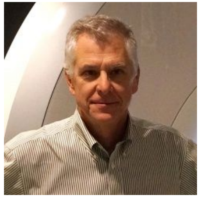
M^{3}L: Language-based Video Editing via Multi-Modal Multi-Level Transformer

Tsu-Jui Fu

Xin Wang



Scott Grafton

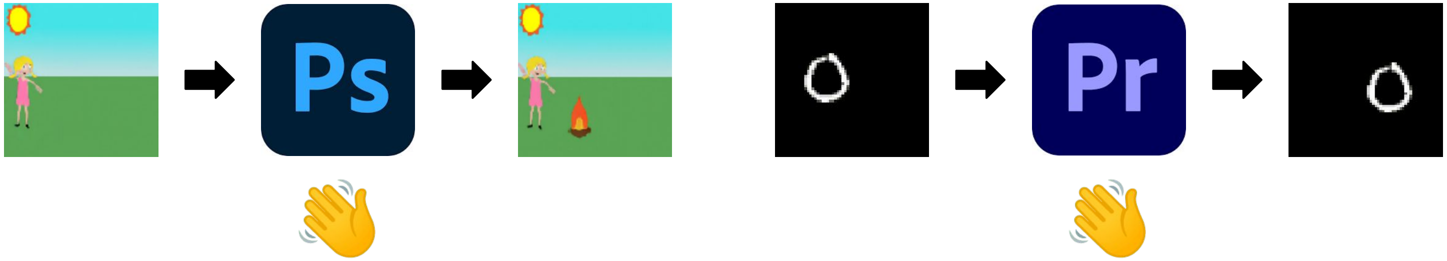
Miguel Eckstein

William Wang

UC Santa Barbara

Visual Editing using Natural Language

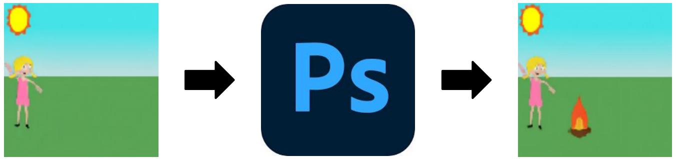
- **Visual editing applications** (Photoshop/Premiere) are widely used but **difficult for novices**



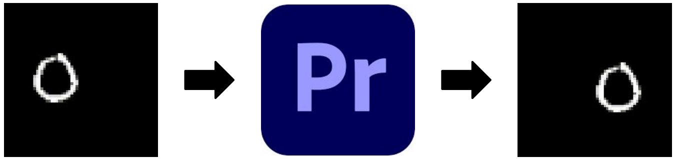
Visual Editing using Natural Language

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

- People can **edit directly using language** and improve accessibility



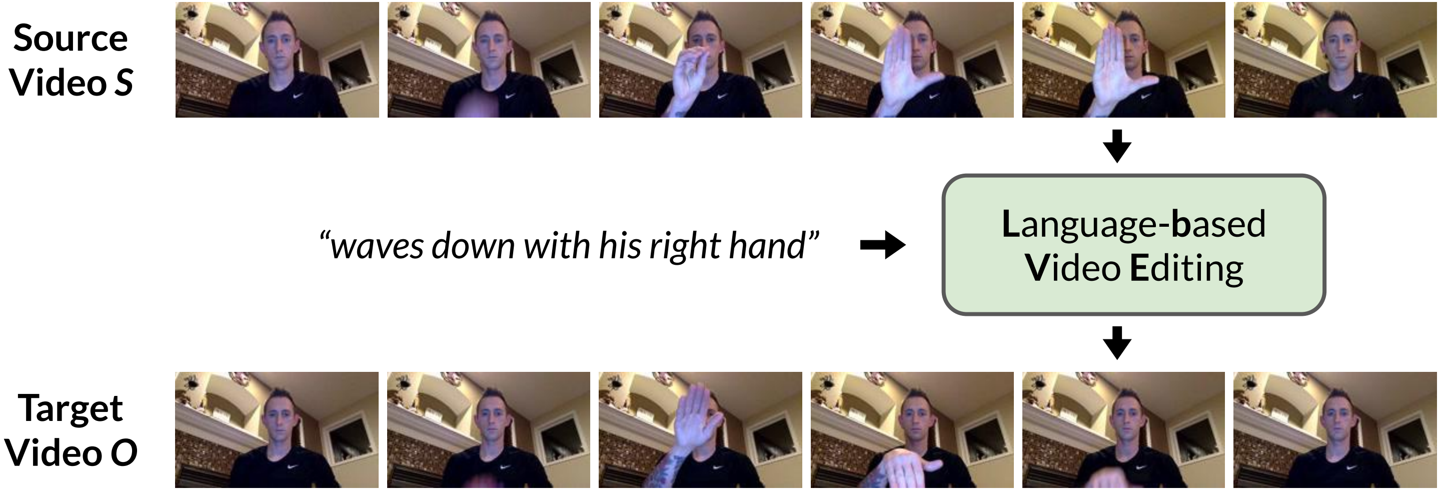
*“a **fire** is on **front** feet of **girl**”*



