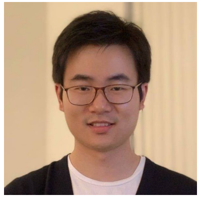
Animatable version:
https://tsujuifu.github.io/slides/cvpr23_tvc.ppsx

Tell Me What Happened: Unifying Text-guided Video Completion via Multimodal Masked Video Generation

Tsu-Jui Fu[1]



Licheng Yu[2]

Ning Zhang[2]

Cheng-Yang Fu[2]

Jong-Chyi Su[3]

William Wang[1]

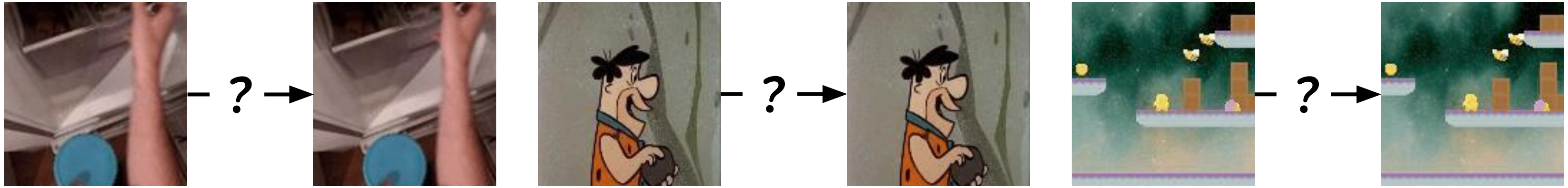
Sean Bell[2]

[1]UC Santa Barbara, [2]Meta, [3]NEC Lab

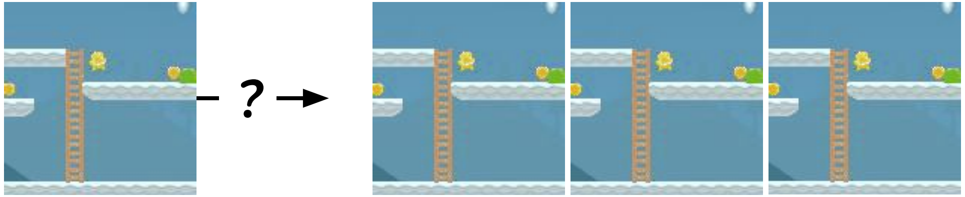
https://tsujuifu.github.io

Video Prediction

- Generate future video frames, **given the past**
  - Maintain **reasonable continuation** and **temporal coherence**
  - Wide range of video applications (*e.g.*, **compression** / **autonomous** / **VR**)

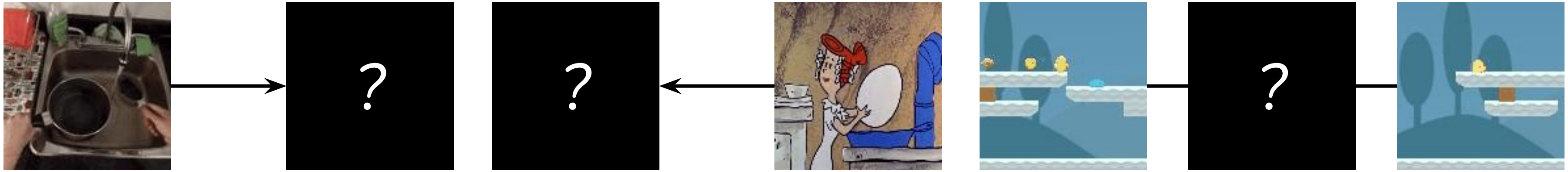


- **Uncontrollable**: there can be different outcomes



Video Completion

- **Not just the first frame** can guide
  - Complete a video from **partial frames at arbitrary time points**
  - **Prediction** (first) / **Rewind** (last) / **Infilling** (head-tail)

