Novelty Assessment Report

Paper: CTRL&SHIFT: High-quality Geometry-Aware Object Manipulation in Visual Generation
PDF URL: https://openreview.net/pdf?id=T6T0JUgGFQ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Object-level manipulation—relocating or reorienting objects in images or videos while preserving scene realism—is central to film post-production, AR, and creative editing. Yet existing methods struggle to jointly achieve three core goals: background preservation, geometric consistency under viewpoint shifts, and user-controllable transformations. Geometry-based approaches offer precise control but require explicit 3D reconstruction and generalize poorly; diffusion-based methods generalize better but lack fine-grained geometric control. We present **Ctrl&Shift**, an end-to-end diffusion framework to achieve geometry-consistent object manipulation without explicit 3D representations. Our key insight is to decompose manipulation into two stages—object removal and reference-guided inpainting under explicit camera pose control—and encode both within a unified diffusion process. To enable precise, disentangled control, we design a multi-task, multi-stage training strategy that separates background, identity, and pose signals across tasks. To improve generalization, we introduce a scalable real-world dataset construction pipeline that generates paired image and video samples with estimated relative camera poses. Extensive experiments demonstrate that **Ctrl&Shift** achieves state-of-the-art results in fidelity, viewpoint consistency, and controllability. To our knowledge, this is the first framework to unify fine-grained geometric control and real-world generalization for object manipulation—without relying on any explicit 3D modeling.

Core Task Landscape

This paper addresses: **Geometry-aware object manipulation in images and videos**
A total of **50 papers** were analyzed and organized into a taxonomy with **41 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Robotic Manipulation with Geometric Reasoning**
- **Visual Content Editing and Generation**
- **Domain-Specific Applications**
- **3D Perception and Representation**

Complete Taxonomy Tree

- Geometry-aware object manipulation in images and videos Survey Taxonomy
- Robotic Manipulation with Geometric Reasoning
  - Vision-Based Grasping and Pose Estimation
  - Hand-Object Interaction and Occlusion Handling (1 papers)
    - [3] Geometry-Aware 3D Handâ□□Object Pose Estimation Under Occlusion via Hierarchical Feature Decoupling (Yuting Cai, 2025) View paper
  - Multi-Property Object Perception (1 papers)
    - [5] Towards Intelligent Object Manipulation: Vision-Based Grasping, Pose Estimation, and Physical Property Identification (Junhao Cai, 2025) View paper
  - Depth Completion for Transparent Objects (1 papers)
    - [25] CAGT: Sim-to-Real Depth Completion with Interactive Embedding Aggregation and Geometry Awareness for Transparent Objects (Xingshuo Jing, 2025) View paper
  - General Vision Systems for Manipulation (1 papers)
    - [29] Vision for robotic object manipulation in domestic settings (Danica Kragic, 2005) View paper
  - Articulated and Multi-Object Manipulation
  - Adaptive Articulated Object Manipulation (1 papers)
    - [2] Adaptive Articulated Object Manipulation On The Fly with Foundation Model Reasoning and Part Grounding (Zhang Xiao-jie, 2025) View paper
  - Multi-Object Scene Modeling (1 papers)
    - [6] Multi-object manipulation via object-centric neural scattering functions (Stephen Tian, 2023) View paper
  - Language-Grounded Spatial Reasoning for Manipulation
  - Spatial Affordance Prediction (2 papers)
    - [7] Robopoint: A vision-language model for spatial affordance prediction for robotics (Yuan, 2024) View paper
    - [44] From Seeing to Doing: Bridging Reasoning and Decision for Robotic Manipulation (Yuan, 2025) View paper
  - Semantic Orientation and Geometric Constraints (2 papers)
    - [18] Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation (Qi, 2025) View paper
    - [40] Geomanip: Geometric constraints as general interfaces for robot manipulation (Tang Wei-liang, 2025) View paper
  - Deformable Object Manipulation (2 papers)
  - [37] Learning latent graph dynamics for visual manipulation of deformable objects (Xiao Ma, 2022) View paper

