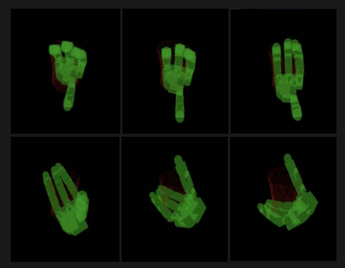
FLASH: Flow-Based Language-Annotated Grasp Synthesis for Dexterous Hands

Hrishit Leen, Jeremy A. Collins, Kunal Aneja, Nhi Nguyen, Priyadarshini Tamilselvan, Sri Siddarth Chakaravarthy P, Animesh Garg

bit.ly/flashgrasp

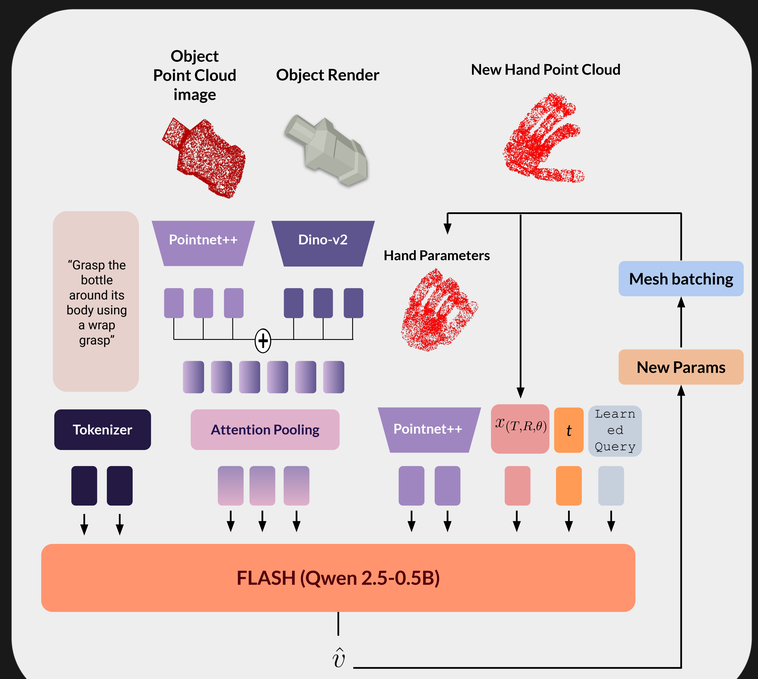
Problem Definition and Motivation

- Previous grasping methods decouple semantic intent from physical plausibility
- Many public grasp datasets have non-watertight meshes, weak language grounding, and unreliable SDFs
- Hand-pose vectors are a narrow information bottleneck; they ignore evolving geometry making models memorize joint statistics without considering contact



We present **FLASH**, a **conditional flow-matching** model that couples an **LLM backbone** with **live-updated hand meshes**, trained on our dataset's **richly annotated assets** to **synthesize dexterous grasps**

FLASH Architecture

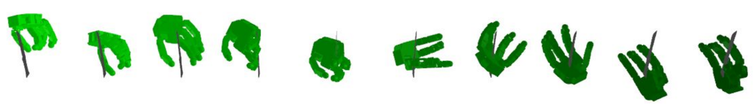


FLASH is a **conditional flow-matching architecture** trained on **FLASH**-Drive and contains:

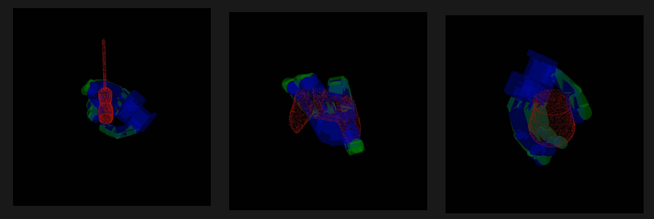
- A pre-trained Qwen2.5 LLM backbone
- Pointnet++ for point cloud processing
- Efficient Mesh Batching that feeds the live hand point cloud during flow inference

FLASH inference:

1. Encode object point cloud and text input
2. Sample x_0 from a gaussian centered around the mean of the hand poses
3. Numerically integrate the predicted velocity field $\hat{v}(x_\square, t, c)$, feeding a live point cloud of hand parameters
4. Return x_1 as the final grasp

