Generating Videos of Zero-Shot Compositions of Actions and Objects

Megha Nawhal[1,2], Mengyao Zhai[2], Andreas Lehrmann[1], Leonid Sigal[1,3,4], Greg Mori[1,2]

[1]Borealis AI, [2]Simon Fraser University, [3]University of British Columbia, [4]Vector Institute

Generating Human-Object Interaction (HOI) Videos

action: **take**
object: **spoon**

action: **remove**
object: **cup**

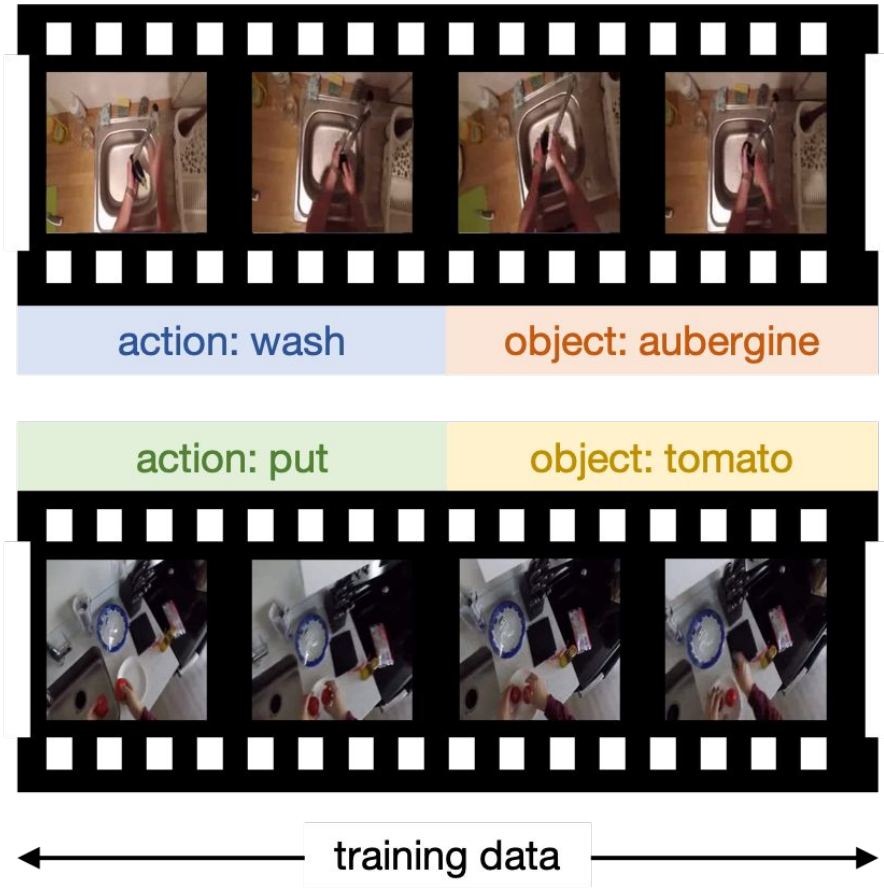
action: **move**
object: **book**

action: **put**
object: **banana**

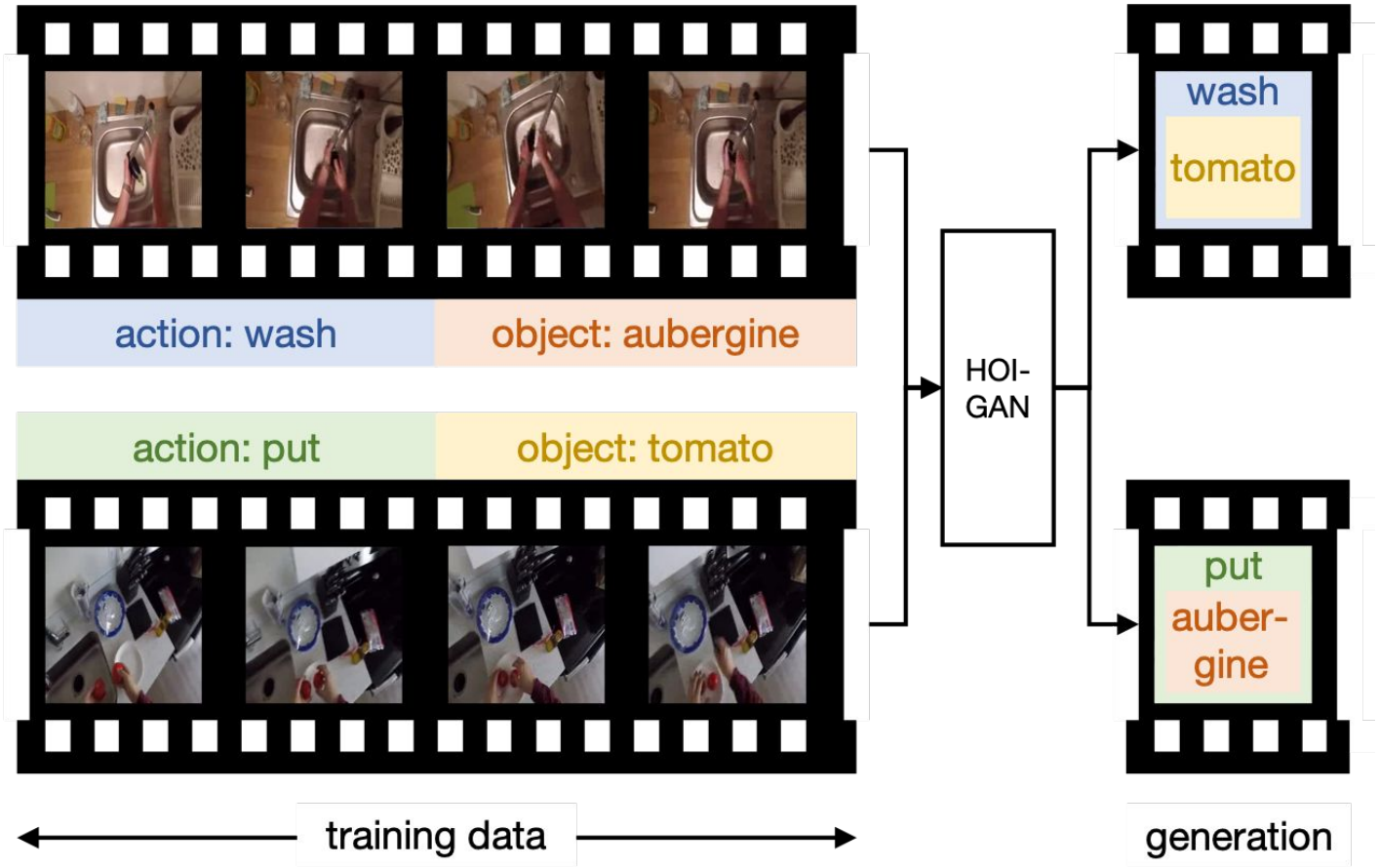
HOI videos as compositions of actions and objects