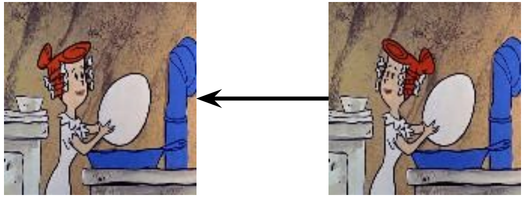


Text-guided Video Completion

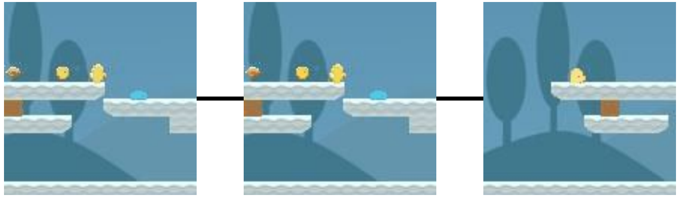
- **Not just the first frame** can guide
  - Complete a video from **partial frames at arbitrary time points**
  - **Prediction** (first) / **Rewind** (last) / **Infilling** (head-tail)

- Use language to **describe missing event**

pour water in pot



Wilma turns her head then she speaks.

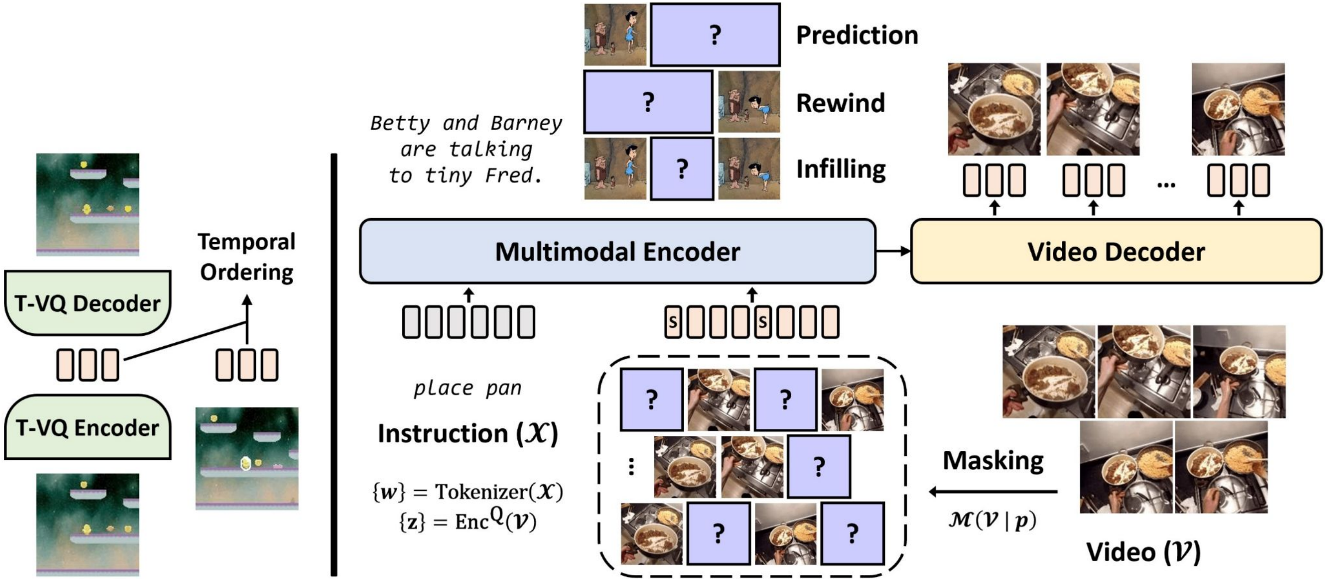


Mugen runs to left. Then collects a coin and a gem.

Multimodal Masked Video Generation (MMVG)

- For **training**, we have full video (V) and caption (X)
- During **inference**, **only partial V and X** are provided

- **Temporal-aware** discrete video representation
- **Mask-then-fill learning** to unify video completion


T-VQ Decoder
Temporal Ordering
T-VQ Encoder
Betty and Barney are talking to tiny Fred.
?
Prediction
Rewind
Infilling
Multimodal Encoder
Video Decoder
...
S
place pan
Instruction ($\mathcal{X}$)
$\{w\} = \text{Tokenizer}(\mathcal{X})$
$\{z\} = \text{Enc}^{Q}(\mathcal{V})$
Masking
$\mathcal{M}(\mathcal{V} \mid p)$
Video ($\mathcal{V}$)