*“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*

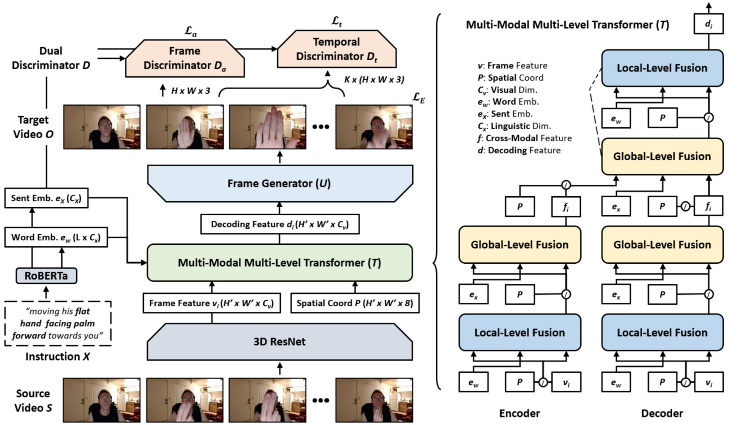
"waves down with his right hand"

Language-**b**ased **V**ideo **E**diting

Target Video O

Multi-Modal Multi-Level Transformer (M^3L)

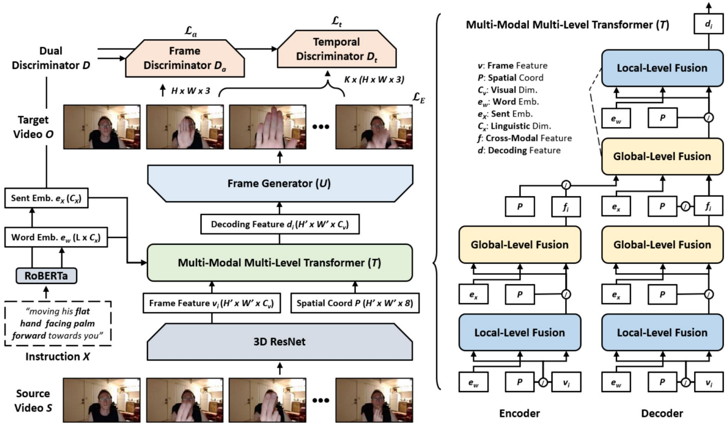
- **Input:** **Source** $S=\{s_1, s_2, \ldots, s_N\}$, **Instruction** X
- **Output:** **Target** $O=\{o_1, o_2, \ldots, o_N\}$



Multi-Modal Multi-Level Transformer (M^3L)

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

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

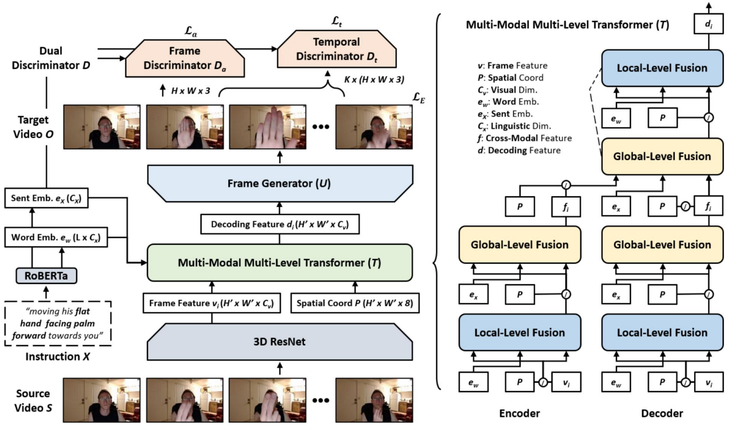


Multi-Modal Multi-Level Transformer (M^{3}L)