Zero-Shot Compositional Generation of HOI Videos

Zero-Shot Compositional Generation of HOI Videos



Zero-Shot Compositional Generation of HOI Videos



Challenges of HOI Video Generation

Challenges of HOI Video Generation

- Object content

target object: ***pizza***

Challenges of HOI Video Generation

- Object content
- Specified scene

target object: ***pizza***

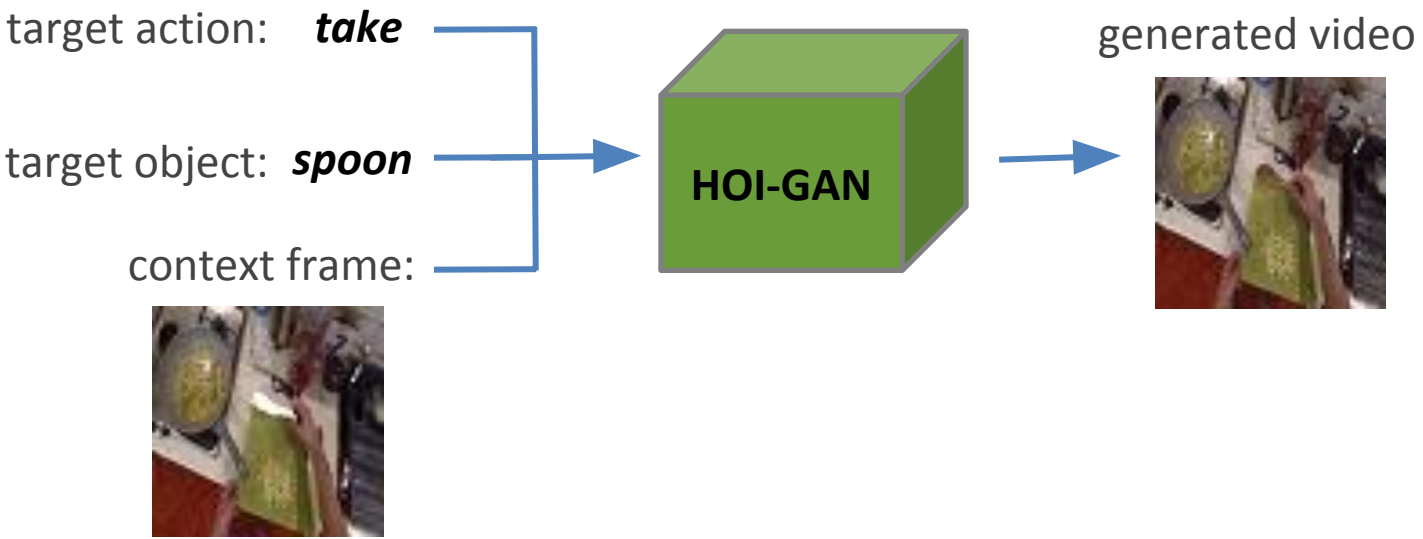
Challenges of HOI Video Generation

- Object content
- Specified scene
- Temporally consistent actions

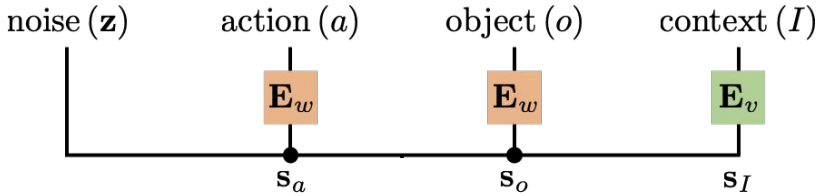
target object: ***pizza***
target action: ***take***