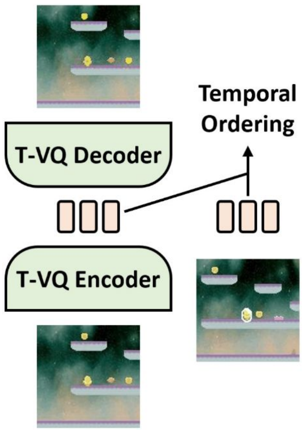
Temporal-aware (T-VQ)

- Built upon standard training of vector quantization (VQ)
  - *v*’ = Dec^{Q}(q(Enc^{Q}(*v*) | C))

- **Inject temporal relationship** into discrete tokens (z)
  - Contrastive **temporal ordering**
  - Lead to **smooth video modeling**

$$\begin{aligned}
\mathrm{z}_i &= \mathrm{q}(\mathrm{Enc}^{\mathrm{Q}}(v_i) \mid C) \\
\hat{v}_i &= \mathrm{Dec}^{\mathrm{Q}}(\mathrm{z}_i) \\
\mathcal{L}_{\mathrm{VQ}} &= \underbrace{||\hat{v}_i - v_i||_1}_{\text{reconstrcution}} + \underbrace{||\mathrm{sg}[\mathrm{Enc}^{\mathrm{Q}}(v_i)] - C_{\mathrm{z}_i}||_2^2}_{\text{codebook}} \\
&\quad + \underbrace{\beta||\mathrm{sg}[C_{\mathrm{z}_i}] - \mathrm{Enc}^{\mathrm{Q}}(v_i)||_2^2}_{\text{commit}} + \underbrace{||\mathcal{F}(\hat{v}_i) - \mathcal{F}(v_i)||_1}_{\text{matching}}
\end{aligned}$$

$$\begin{aligned}
o_i &= \mathrm{FC}^{\mathrm{T}}(\mathrm{z}_i, \mathrm{z}_j) \\
\mathcal{L}_{\mathrm{T}} &= \mathrm{BCELoss}(o_i, 0 \text{ if } i > j \text{ else } 1)
\end{aligned}$$



Generation from Masked Video

- Masking strategy (*M*) **masks out** V with probability (*p*)
  - **Replace each fragment** as unique [SPAN]
  - For example: *M* **reserves the 3rd and 5th** for 5-length

- Multimodal Encoder (Enc$^{\mathrm{M}}$) performs **cross-modal fusion**
  - **Tokenize** V via Enc$^{\mathrm{Q}}$ and X
  - **Extract** encoding feature (*h*) via Transformer
- Video Decoder (Dec$^{\mathrm{M}}$) produces **entire V**
  - **Auto-regressively decode** tokens upon *h*
  - **Reconstruct** all frames via Dec$^{\mathrm{Q}}$

$$\overline{\mathcal{V}}: \{[\mathrm{S}], v_3, [\mathrm{S}], v_5\} = \mathcal{M}(\mathcal{V} \mid p)$$

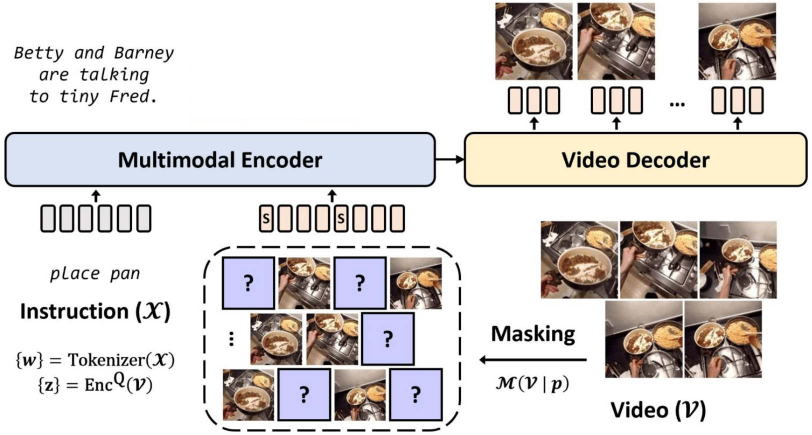
$$\{[\mathrm{S}], \mathrm{z}_3, [\mathrm{S}], \mathrm{z}_5\} \quad f_i^w, f_j^v = \mathrm{LP}^w(w_i), \mathrm{LP}^v(\mathrm{z}_j)$$

