Interactive 3D Vision & Learning Lab (IVL), Brown University
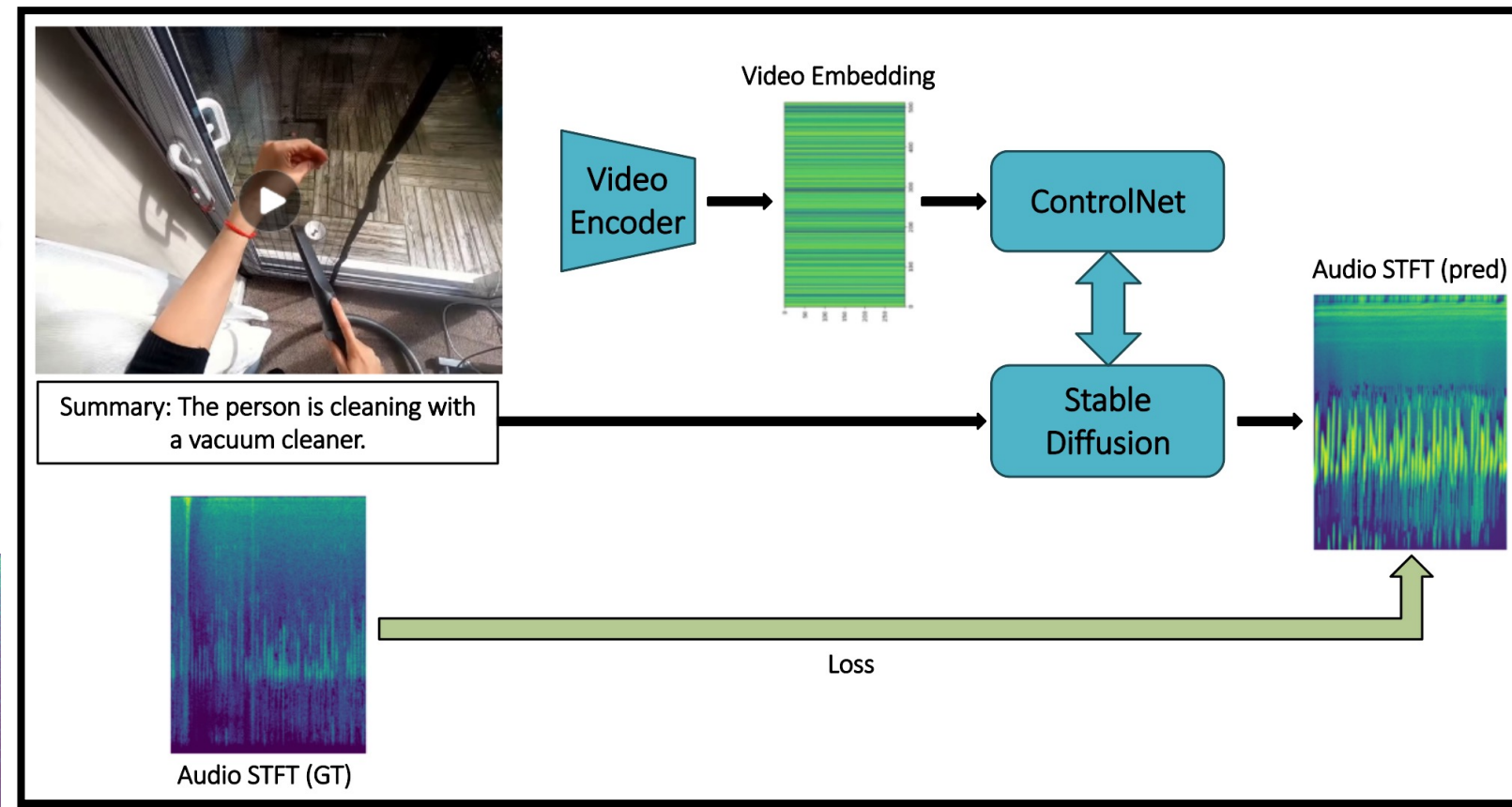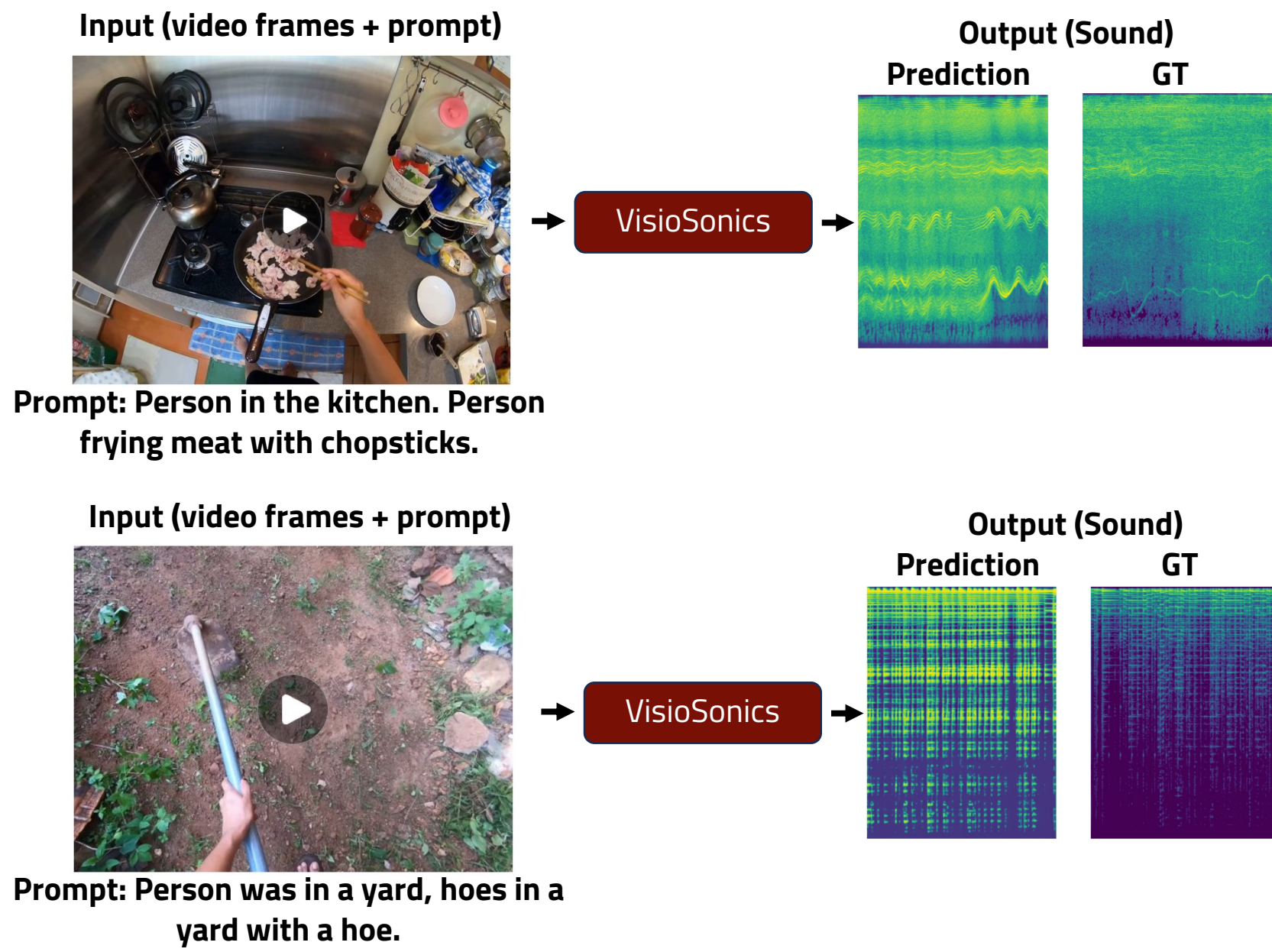
BROWN