Our Solution

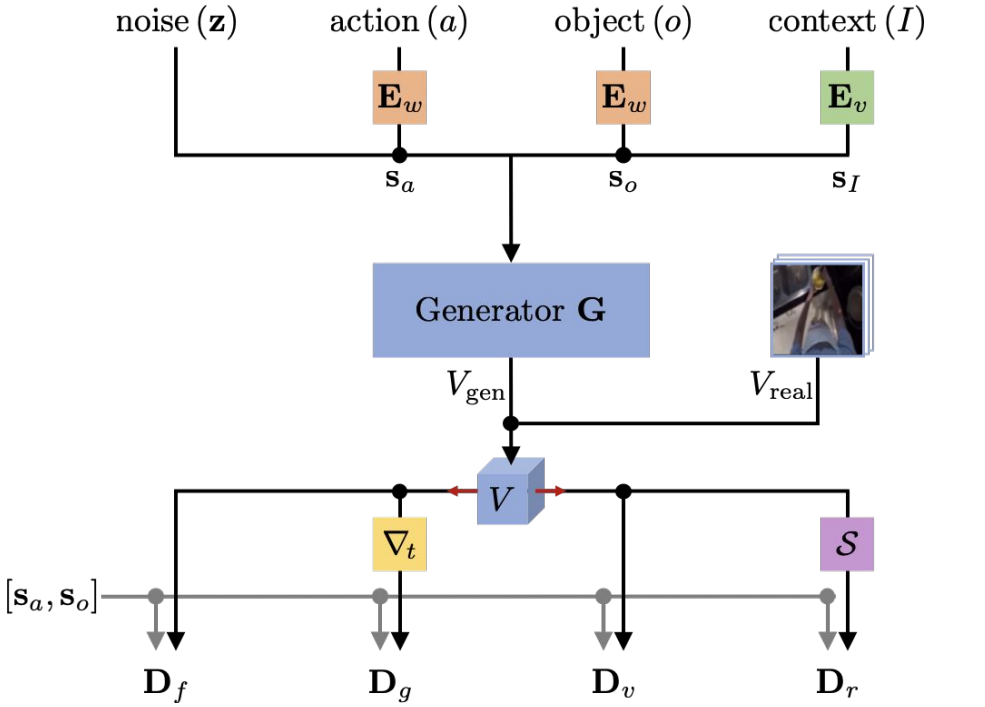
HOI-GAN generates a video depicting the target action on the target object using a context image (with an object mask) as background