$$\{h\} = \mathrm{Enc}^{\mathrm{M}}([\{f^w\}, \{f^v\}])$$

$$\hat{\mathrm{z}_t} = \mathrm{Dec}^{\mathrm{M}}(\{\hat{\mathrm{z}_1}, ..., \hat{\mathrm{z}_{t-1}}\} \mid \{h\})$$

$$\mathcal{L}_t = \mathrm{CELoss}(\hat{\mathrm{z}_t}, \mathrm{z}_t)$$

$$\hat{\mathcal{V}} = \mathrm{Dec}^{\mathrm{Q}}(\{\hat{\mathrm{z}}\}_{t=1}^{N})$$

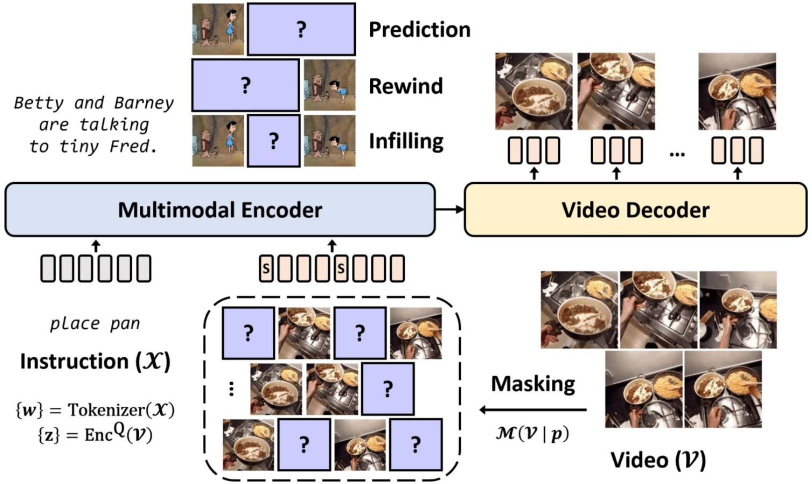

Betty and Barney are talking to tiny Fred.
Multimodal Encoder
Video Decoder
...
S
S
place pan
Instruction ($\mathcal{X}$)
$\{w\}$ = Tokenizer($\mathcal{X}$)
$\{\mathrm{z}\}$ = Enc$^{\mathrm{Q}}$($\mathcal{V}$)
?
?
?
?
?
Masking
$\mathcal{M}(\mathcal{V} \mid p)$
Video ($\mathcal{V}$)

Generation from Masked Video

- Masking strategy (*M*) **masks out** V with probability (*p*)
  - **Replace each fragment** as unique [SPAN]
  - For example: *M* **reserves the 3rd and 5th** for 5-length

- Multimodal Encoder (Enc^{M}) performs **cross-modal fusion**
  - **Tokenize** V via Enc^{Q} and X
  - **Extract** encoding feature (*h*) via Transformer
- Video Decoder (Dec^{M}) produces **entire V**
  - **Auto-regressively decode** tokens upon *h*
  - **Reconstruct** all frames via Dec^{Q}

- **Unifying video completion** during inference
  - **Prediction**: [{*w*}, {z_1, [S]}]
  - **Rewind**: [{*w*}, {[S], z_N}]
  - **Infilling**: [{*w*}, {z_1, [S], z_N}]


?
Prediction
Rewind
Infilling
Betty and Barney are talking to tiny Fred.
...
Multimodal Encoder
Video Decoder
S
S
place pan
Instruction ($\mathcal{X}$)
$\{w\} = \text{Tokenizer}(\mathcal{X})$
$\{z\} = \text{Enc}^{Q}(\mathcal{V})$
⋮
Masking
$\mathcal{M}(\mathcal{V} \mid p)$
Video ($\mathcal{V}$)