BVC

VisioSonics: Bridging Vision and Audio

Aashish Rai[1] and Srinath Sridhar[1]
[1]Brown University
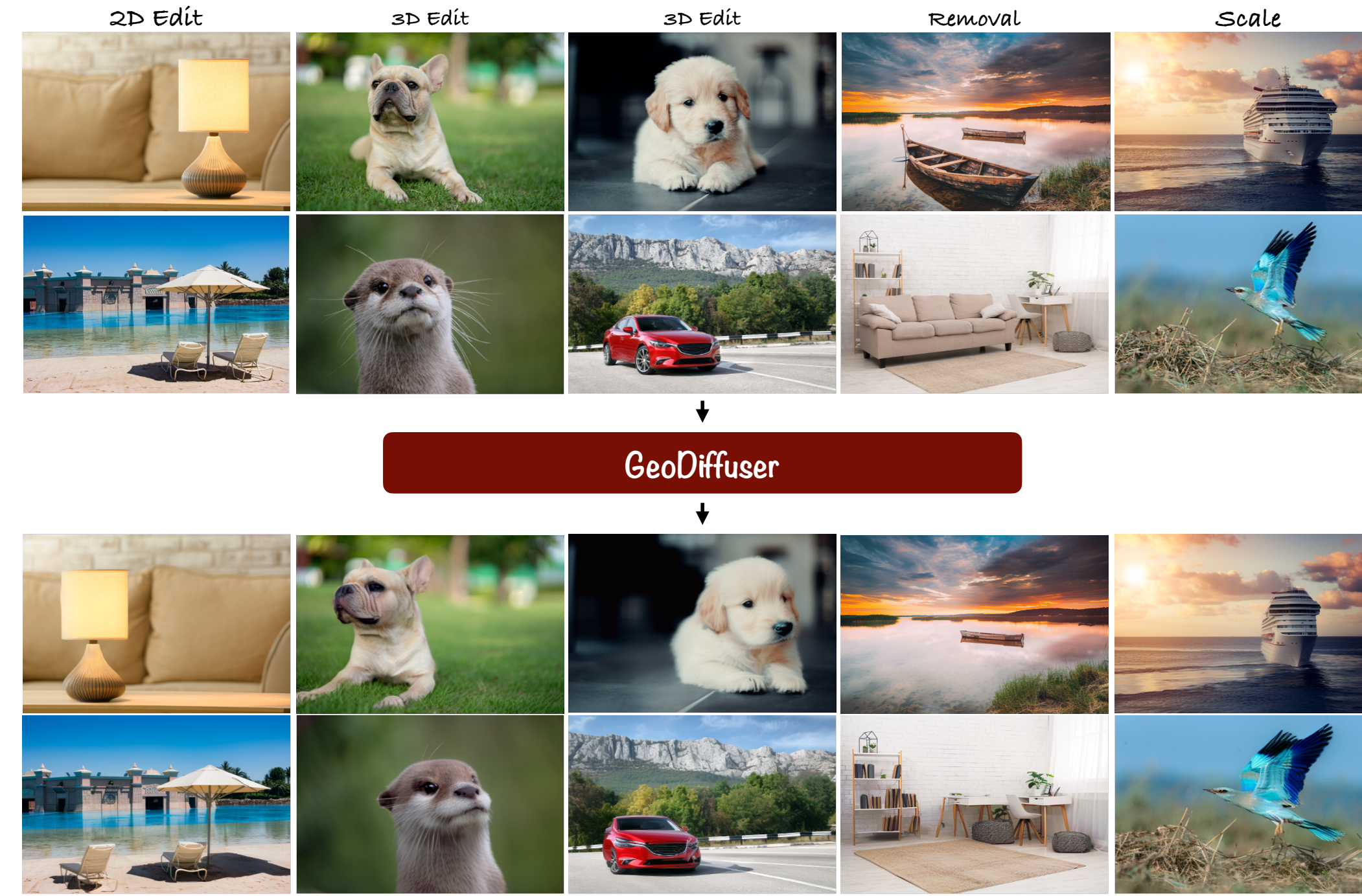
Given a muted video (image sequences) and a short summary of the scene (prompt), our method can predict the corresponding synchronized audio.

Input (video frames + prompt)

Output (Sound)
Prediction   GT

Prompt: Person in the kitchen. Person frying meat with chopsticks.

Input (video frames + prompt)

Output (Sound)
Prediction   GT

Prompt: Person was in a yard, hoes in a yard with a hoe.

Video Embedding

Video Encoder → ControlNet

Summary: The person is cleaning with a vacuum cleaner.

Stable Diffusion → Audio STFT (pred)

Audio STFT (GT)

Loss

VisioSonics leverages CLIP Encoders, Stable Diffusion, and ControlNet to generate video conditioned audio spectrograms.

GeoDiffuser: Geometry-Based Image Editing with Diffusion Models

Rahul Sajnani[1,2], Jeroen Vanbaar[2], Jie Min[2], Kapil Katyal[2], and Srinath Sridhar[1,2]
[1]Brown University and [2]Amazon Robotics