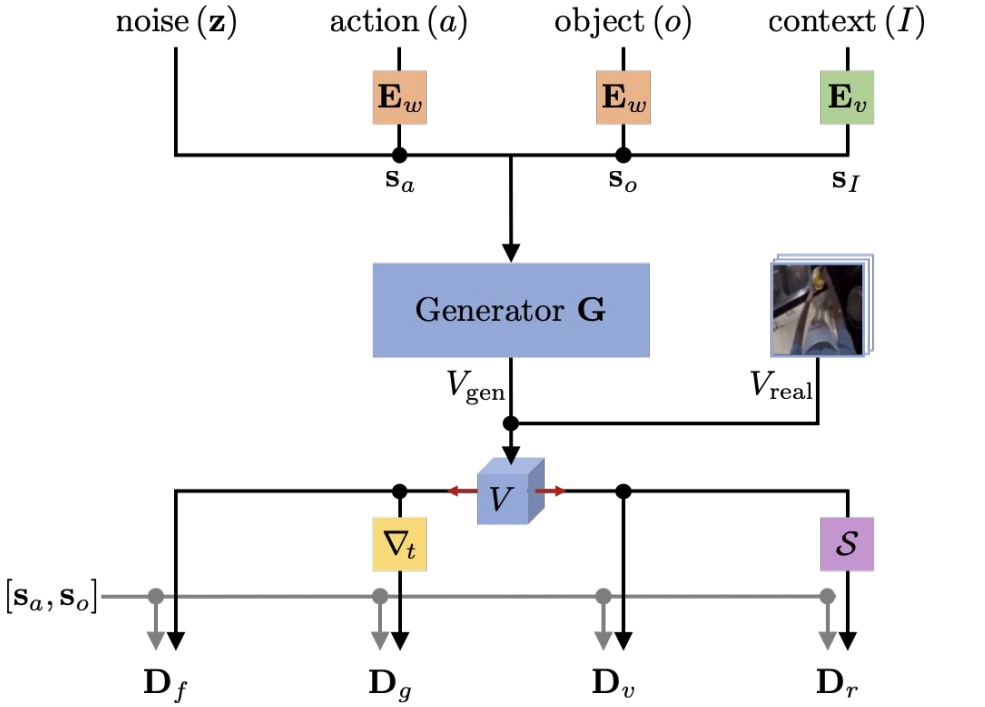
HOI-GAN



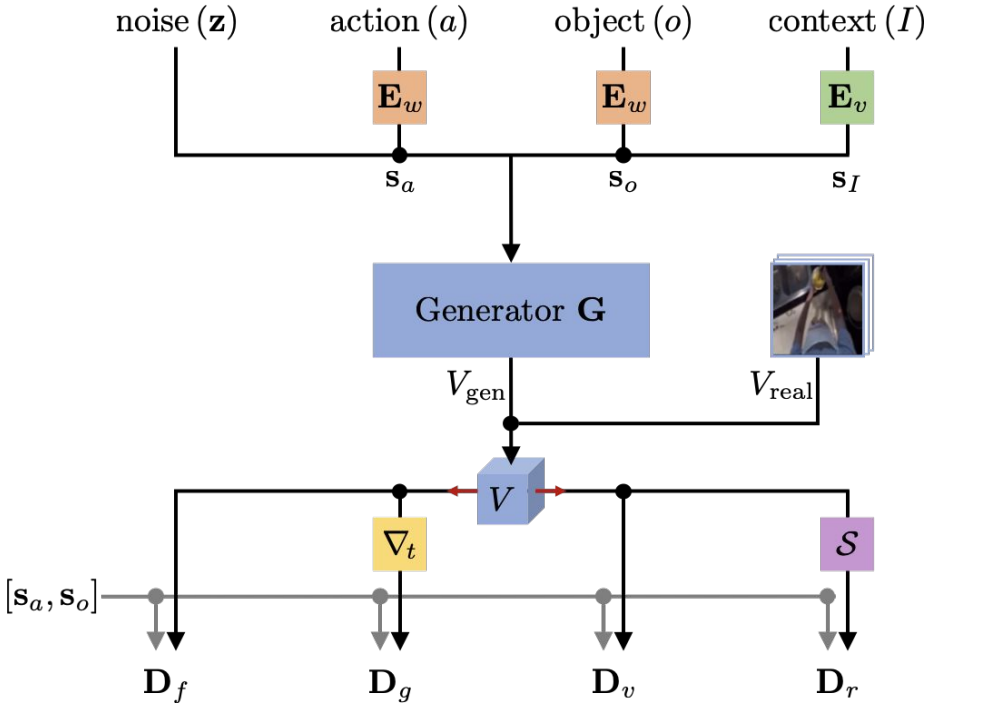
HOI-GAN


noise (z)
action (a)
object (o)
context (I)
$\mathbf{E}_w$
$\mathbf{E}_w$
$\mathbf{E}_v$
$\mathbf{s}_a$
$\mathbf{s}_o$
$\mathbf{s}_I$
Generator G
V_{gen}
V_{real}
V
∇_t
$\mathcal{S}$
$[\mathbf{s}_a, \mathbf{s}_o]$
$\mathbf{D}_f$
$\mathbf{D}_g$
$\mathbf{D}_v$
$\mathbf{D}_r$
Frame Discriminator
Gradient Discriminator
Video Discriminator
Relational Discriminator

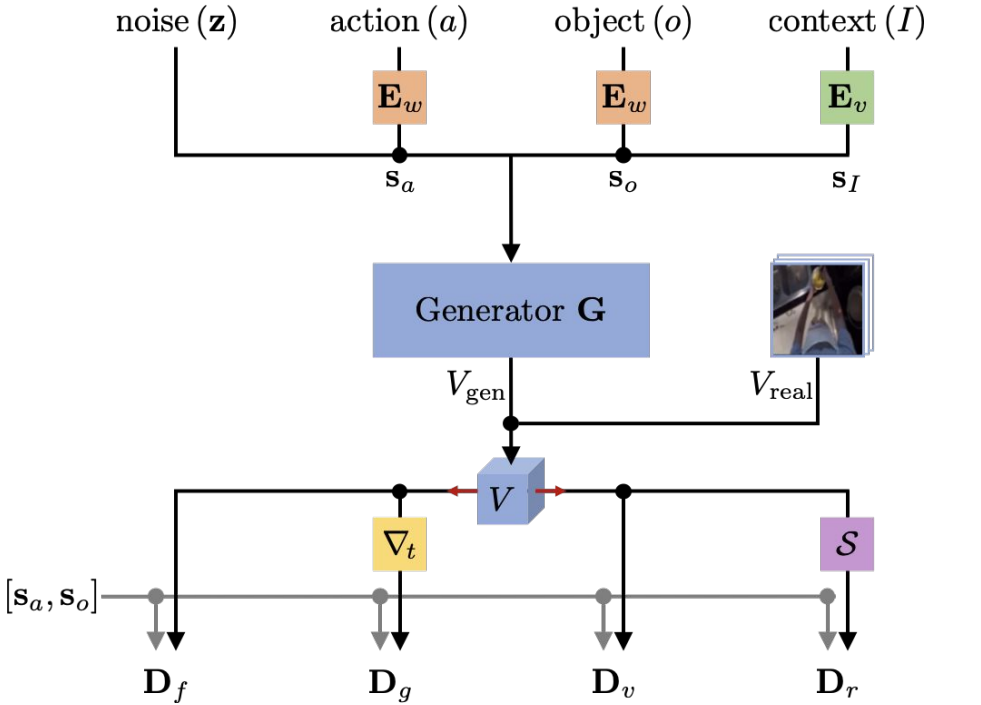
HOI-GAN


noise (z)
action (a)
object (o)
context (I)
E_w
E_w
E_v
s_a
s_o
s_I
Generator G
V_gen
V_real
V
∇_t
S
[s_a, s_o]
D_f
D_g
D_v
D_r
Frame Discriminator
Gradient Discriminator
Video Discriminator
Relational Discriminator