- [45] Learning Rope Manipulation Policies Using Dense Object Descriptors Trained on Synthetic Depth Data (Priya Sundaresan, 2020) View paper
- Task-Oriented Manipulation Planning
- Zero-Shot Task-Oriented Grasping (1 papers)
  - [46] Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition (Samuel Li, 2024) View paper
- Affordance-Based Cognitive Planning (1 papers)
  - [47] Task-oriented robot cognitive manipulation planning using affordance segmentation and logic reasoning (Zhongli Wang, 2023) View paper
- Trajectory Optimization and Mobile Manipulation (1 papers)
- [32] M2 Diffuser: Diffusion-based Trajectory Optimization for Mobile Manipulation in 3D Scenes (Sixu Yan, 2025) View paper
- Learning from Demonstrations and Benchmarks
- Manipulation Skill Benchmarks (2 papers)
  - [16] Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations (Mu, 2021) View paper
  - [24] Maniskill2: A unified benchmark for generalizable manipulation skills (Gu, 2023) View paper
- Learning from Human Video Demonstrations (1 papers)
  - [50] Vision-based Manipulation from Single Human Video with Open-World Object Graphs (Zhu Yi-feng, 2024) View paper
- Interactive Scene Exploration (1 papers)
  - [31] Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation (Jiang, 2024) View paper
- Diffusion-Based Robotic Planning
- Image-Editing Diffusion for Subgoal Planning (1 papers)
  - [12] Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models (Black, 2023) View paper
- Video Generation for Manipulation (3 papers)
  - [1] Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping (Youxin Pang, 2025) View paper
  - [8] Learning Video Generation for Robotic Manipulation with Collaborative Trajectory Control (Fu Xiao, 2025) View paper
  - [22] Geometry-aware 4D Video Generation for Robot Manipulation (Liu Zeyi, 2025) View paper
- Geometric Reasoning for Precise Placement (2 papers)
- [30] Deep SE(3)-Equivariant Geometric Reasoning for Precise Placement Tasks (Eisner, 2024) View paper
- [43] Visual-Tactile Perception Based Control Strategy for Complex Robot Peg-in-Hole Process via Topological and Geometric Reasoning (Gaozhao Wang, 2024) View paper
- Open-Instruction 6-DoF Rearrangement (1 papers)
- [49] Open6DOR: Benchmarking open-instruction 6-DoF object rearrangement and a VLM-based approach (Yufei Ding, 2024) View paper
- Multi-View Understanding for Embodied AI (1 papers)
- [13] Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs (Yeh, 2025) View paper

• Visual Content Editing and Generation
- Geometry-Aware Image Editing
- 3D Geometry-Based Image Editing ★ (2 papers)
  - [0] CTRL&SHIFT: High-quality Geometry-Aware Object Manipulation in Visual Generation (Anon et al., 2026) View paper
  - [21] Image sculpting: Precise object editing with 3d geometry control (Jiraphon Yenphraphai, 2024) View paper
- Drag-Based Image Editing with Mesh Guidance (1 papers)
  - [17] Flowdrag: 3d-aware drag-based image editing with mesh-guided deformation vector flow fields (Yoon, 2025) View paper
- Learning from Dynamic Videos for Editing (1 papers)
  - [15] Magic fixup: Streamlining photo editing by watching dynamic videos (Hadi Alzayer, 2025) View paper
- Geometry-Aware Video Editing
- Layered Video Editing with Occlusion Modeling (1 papers)
  - [4] Text2LIVE: Text-Driven Layered Image and Video Editing (Omer Bar-Tal, 2022) View paper
- 3D Object-Centric Video Editing (1 papers)
  - [9] Playable environments: Video manipulation in space and time (Willi Menapace, 2022) View paper
- Motion and Appearance Editing with Diffusion (1 papers)
  - [10] Uniedit: A unified tuning-free framework for video motion and appearance editing (Jianhong Bai, 2025) View paper
- Appearance-Focused Video Editing (1 papers)
  - [26] Adapedit: Spatio-temporal guided adaptive editing algorithm for text-based continuity-sensitive image editing (Zhiyuan Ma, 2024) View paper
- Zero-Shot and One-Shot Video Editing
- Spatiotemporal Slice-Based Editing (1 papers)
  - [27] Slicedit: Zero-Shot Video Editing With Text-to-Image Diffusion Models Using Spatio-Temporal Slices (Cohen, 2024) View paper
- Attention-Controlled Video Editing (2 papers)
  - [35] RealCraft: Attention Control as A Tool for Zero-Shot Consistent Video Editing (Shutong Jin, 2025) View paper
  - [42] Visual Prompting for One-shot Controllable Video Editing without Inversion (ZhengBo Zhang, 2025) View paper
- Localized Semantic Video Editing (1 papers)
  - [36] Videoshop: Localized Semantic Video Editing with Noise-Extrapolated Diffusion Inversion (Fan Xiang, 2024) View paper
- Randomized Noise Shuffling for Editing (1 papers)
  - [38] RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models (Ozgur Kara, 2023) View paper
- Action and Reasoning-Centric Editing (1 papers)
- [39] Learning Action and Reasoning-Centric Image Editing from Videos and Simulations (Krojer, 2024) View paper
- Unified Image and Video Generation (1 papers)
- [23] UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics (Xi Chen, 2024) View paper
- 3D-Aware Multi-Object Scene Synthesis (1 papers)
- [33] Neural Assets: 3D-Aware Multi-Object Scene Synthesis with Image Diffusion Models (Kelsey Allen, 2024) View paper