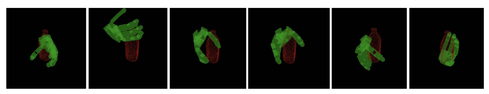


FLASH-Drive Dataset Quality

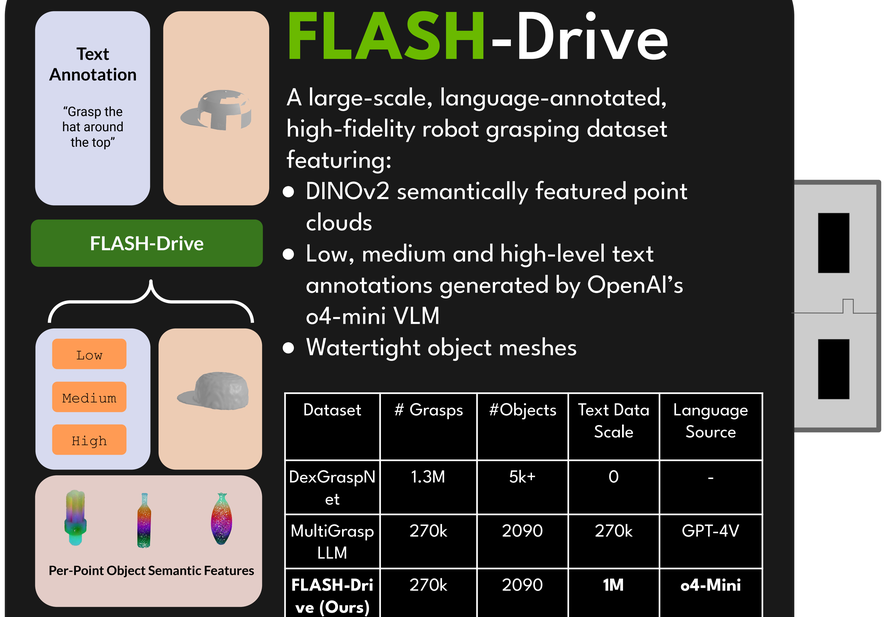


Qualitative comparisons on the improvement made on the original grasp quality of the MultiGraspLLM dataset. Our grasps (**blue**) reduce object penetration and increase grasp contact area.

Sample Trajectory from diffusion head baseline:



FLASH-Drive

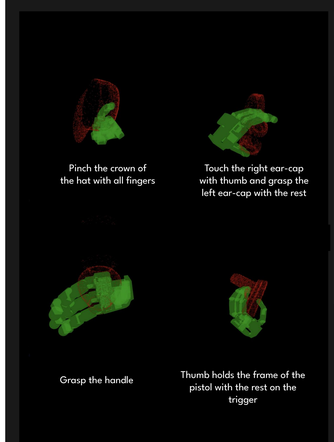


A large-scale, language-annotated, high-fidelity robot grasping dataset featuring:

- DINOv2 semantically featured point clouds
- Low, medium and high-level text annotations generated by OpenAI's o4-mini VLM
- Watertight object meshes

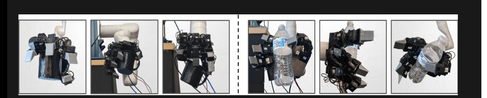
Dataset	# Grasps	#Objects	Text Data Scale	Language Source
DexGraspNet	1.3M	5k+	0	-
MultiGraspLLM	270k	2090	270k	GPT-4V
FLASH-Drive (Ours)	270k	2090	**1M**	**o4-Mini**

Simulation Results and Unseen Prompt Generalization



A grasp trajectory from one of the 1024 vectorized environments in our IsaacLab simulation evaluation setup

Method	Chamfer Dist.	Max Pen Dist.	Succ Rate	GPT Score (Align/Feas)
MultiGraspLLM	0.37	1.04	31.98	_/_
DexGraspNet	0.62	1.27	_	_/_
FLASH-Drive (Ours)	0.43	**0.36**	31.34	**55.2/79.0**



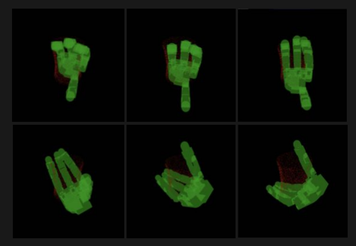
Real-world demonstration setup using a LEAP hand mounted on a Kinova Gen 3 arm used for qualitative validation.