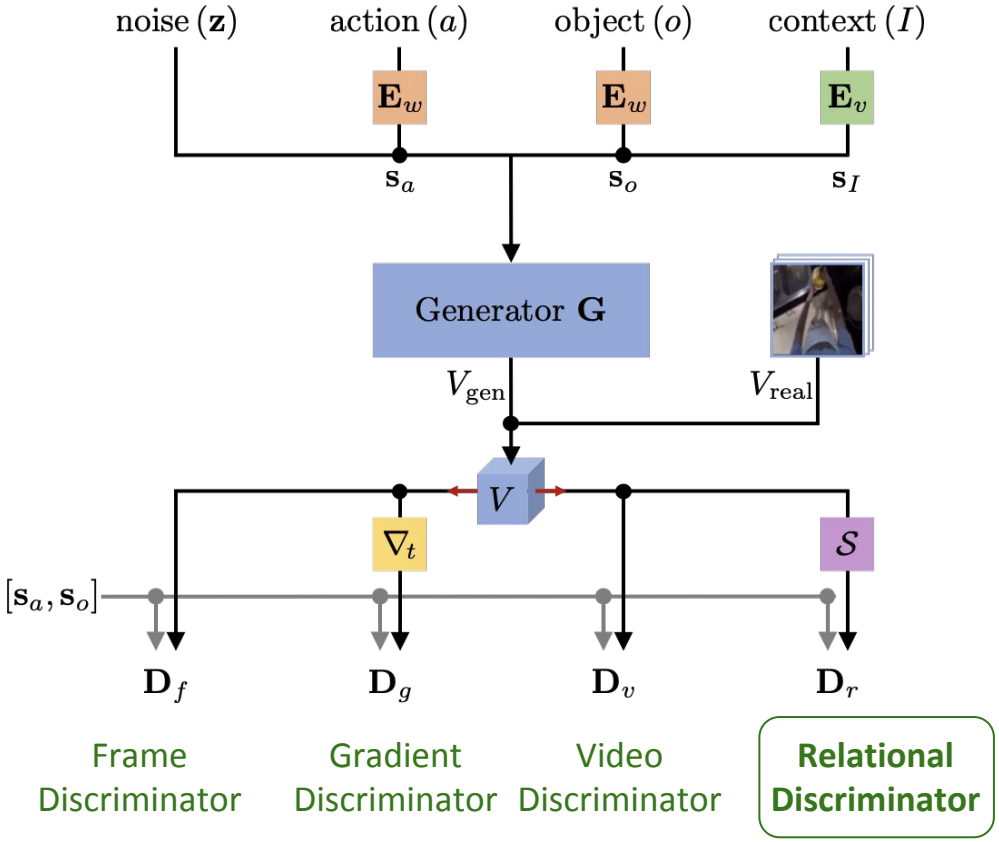
HOI-GAN


noise (z)
action (a)
object (o)
context (I)
E_w
E_w
E_v
s_a
s_o
s_I
Generator G
V_gen
V_real
V
∇_t
S
[s_a, s_o]
D_f
D_g
D_v
D_r
Frame Discriminator
Gradient Discriminator
Video Discriminator
Relational Discriminator

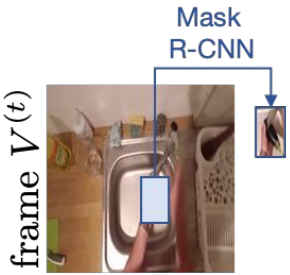
HOI-GAN


noise (z)
action (a)
object (o)
context (I)
E_w
E_w
E_v
s_a
s_o
s_I
Generator G
V_gen
V_real
V
∇_t
S
[s_a, s_o]
D_f
D_g
D_v
D_r
Frame Discriminator
Gradient Discriminator
Video Discriminator
Relational Discriminator

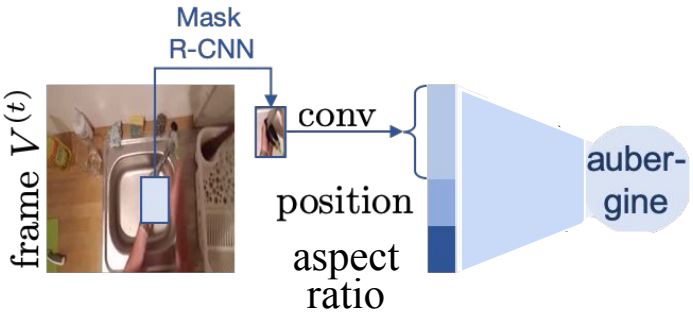
HOI-GAN


noise (z)
action (a)
object (o)
context (I)
$\mathbf{E}_w$
$\mathbf{E}_w$
$\mathbf{E}_v$
$\mathbf{s}_a$
$\mathbf{s}_o$
$\mathbf{s}_I$
Generator G
V_{gen}
V_{real}
V
∇_t
$\mathcal{S}$
$[\mathbf{s}_a, \mathbf{s}_o]$
$\mathbf{D}_f$
$\mathbf{D}_g$
$\mathbf{D}_v$
$\mathbf{D}_r$
Frame Discriminator
Gradient Discriminator
Video Discriminator
Relational Discriminator