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

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

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



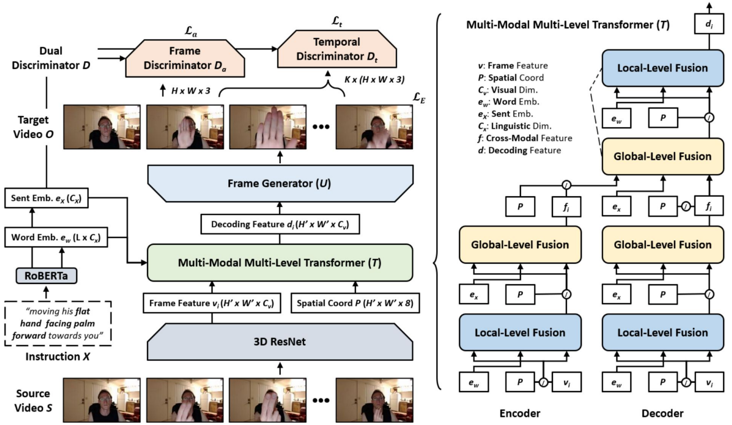
Multi-Modal Multi-Level Transformer (M^{3}L)

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

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

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

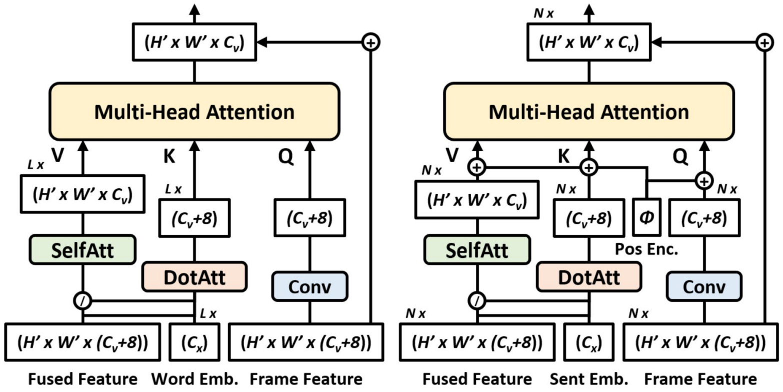
- **Frame Generation:** $o_i = U(d_i)$



Multi-Level Fusion

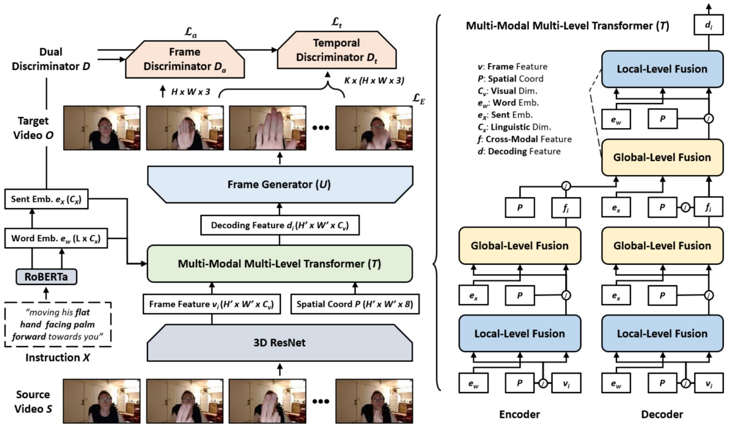
- Both video and language are **multi-level conveyed**

- Follow **multi-head attention (MHA)**
  - **Local-level Fusion (LF)**: single frame ↔ word token
  - **Global-level Fusion (GF)**: video sequence ↔ whole instruction



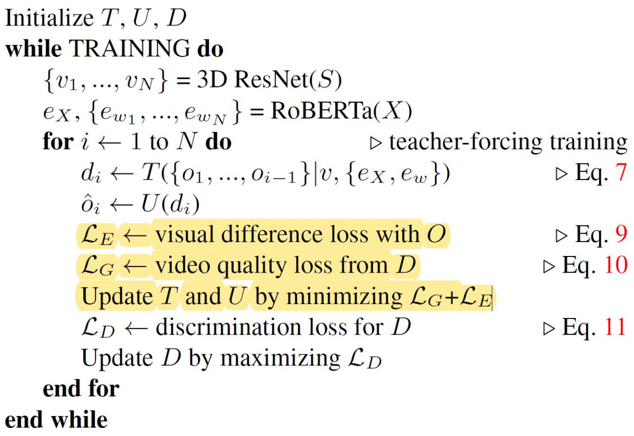
(a) Local-Level Fusion (LF)

