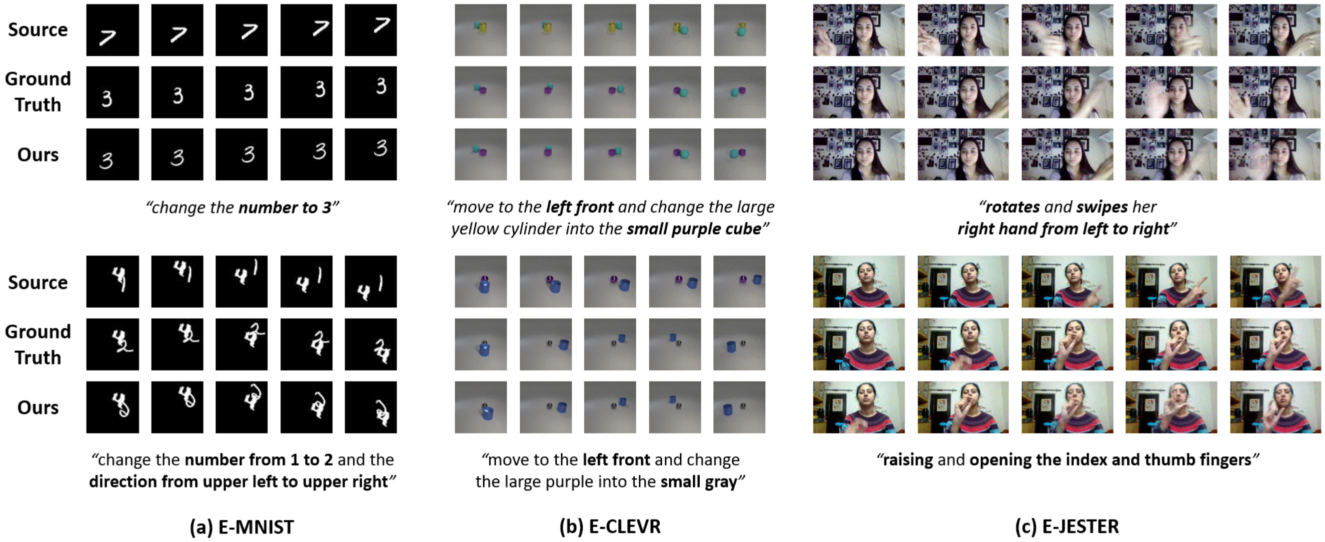
Supplementary Material for
M3L: Language-based Video Editing via Multi-Modal Multi-Level Transformers

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

[†]UC Santa Barbara [‡]UC Santa Cruz

tsu-juifu@ucsb.edu, {scott.grafton, miguel.eckstein}@psych.ucsb.edu
william@cs.ucsb.edu, xwang366@ucsc.edu



"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****"*

"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**"

(a) E-MNIST **(b) E-CLEVR** **(c) E-JESTER**

Figure 1. The sampled source videos, the ground-truth target videos, and the generated LBVE videos on all three datasets.

A. Zero-shot Generalization under E-JESTER

We conduct the zero-shot setting on E-JESTER, where the people in the testing set do not exist during training. We evaluate the generalizability of a model through editing an unseen person with a specific gesture. The results are summarized in Table 1. pix2pix [2], which only treats single frame translation, performs the worst. Both vid2vid [4] and E3D-LSTM [5] result in a significant performance drop under the zero-shot setting (*e.g.,* vid2vid drops from 82.0 GA to 73.8 and E3D-LSTM ups from 1.55 VAD to 1.79). In contrast, with the multi-level fusion (MLF) over different levels of video-and-language reasoning, our M^{3}L still maintains the lowest 1.51 VAD and the highest 86.0 GA, even encountering an unseen person.

	E-JESTER (Full)		**E-JESTER** (Zero-shot)	
	VAD ↓	GA ↑	VAD ↓	GA ↑
pix2pix [2]	2.00	8.6	2.42	8.7
vid2vid [4]	1.62	82.0	1.84	73.8
E3D-LSTM [5]	1.55	83.6	1.79	78.4
M^{3}L (Ours)	**1.44**	**89.3**	**1.51**	**86.0**

Table 1. Zero-shot Generalization under E-JESTER.

B. Human Evaluation of Baselines

We conduct a human evaluation with 30 E-JESTER examples over all baselines. Table 2 shows the mean ranking score (from 1 to 4, the higher is better) under different aspects. In general, videos produced by our M^{3}L have higher quality. Furthermore, the proposed MLF makes the editing result more related to the guided text.

	pix2pix	vid2vid	E3D-LSTM	M^{3}L
Video Quality	2.07	2.47	2.50	**2.97**
Video-Instruction Alignment	1.67	2.27	2.37	**3.67**
Similarity to GT Video	1.60	2.40	2.63	**3.37**

Table 2. Human evaluation (mean ranking score from 1 to 4, the higher is better) on E-JESTER.

C. Ablation of MLF/Discriminator

Table 3 illustrates the ablation study of multi-level fusion (MLF), including local-level (LF) and global-level fusion (GF), and dual discriminator (Dual-D) on E-CLEVR. Comparing row (b) and (c) with (a), LF contains better local perception (higher OA) between object properties and word tokens, and GF benefits the global motion (lower VAD and higher mIoU). Row (d) further shows that combining LF and GF as MLF can help both. In the end (row (e)), Dual-D enhances the video quality, leading to a comprehensive improvement.

	LF	GF	Dual-D	VAD ↓	OA ↑	mIoU ↑
(a)	✗	✗	✗	2.19	82.4	70.5
(b)	✓	✗	✗	2.25	83.4	71.7
(c)	✗	✓	✗	2.04	83.1	74.6
(d)	✓	✓	✗	2.02	83.6	75.3
(e)	✓	✓	✓	**1.96**	**84.5**	**78.4**

Table 3. Ablation study of MLF/Discriminator on E-CLEVR.

D. Multi-Modal Baseline

We consider GeNeVA [1], iterative-base LBIE, as the multi-modal baseline. For each turn, we feed in the instruction and generate a frame based on previous results and the encoded source video from LSTM. Then we compose all iterative frames as the editing video. Table 4 shows the evaluation on E-CLVER. GeNeVA has better OA and MIoU than E3D-LSTM by the self-attention module over the visual-and-linguistic feature. Upon cross-modal attention, M^{3}L further considers multi-level fusion (MLF), leading to the best results on all metrics.

Method	VAD ↓	OA ↑	mIoU ↑
E3D-LSTM	2.11	83.1	72.2
GeNeVA$^{??}$	2.13	83.3	74.5
M^{3}L	**1.96**	**84.5**	**78.4**

Table 4. The testing results of GeNeVA on E-CLEVR.

E. Limitation and Social Impact

Our M^{3}L framework treats source/target videos as fully-supervised training, which may fail for out-domain scenes and instructions. We can exploit pretrained visual-linguistic alignment (*e.g.,* CLIP [3]) to boost the editing result weakly-supervisedly. Besides, there may be an authenticity doubt for those edited videos. To mitigate this issue, we train a binary video classifier, which achieves 93% real/fake accuracy on E-JESTER. It shows that such video forensics can help video authentication of the potential negative impact.

References

[1] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W.Taylor. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *ICCV*, 2019. 2

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Nets. In *CVPR*, 2017. 1

[3] Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. Contrastive Visual-Linguistic Pretraining. In *arXiv:2007.13135*, 2020. 2

[4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *NeurIPS*, 2018. 1

[5] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3D LSTM: A Model for Video Prediction and Beyond. In *ICLR*, 2019. 1