Experimental Setup

- **Datasets**
  - Kitchen / Flintstones / MUGEN

Resolution: 128^2

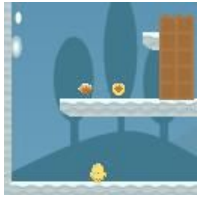
Dataset	Domain	#Frame	#Word	FPS
Kitchen	Egocentric	8.3	2.8	6
Flintstones	Animation	15	16.5	5
MUGEN	Gaming	16	20.6	5

pick up meat

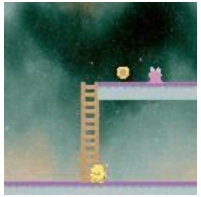
open fridge

Fred is driving. Then Barney talks to Fred.

Barney talks to Betty in a room.



Mugen jumps again. Move to the right to collect a gem.

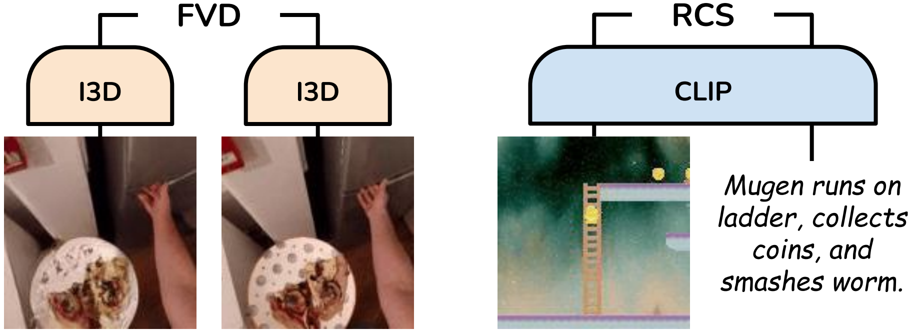


Mugen climbs up and jumps to collect the coin.

Experimental Setup

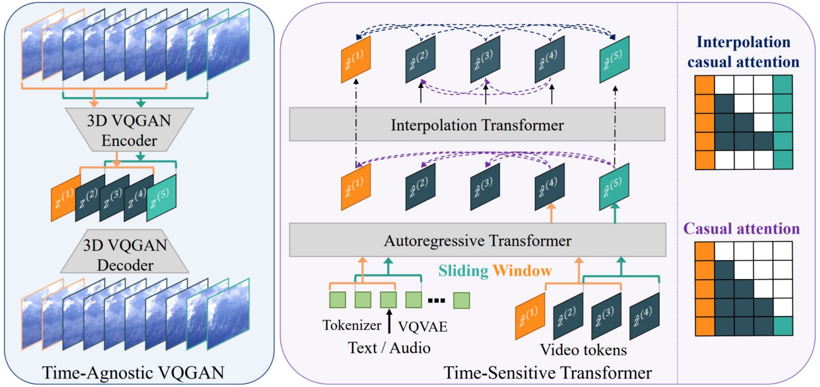
- **Datasets**
  - Kitchen / Flintstones / MUGEN

- **Evaluation Metrics**
  - **FVD (↓)**: distance of **I3D video feature** (*vs.* GT)
  - **RCS (↑)**: **relative visual-text similarity** from CLIP (*vs.* instruction)