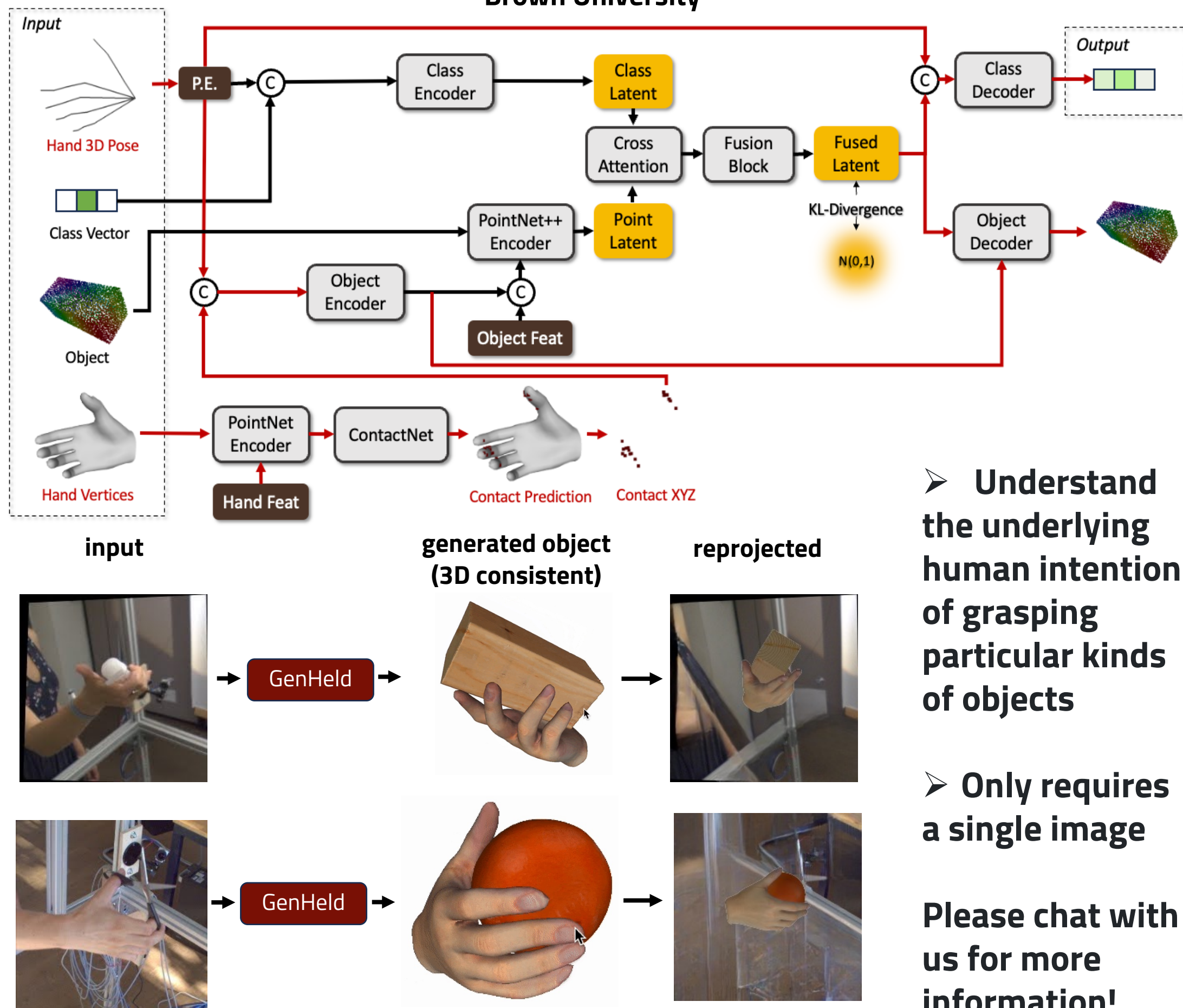2D Edit   3D Edit   3D Edit   Removal   Scale