- Positional Encoding for Visual Generation (1 papers)
- [34] Positional encoding field (Bai Yun-peng, 2025) View paper

- Domain-Specific Applications
  - Autonomous Driving Scene Editing (2 papers)
  - [14] Vehiclesim: realistic and 3D-aware video editing with one image for autonomous driving (Beike Yu, 2025) View paper
  - [20] DriveEditor: A Unified 3D Information-Guided Framework for Controllable Object Editing in Driving Scenes (Yiyuan Liang, 2025) View paper
  - Geometry-Aware Driving Scene Simulation (1 papers)
  - [41] Geosim: Realistic video simulation via geometry-aware composition for self-driving (Chen Yun, 2021) View paper
  - Aerial Spatial Reasoning (1 papers)
  - [11] AirSpatialBot: A Spatially-Aware Aerial Agent for Fine-Grained Vehicle Attribute Recognization and Retrieval (Y Zhou, 2025) View paper
- 3D Perception and Representation
  - Multi-View 3D Object Detection (1 papers)
  - [19] Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection (Tao Tu, 2023) View paper
  - Diffusion-Based 3D Detection (1 papers)
  - [28] 3difftection: 3d object detection with geometry-aware diffusion features (ChenFeng Xu, 2024) View paper
  - Geometry-Aware Representation Learning (1 papers)
  - [48] Learning geometry-aware representations by sketching (Hyundo Lee, 2023) View paper

Narrative

Core task: Geometry-aware object manipulation in images and videos. This field spans a diverse set of challenges, from enabling robots to reason about spatial relationships during physical manipulation to editing visual content in ways that respect underlying 3D structure. The taxonomy reflects four main branches: Robotic Manipulation with Geometric Reasoning focuses on planning and control for physical systems, often leveraging simulation environments like Maniskill[16] and methods that integrate spatial understanding into policy learning. Visual Content Editing and Generation addresses how to modify images or videos while preserving geometric consistency, including approaches that use 3D priors or learned representations to guide edits. Domain-Specific Applications target specialized settings such as autonomous driving scene editing (DriveEditor[20]) or hand-object interaction modeling (Geometry Aware Hand[3]). Finally, 3D Perception and Representation explores foundational techniques for extracting and representing geometric information from visual data, which underpins both robotic and editing pipelines.