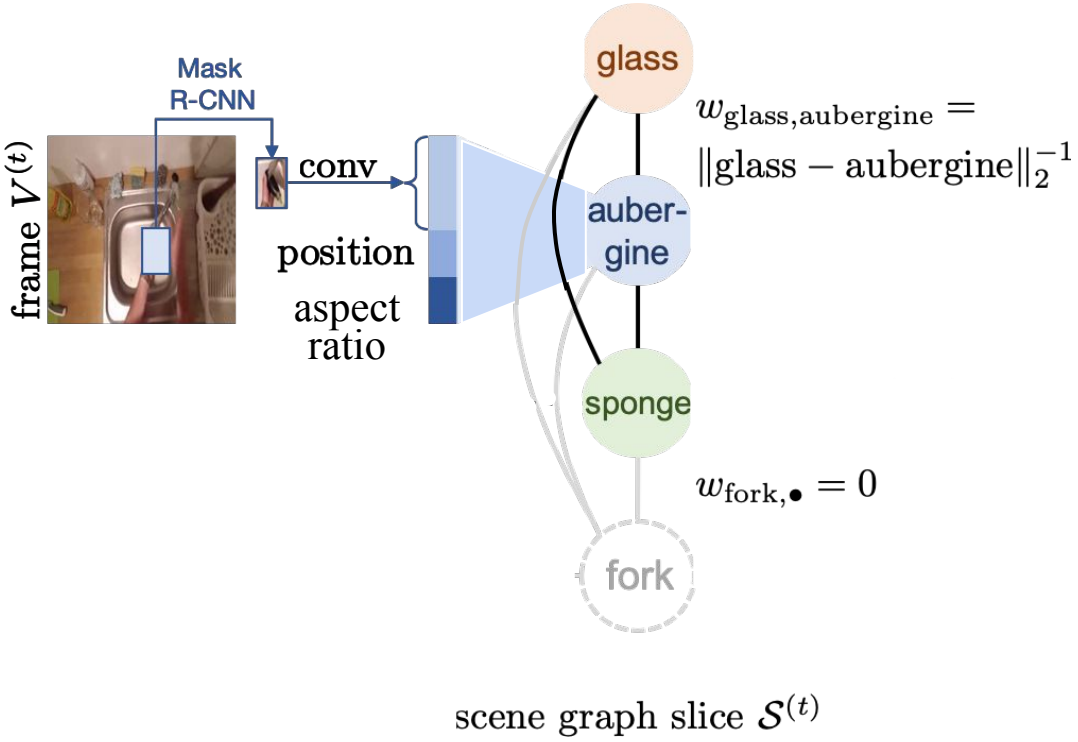
HOI-GAN: Relational Discriminator



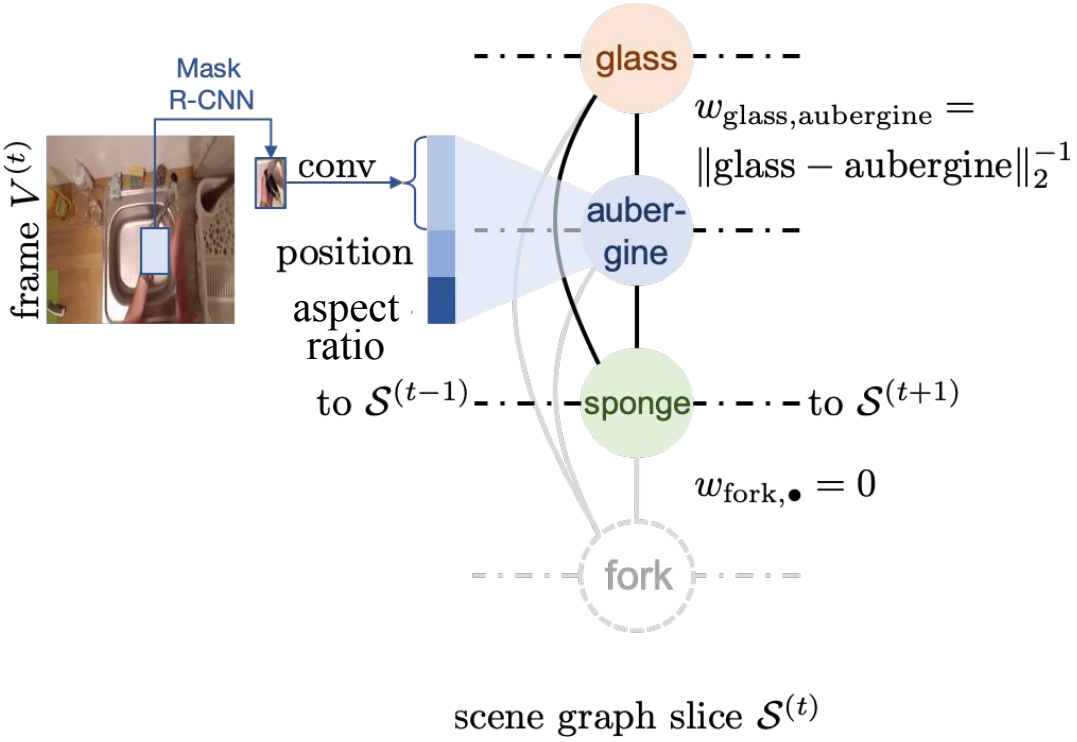
HOI-GAN: Relational Discriminator



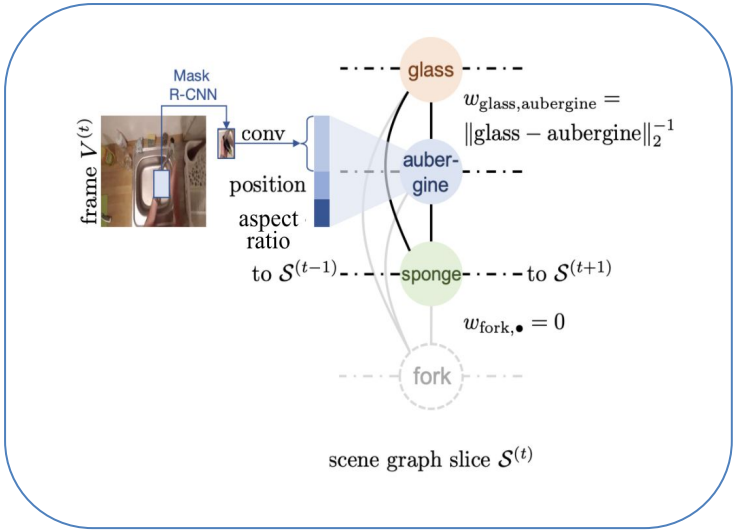
HOI-GAN: Relational Discriminator



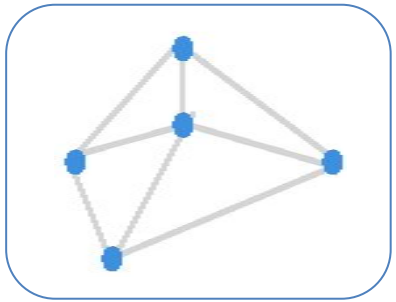