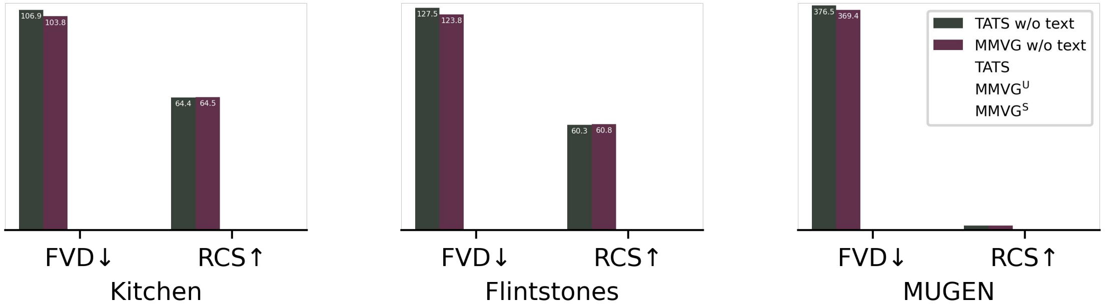


Experimental Setup

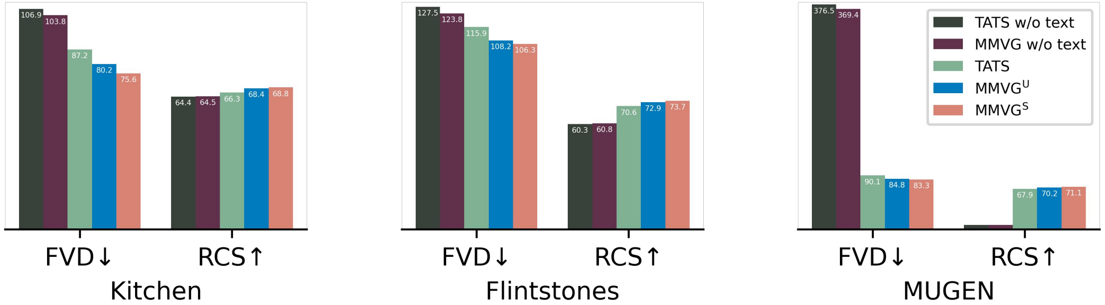
- **Datasets**
  - Kitchen / Flintstones / MUGEN
- **Evaluation Metrics**
  - **FVD (↓)**: distance of **I3D video feature** (*vs.* GT)
  - **RCS (↑)**: **relative visual-text similarity** from CLIP (*vs.* instruction)
- **Baselines**
  - **Auto-regressive VQ**: TATS (**requires specific training** for rewind / infilling)



Text-guided Video Prediction



Text-guided Video Prediction



place pan

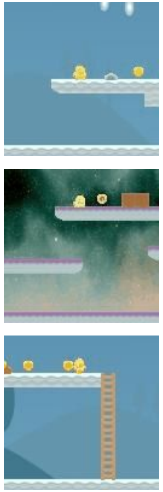
put down cloth

rinse plate

Betty quickly opens and closes a window blind.

Fred talks to Barney and turns his head.

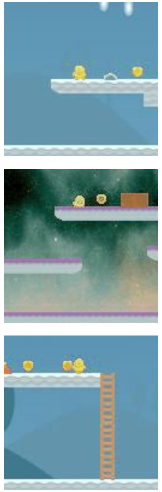
Wilma is in a car and talking.



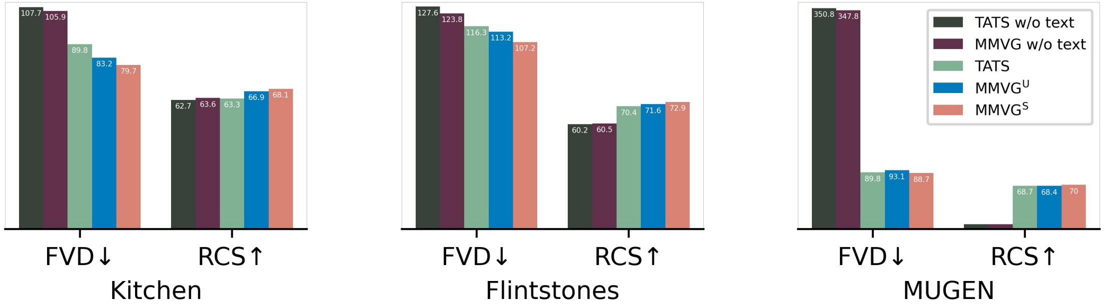
Mugen jumps over a gear and then collects coin.