(b) Global-Level Fusion (GF)

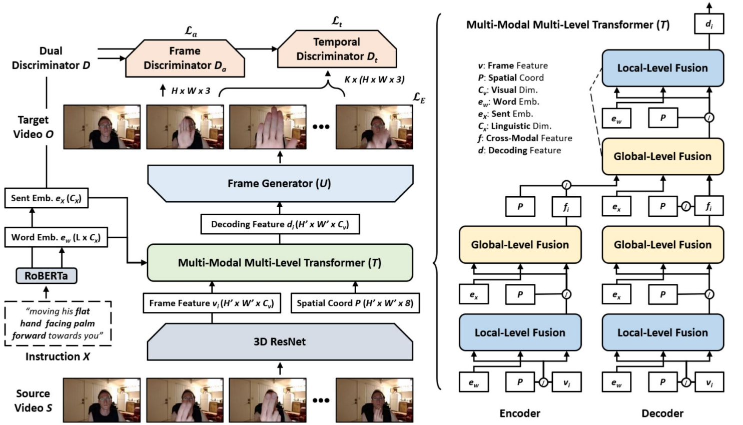


Learning of M^{3}L

- **Editing Loss *L*$_E$**: 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}}))

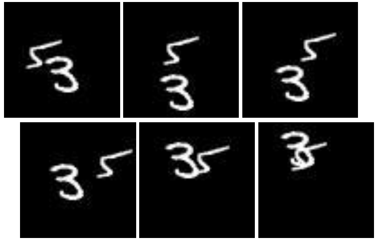


```
Initialize $T, U, D$
while TRAINING do
    $\{v_1, ..., v_N\}$ = 3D ResNet($S$)
    $e_X, \{e_{w_1}, ..., e_{w_N}\}$ = RoBERTa($X$)
    for $i \leftarrow 1$ to $N$ do                     ▷ teacher-forcing training
        $d_i \leftarrow T(\{o_1, ..., o_{i-1}\}|v, \{e_X, e_w\})$        ▷ Eq. 7
        $\hat{o}_i \leftarrow U(d_i)$
        $\mathcal{L}_E \leftarrow$ visual difference loss with $O$          ▷ Eq. 9
        $\mathcal{L}_G \leftarrow$ video quality loss from $D$            ▷ Eq. 10
        Update $T$ and $U$ by minimizing $\mathcal{L}_G$+$\mathcal{L}_E$
        $\mathcal{L}_D \leftarrow$ discrimination loss for $D$            ▷ Eq. 11
        Update $D$ by maximizing $\mathcal{L}_D$
    end for
end while
```

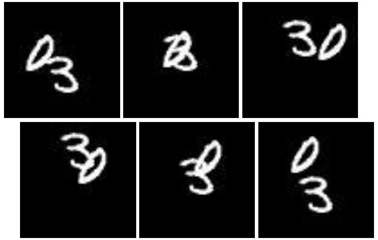


Dataset

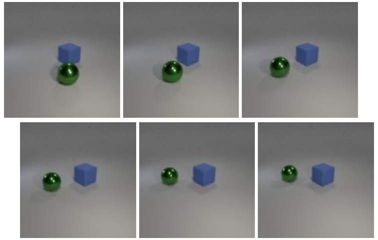
M-MNIST



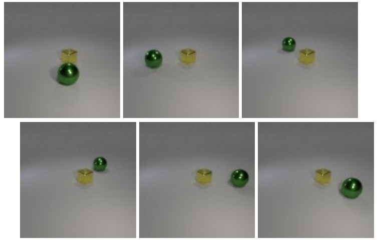
*“change the **direction from lower left to upper right** and the **number from 5 to 0**”*