HOI-GAN: Relational Discriminator



HOI-GAN: Relational Discriminator

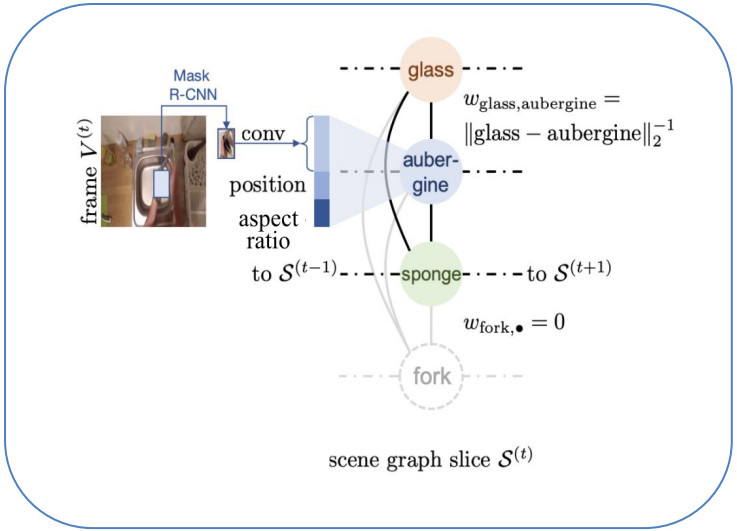


spatio-temporal scene graph

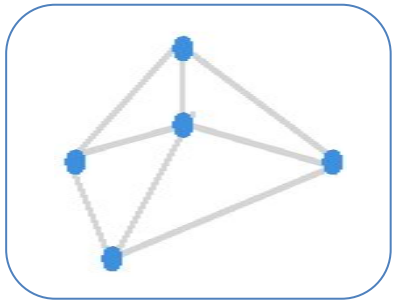


graph convolutions (GCN)

HOI-GAN: Relational Discriminator



spatio-temporal scene graph



graph convolutions (GCN)