Within Visual Content Editing and Generation, a particularly active line of work centers on geometry-based image editing, where methods must balance creative flexibility with physical plausibility. Some approaches, like Text2LIVE[4], emphasize text-driven edits that adapt to scene structure, while others such as Image Sculpting[21] and CTRL SHIFT[0] focus more explicitly on manipulating objects in a manner consistent with their 3D geometry. CTRL SHIFT[0] sits within the 3D Geometry-Based Image Editing cluster, sharing conceptual ground with Image Sculpting[21] in its emphasis on respecting spatial constraints during edits. Compared to broader editing frameworks like Uniedit[10] or video-focused methods such as Manivideo[1], CTRL SHIFT[0] prioritizes geometric fidelity over purely appearance-driven transformations. This positioning highlights an ongoing tension in the field: how to integrate strong geometric priors without sacrificing the ease and expressiveness that make generative editing tools appealing to practitioners.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Image sculpting: Precise object editing with 3d geometry control

Authors: Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, Saining Xie, et al. (6 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We present Image Sculpting, a new framework for editing 2D images by incorporating tools from 3D geometry and graphics. This approach differs markedly from existing methods, which are confined to 2D spaces and typically rely on textual instructions, leading to ambiguity and limited control. Image Sculpting converts 2D objects into 3D, enabling direct interaction with their 3D geometry. Post-editing, these objects are re-rendered into 2D, merging into the original image to produce high-fidelity r...

Relationship Analysis

Both papers belong to the 3D Geometry-Based Image Editing category, converting 2D objects to 3D for geometric manipulation and re-rendering. They share the core approach of lifting 2D images to 3D representations (NeRF-based reconstruction), enabling precise geometric transformations (rotation, translation, pose editing), and rendering back to 2D with enhancement pipelines. The key difference is that CTRL&SHIFT appears to focus on a streamlined editing workflow with geometry-aware controls, while Image Sculpting emphasizes interactive 3D mesh deformation techniques (ARAP, linear blend skinning) combined with a coarse-to-fine generative enhancement process using feature injection and depth control for high-fidelity results.

Contributions Analysis

Overall novelty summary. The paper introduces Ctrl&Shift, a diffusion-based framework for geometry-consistent object manipulation in images and videos. It resides in the '3D Geometry-Based Image Editing' leaf of the taxonomy, which contains only two papers total (including this one). This sparse population suggests the specific combination of diffusion models with explicit geometric control for object-level editing remains relatively underexplored. The sibling paper (Image Sculpting) shares the goal of geometry-aware manipulation but differs in technical approach, indicating this research direction is nascent rather than saturated.

The taxonomy reveals that Ctrl&Shift sits within 'Visual Content Editing and Generation', adjacent to several related but distinct directions. Neighboring leaves include 'Drag-Based Image Editing with Mesh Guidance' (which uses explicit mesh deformation) and 'Learning from Dynamic Videos for Editing' (which focuses on photorealistic lighting from video). The broader 'Geometry-Aware Video Editing' branch contains methods like layered representations and volumetric rendering, but these typically require different technical machinery. The taxonomy's scope notes clarify that Ctrl&Shift excludes robotic execution (unlike the 'Robotic Manipulation' branch) and focuses on visual editing without physical interaction.

Among ten candidates examined for the single analyzed contribution, zero were found to clearly refute the approach. This limited search scope—covering top-K semantic matches plus citation expansion—suggests that within the immediate neighborhood of related work, no prior method appears to provide the same combination of diffusion-based manipulation with explicit camera pose control and two-stage decomposition. However, the small candidate pool (ten papers) means the analysis cannot claim exhaustive coverage of all potentially overlapping prior work. The contribution appears more novel within this constrained search than it might under broader examination.

Based on the limited literature search (ten candidates), the work occupies a sparsely populated research direction at the intersection of diffusion models and geometry-aware editing. The taxonomy structure confirms this is an emerging area rather than a crowded field.

While the analysis provides useful context, the restricted search scope means definitive novelty claims require validation against a more comprehensive survey of related diffusion-based editing and 3D-aware generation methods.

This paper presents **1 main contributions**, each analyzed against relevant prior work:

Contribution 1: CTRL&SHIFT framework for geometry-aware object manipulation

Description: A framework that enables high-quality manipulation of objects in images while maintaining geometric awareness. The system allows for precise control over object positioning and transformations in visual generation tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Realistic and Controllable 3D Gaussian-Guided Object Editing for Driving Video Generation

URL: View paper

Brief Assessment

Gaussian Guided Driving[59] focuses on object editing in driving videos using 3D Gaussian Splatting combined with diffusion models, while the original paper addresses general geometry-aware object manipulation across diverse visual content without requiring explicit 3D representations at inference.