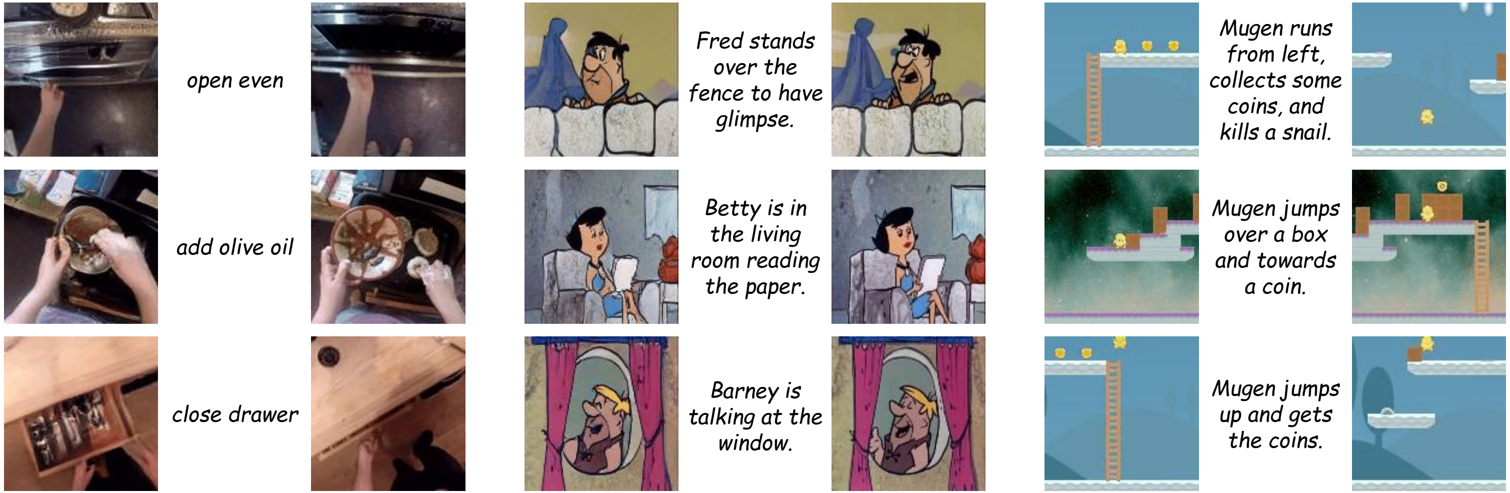
Mugen runs to right, jumps to hit face, and jumps down.

Mugen gets coins. It saw a ladybug so it runs to right.



Text-guided Video Rewind





open even

add olive oil

close drawer

Fred stands over the fence to have glimpse.

Betty is in the living room reading the paper.

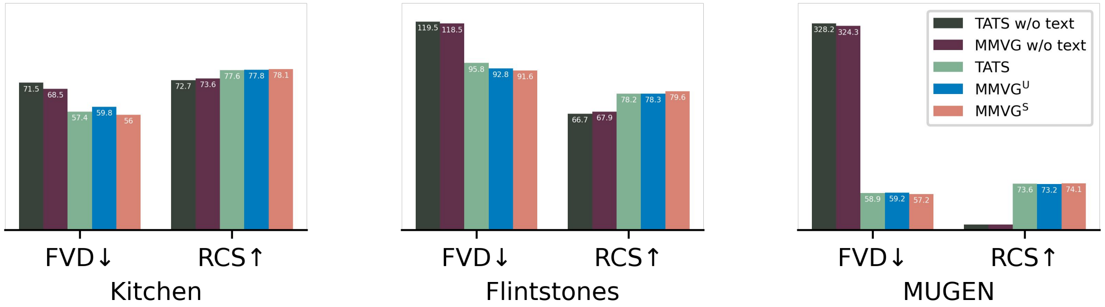
Barney is talking at the window.

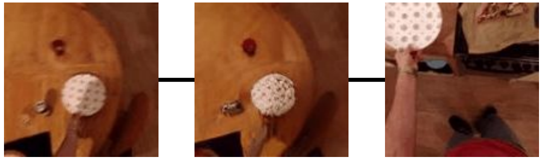
Mugen runs from left, collects some coins, and kills a snail.

Mugen jumps over a box and towards a coin.

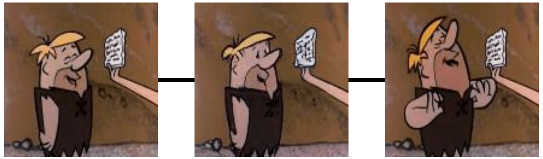
Mugen jumps up and gets the coins.