GeoDiffuser

➤ Geometry Editing on Real Images
➤ Training-Free
➤ Prompt-Free
➤ Novel Geometry-Based Shared Attention
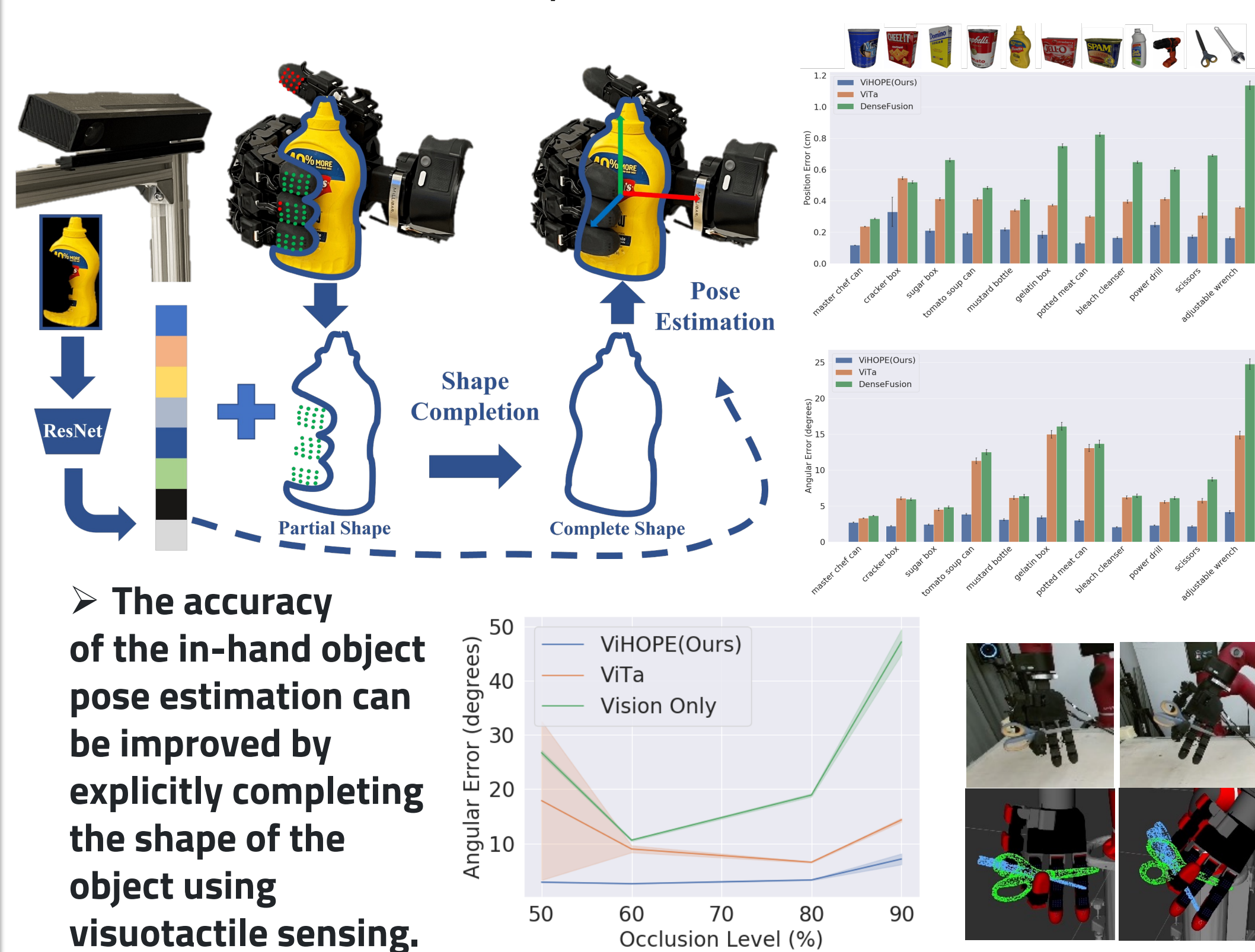
Please chat with us for more information!

GenHeld: Generate Held Objects

Chaerin Min[1], and Srinath Sridhar[1]
[1]Brown University

Input

P.E.

Hand 3D Pose

Class Vector

Object

Class Encoder → Class Latent

Cross Attention → Fusion Block → Fused Latent

PointNet++ Encoder → Point Latent

KL-Divergence

N(0,1)

Class Decoder → Output

Object Encoder → Object Feat

Object Decoder

Hand Vertices

PointNet Encoder → ContactNet → Contact Prediction   Contact XYZ

Hand Feat

input   generated object (3D consistent)   reprojected

GenHeld

GenHeld

➤ Understand the underlying human intention of grasping particular kinds of objects

➤ Only requires a single image

Please chat with us for more information!

ViHOPE: Visuotactile In-Hand Object 6D Pose Estimation with Shape Completion

Hongyu Li[1,2], Snehal Dikhale[2], Soshi Iba[2], and Nawid Jamali[2]
[1]Brown University and [2]Honda Research Institute

ResNet

Partial Shape → Shape Completion → Complete Shape

Pose Estimation

VIHOPE(Ours)   ViTa   DenseFusion

➤ The accuracy of the in-hand object pose estimation can be improved by explicitly completing the shape of the object using visuotactile sensing.

ViHOPE(Ours)   ViTa   Vision Only

Angular Error (degrees)

Occlusion Level (%)

Indoor Scene Understanding

Qiuhong Anna Wei[1] and Srinath Sridhar[1]
[1]Brown University

➤ Scene understanding using 3D or 2D information

Messy Scene

Cleaned "Regular" Scene

Object Attribute Encoder

c_i PE
t_i PE
r_i PE
b_i PE
h_i PE

δ_i

Transformer F_θ → $\Delta t_i, \Delta r_i$

Floor Plan → PointNet

Denoising Transformer Architecture

NSF   AFOSR   Office of Naval Research   NASA   Meta   amazon   Google