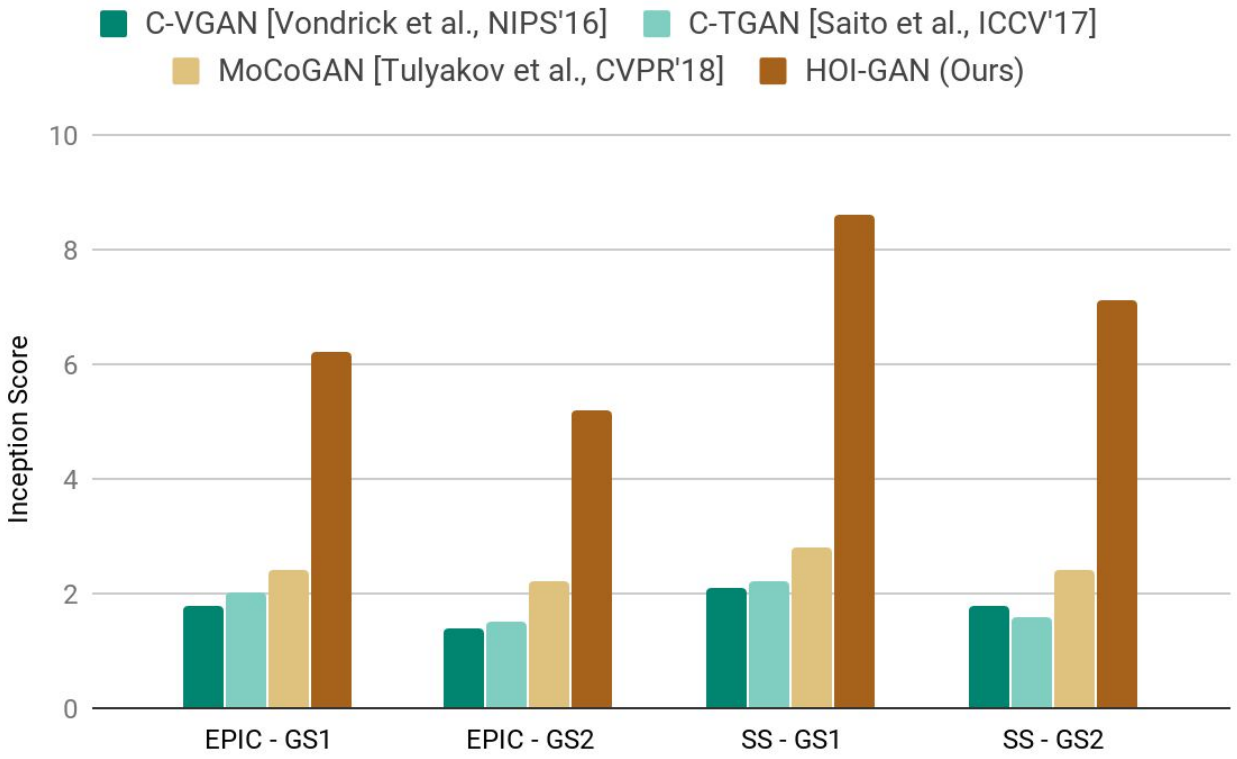
pooling

adversarial loss computation

Evaluation

- Large scale HOI video datasets
  - EPIC-Kitchens *[Damen et al., ECCV'18]*
  - 20BN-Something-Something V2 *[Goyal et al., ICCV'17]*

Quantitative Evaluation: Inception Score



Videos generated using HOI-GAN

	context frame	generated video
move book		
put banana		
remove cup		

Videos generated using HOI-GAN

	context frame	generated video
lift apple		
cut carrot		
turn vase		

Videos generated using HOI-GAN

context frame:

generated video - target action: ***lift***

handbag

apple

banana

scissors

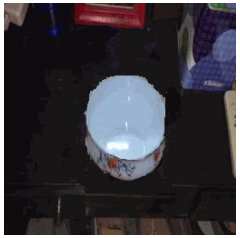
mouse

spoon

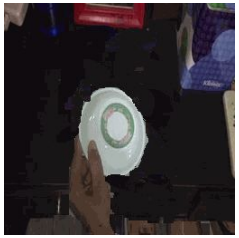
Videos generated using HOI-GAN

context frame:

generated video - target object: ***bowl***



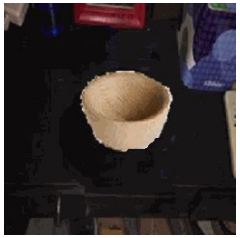
hold



move

push

put

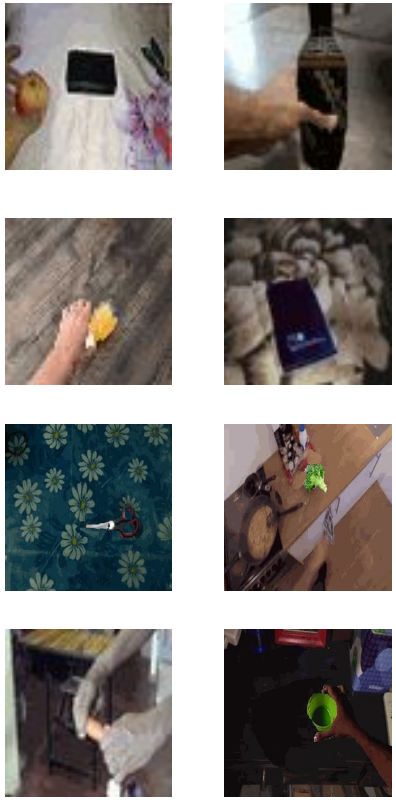


remove

throw

Summary

- Zero-shot compositional generation of human-object interaction videos
- HOI-GAN framework using pixel-level and object-level information in videos
- Generation of realistic videos depicting diverse actions on objects



Thank You!

Generating Videos of Zero-Shot Compositions of Actions and Objects

Megha Nawhal, Mengyao Zhai, Andreas Lehrmann,
Leonid Sigal, Greg Mori

http://www.sfu.ca/~mnawhal/projects/zs_hoi_generation.html