---

2. ASIMO: Agent-centric scene representation in multi-object manipulation

URL: View paper

Brief Assessment

ASIMO[54] focuses on vision-based RL for multi-object manipulation through scene decomposition and agent training, not on geometry-aware visual generation or image editing tasks that CTRL&SHIFT addresses.

---

3. HOSIG: Full-Body Human-Object-Scene Interaction Generation with Hierarchical Scene Perception

URL: View paper

Brief Assessment

HOSIG[56] focuses on full-body human-object-scene interaction generation with navigation and grasp synthesis, not on geometry-aware object manipulation in visual generation tasks like image/video editing.

---

4. Geometry-aware 4D Video Generation for Robot Manipulation

URL: View paper

Brief Assessment

Geometry Aware 4D[22] focuses on 4D video generation for robot manipulation with multi-view consistency through pointmap alignment, not on object manipulation in static images/videos for creative editing tasks.

---

5. Novel Demonstration Generation with Gaussian Splatting Enables Robust One-Shot Manipulation

URL: View paper

Brief Assessment

Gaussian Splatting Demo[55] focuses on robotic visuomotor policy learning through 3D Gaussian manipulation for demonstration augmentation, not on geometry-aware object manipulation in visual generation tasks with precise positioning control for image/video editing.

---

6. Manipnet: neural manipulation synthesis with a hand-object spatial representation

URL: View paper

Brief Assessment

Manipnet[53] focuses on neural manipulation synthesis using hand-object spatial representations with voxel occupancies and distance samples, not on geometry-aware object manipulation in visual generation with camera pose control for image editing tasks.

---

7. 6-dof graspnet: Variational grasp generation for object manipulation

URL: View paper

Brief Assessment

6DOF Graspnet[51] focuses on robotic grasp pose generation for object manipulation using variational autoencoders, not on visual generation or image editing tasks. The candidate addresses physical object grasping in robotics, while the original paper addresses geometry-aware manipulation in visual content generation.

---

8. G3Flow: Generative 3D Semantic Flow for Pose-aware and Generalizable Object Manipulation

URL: View paper

Brief Assessment

G3Flow[57] focuses on robotic manipulation with semantic flow for 3D object tracking and diffusion-based policies, not on geometry-aware image/video editing with camera pose control.

---

9. CineMaster: A 3D-Aware and Controllable Framework for Cinematic Text-to-Video Generation

URL: View paper

Brief Assessment

CineMaster[58] focuses on text-to-video generation with 3D scene construction and camera control, not on geometry-aware object manipulation in existing images/videos as CTRL&SHIFT does.