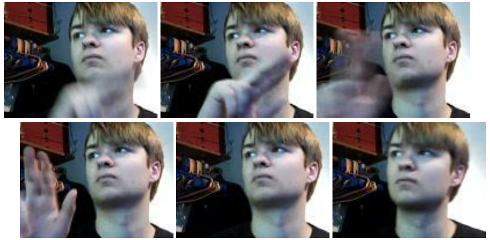
M-CLEVR



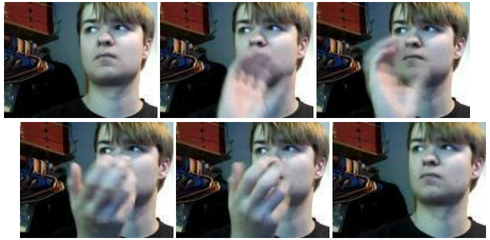
*“move to the **right front** and change the **large blue rubber into the small yellow metal**”*



E-JESTER



*“makes a **cup gesture** and **turns his hand in a circle**”*



Experiments

- Collected Dataset

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

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^{3}L: incorporate **multi-level fusion** to achieve the best performance

	M-MNIST			M-CLEVR			E-JESTER	
Method	**VAD ↓**	**OA ↑**	**mIoU ↑**	**VAD ↓**	**OA ↑**	**mIoU ↑**	**VAD ↓**	**GA ↑**
pix2pix	3.05	87.7	64.1	2.84	80.4	60.5	2.00	8.6
vid2vid	2.30	87.5	77.9	2.21	80.5	69.3	1.62	82.0
E3D-LSTM	2.10	90.4	81.3	2.11	83.1	72.2	1.55	83.6
M^{3}L	**1.90**	**93.2**	**84.7**	**1.96**	**84.5**	**78.4**	**1.44**	**89.3**

Experiments

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

Ablation Settings		E-JESTER	
Instruction	MLF	VAD ↓	GA ↑
✘	✘	1.99	4.7
✔	✘	1.50	85.4
✔	✔	**1.44**	**89.3**

Experiments

- Ablation Study

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

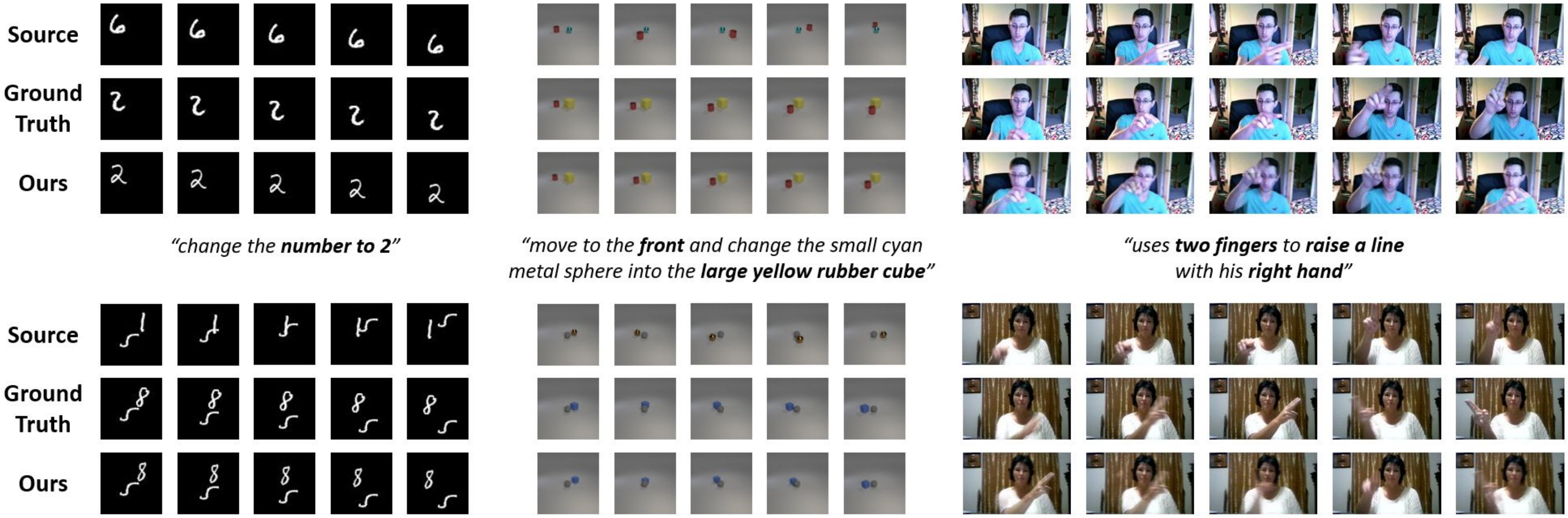
Ablation	M-MNIST			M-CLEVR		
MLF	VAD ↓	OA ↑	mIoU ↑	VAD ↓	OA ↑	mIoU ↑
✘	2.64	82.6	73.6	2.32	70.1	66.6
✔	**2.35**	**87.5**	**79.1**	**2.29**	**76.7**	**71.5**