Text-guided Video Infilling

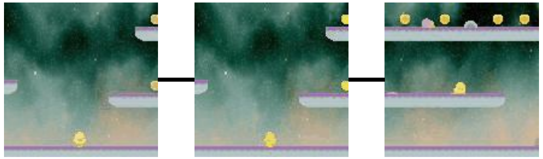




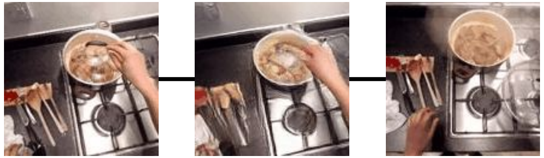
take plate



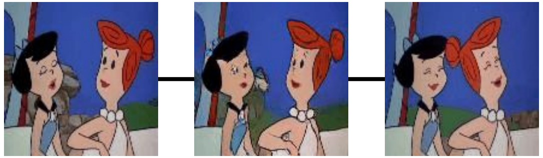
Barney is trying to sing in a living room.



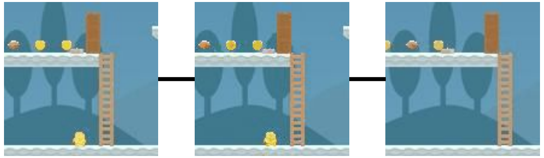
Mugen jumps up the stage. It runs from left to right and jumps on a worm.



take lid off



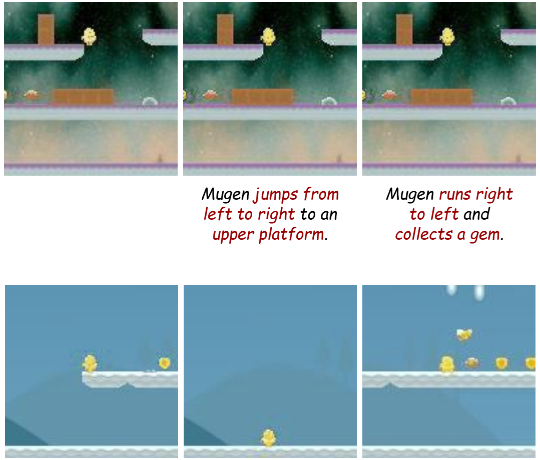
Wilma looks back at Betty who is speaking then she turns her head.



Mugen climbs up a ladder. It jumps onto a stack of boxes, drops down, and is killed by a worm.

Summary

- Text-guided Video Completion
  - **Control video generation from partial frames** via language command
- Multimodal Masked Video Generation (MMVG)
  - T-VQ for **temporal-aware** discrete frame tokens
  - **Unified masked training** to support all prediction / rewind / infilling
  - **Benefit general video generation** as well

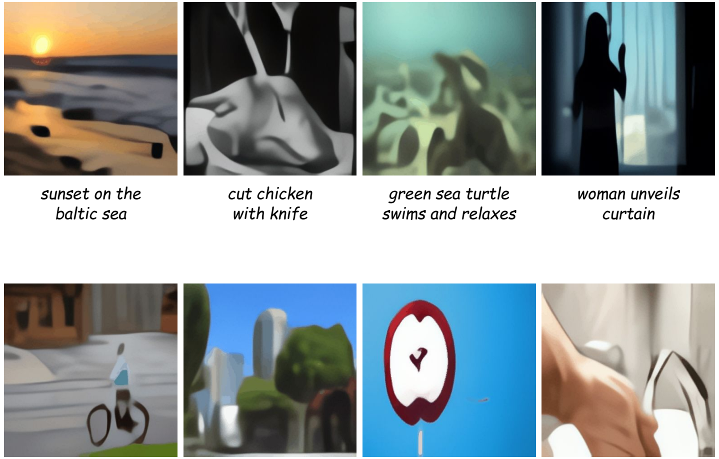


Mugen jumps from left to right to an upper platform.

Mugen runs right to left and collects a gem.

Mugen walks to the right and jumps over a mouse to collect coins.

Mugen walks to the right and jumps to collect coins.



sunset on the baltic sea

cut chicken with knife

green sea turtle swims and relaxes

woman unveils curtain

child rides a bike

downtown city with traffic car

rotates apple lollipop

wash hand