---

10. Object-Centric Instruction Augmentation for Robotic Manipulation

URL: View paper

Brief Assessment

Object Centric Instruction[52] focuses on robotic manipulation tasks using language instruction augmentation with position cues for pick-and-place operations, not visual generation or image/video editing with geometric control.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] CTRL&SHIFT: High-quality Geometry-Aware Object Manipulation in Visual Generation View paper
- [1] Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping View paper
- [2] Adaptive Articulated Object Manipulation On The Fly with Foundation Model Reasoning and Part Grounding View paper
- [3] Geometry-Aware 3D Handâ□□Object Pose Estimation Under Occlusion via Hierarchical Feature Decoupling View paper
- [4] Text2LIVE: Text-Driven Layered Image and Video Editing View paper
- [5] Towards Intelligent Object Manipulation: Vision-Based Grasping, Pose Estimation, and Physical Property Identification View paper
- [6] Multi-object manipulation via object-centric neural scattering functions View paper
- [7] Robopoint: A vision-language model for spatial affordance prediction for robotics View paper
- [8] Learning Video Generation for Robotic Manipulation with Collaborative Trajectory Control View paper
- [9] Playable environments: Video manipulation in space and time View paper
- [10] Uniedit: A unified tuning-free framework for video motion and appearance editing View paper
- [11] AirSpatialBot: A Spatially-Aware Aerial Agent for Fine-Grained Vehicle Attribute Recognization and Retrieval View paper
- [12] Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models View paper
- [13] Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs View paper
- [14] Vehiclesim: realistic and 3D-aware video editing with one image for autonomous driving View paper
- [15] Magic fixup: Streamlining photo editing by watching dynamic videos View paper
- [16] Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations View paper
- [17] Flowdrag: 3d-aware drag-based image editing with mesh-guided deformation vector flow fields View paper
- [18] Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation View paper
- [19] Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection View paper
- [20] DriveEditor: A Unified 3D Information-Guided Framework for Controllable Object Editing in Driving Scenes View paper
- [21] Image sculpting: Precise object editing with 3d geometry control View paper
- [22] Geometry-aware 4D Video Generation for Robot Manipulation View paper
- [23] UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics View paper
- [24] Maniskill2: A unified benchmark for generalizable manipulation skills View paper
- [25] CAGT: Sim-to-Real Depth Completion with Interactive Embedding Aggregation and Geometry Awareness for Transparent Objects View paper
- [26] Adapedit: Spatio-temporal guided adaptive editing algorithm for text-based continuity-sensitive image editing View paper
- [27] Slicedit: Zero-Shot Video Editing With Text-to-Image Diffusion Models Using Spatio-Temporal Slices View paper
- [28] 3difftection: 3d object detection with geometry-aware diffusion features View paper
- [29] Vision for robotic object manipulation in domestic settings View paper
- [30] Deep SE(3)-Equivariant Geometric Reasoning for Precise Placement Tasks View paper
- [31] Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation View paper
- [32] M2 Diffuser: Diffusion-based Trajectory Optimization for Mobile Manipulation in 3D Scenes View paper
- [33] Neural Assets: 3D-Aware Multi-Object Scene Synthesis with Image Diffusion Models View paper
- [34] Positional encoding field View paper
- [35] RealCraft: Attention Control as A Tool for Zero-Shot Consistent Video Editing View paper
- [36] Videoshop: Localized Semantic Video Editing with Noise-Extrapolated Diffusion Inversion View paper
- [37] Learning latent graph dynamics for visual manipulation of deformable objects View paper
- [38] RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models View paper
- [39] Learning Action and Reasoning-Centric Image Editing from Videos and Simulations View paper
- [40] Geomanip: Geometric constraints as general interfaces for robot manipulation View paper
- [41] Geosim: Realistic video simulation via geometry-aware composition for self-driving View paper
- [42] Visual Prompting for One-shot Controllable Video Editing without Inversion View paper
- [43] Visual-Tactile Perception Based Control Strategy for Complex Robot Peg-in-Hole Process via Topological and Geometric Reasoning View paper
- [44] From Seeing to Doing: Bridging Reasoning and Decision for Robotic Manipulation View paper
- [45] Learning Rope Manipulation Policies Using Dense Object Descriptors Trained on Synthetic Depth Data View paper
- [46] Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition View paper
- [47] Task-oriented robot cognitive manipulation planning using affordance segmentation and logic reasoning View paper
- [48] Learning geometry-aware representations by sketching View paper
- [49] Open6DOR: Benchmarking open-instruction 6-DoF object rearrangement and a VLM-based approach View paper
- [50] Vision-based Manipulation from Single Human Video with Open-World Object Graphs View paper
- [51] 6-dof graspnet: Variational grasp generation for object manipulation View paper
- [52] Object-Centric Instruction Augmentation for Robotic Manipulation View paper
- [53] Manipnet: neural manipulation synthesis with a hand-object spatial representation View paper
- [54] ASIMO: Agent-centric scene representation in multi-object manipulation View paper
- [55] Novel Demonstration Generation with Gaussian Splatting Enables Robust One-Shot Manipulation View paper
- [56] HOSIG: Full-Body Human-Object-Scene Interaction Generation with Hierarchical Scene Perception View paper
- [57] G3Flow: Generative 3D Semantic Flow for Pose-aware and Generalizable Object Manipulation View paper
- [58] CineMaster: A 3D-Aware and Controllable Framework for Cinematic Text-to-Video Generation View paper
- [59] Realistic and Controllable 3D Gaussian-Guided Object Editing for Driving Video Generation View paper