Experiments

- Ablation Study
- Zero-shot Generalization
- Human Evaluation

	w/ MLF	w/o MLF	Tie
Video Quality	**67.1%**	27.1%	5.8%
Video-Instruction Align.	**53.3%**	35.1%	11.6%
Simil. to GT Video	**59.6%**	28.9%	11.6%

Qualitative Examples



*“change the **number to 2**”*

*“move to the **front** and change the small cyan metal sphere into the **large yellow rubber cube**”*

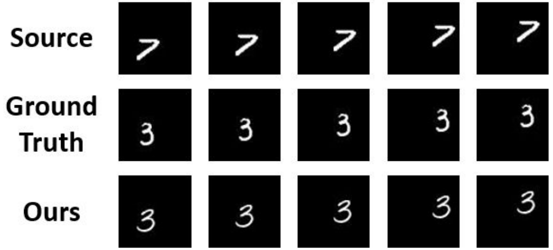
*“uses **two fingers** to **raise a line** with his **right hand**”*

*“change the **direction from upper right to lower right** and the **number from 1 to 8**”*

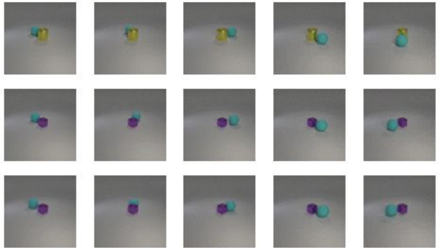
*“change the brown metal sphere into the **blue rubber cube** and move it to the **left**”*

*“motions her **right hand** from **left to right** while **showing two fingers**”*

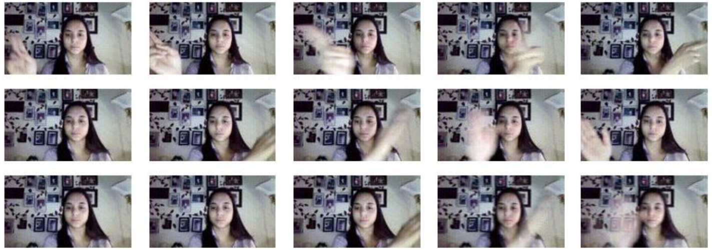
Qualitative Examples



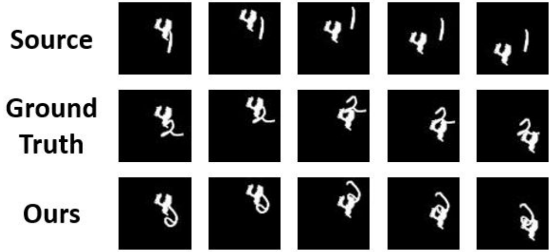
*“change the **number to 3**”*



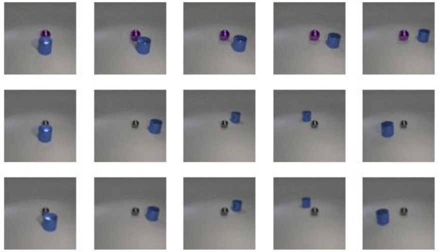
*“move to the **left front** and change the large yellow cylinder into the **small purple cube**”*



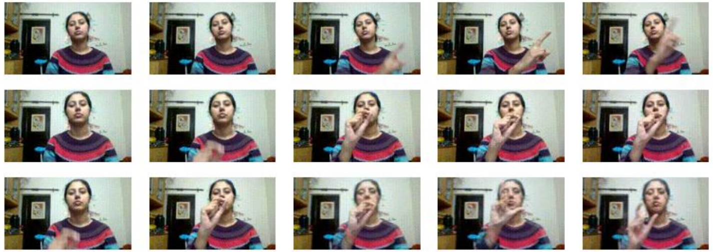
*“**rotates** and **swipes** her **right hand from left to right**”*



“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**”