FLASH: Flow-Based Language- An
Synthesis for Dexterous H

Hrishit Leen, Jeremy A. Collins, Kunal Aneja, Nhi Nguyen, Priyadarshini Tamilselvan, Sri

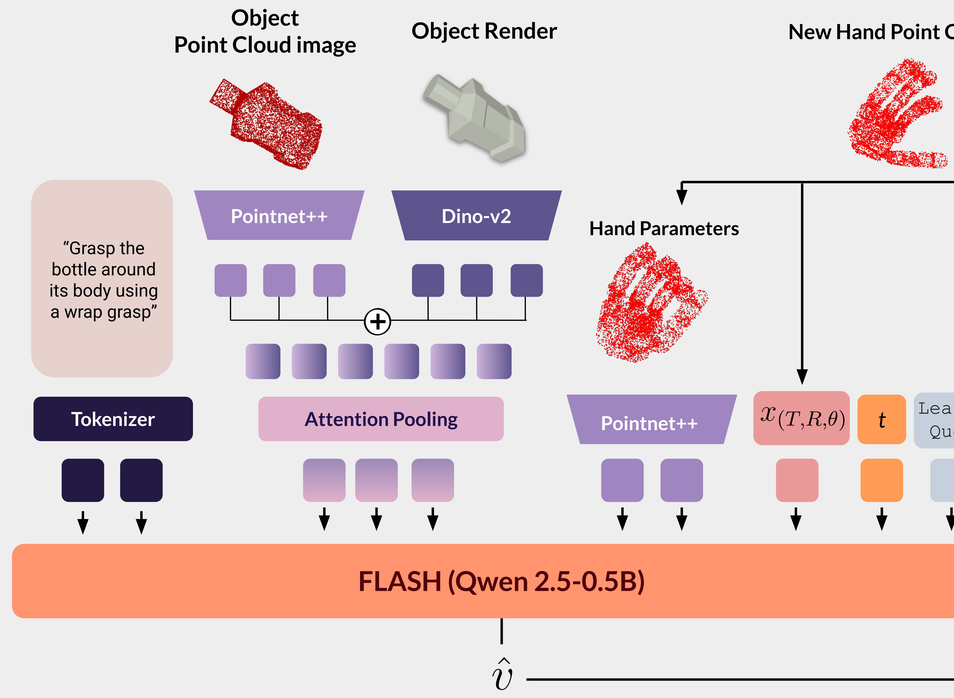
Problem Definition and Motivation

- Previous grasping methods decouple semantic intent from physical plausibility
- Many public grasp datasets have non-watertight meshes, weak language grounding, and unreliable SDFs
- Hand-pose vectors are a narrow information bottleneck; they ignore evolving geometry making models memorize joint statistics without considering contact



We present **FLASH**, a **conditional flow-matching** model that couples an **LLM backbone** with **live-updated hand meshes**, trained on our dataset's **richly annotated assets** to **synthesize dexterous grasps**

FLASH Architectur

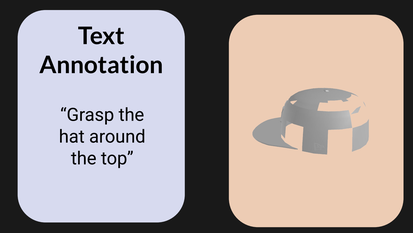


FLASH is a **conditional flow-matching architecture** trained on **FLASH**-Drive and contains:

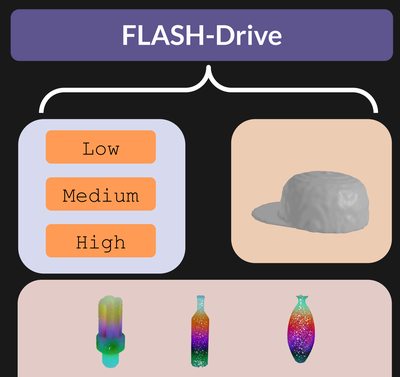
- A pre-trained Qwen2.5 LLM backbone
- Pointnet++ for point cloud processing
- Efficient Mesh Batching that feeds the live hand point cloud during flow inference

FLASH inference:

1. Encode object point clou
2. Sample x_0 from a gaussi around the mean of the
3. Numerically integrate th velocity field $\hat{v}(x$□, t, c), point cloud of hand para
4. Return x_1 as the final gr



FLASH-Drive



A large-scale, language-annotated, high-fidelity robot grasping dataset featuring:

- DINOv2 semantically featured point clouds
- Low, medium and high-level text annotations generated by OpenAI's o4-mini VLM
- Watertight object meshes

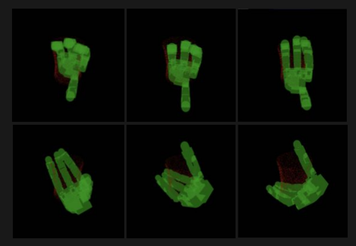
Dataset	# Grasps	#Objects	Text Data Scale	Language Source
DexGraspNet	1.3M	5k+	0	-
MultiGraspLLM	270k	2090	270k	GPT-4V

FLASH: Flow-Based Language- An
Synthesis for Dexterous H

Hrishit Leen, Jeremy A. Collins, Kunal Aneja, Nhi Nguyen, Priyadarshini Tamilselvan, Sri

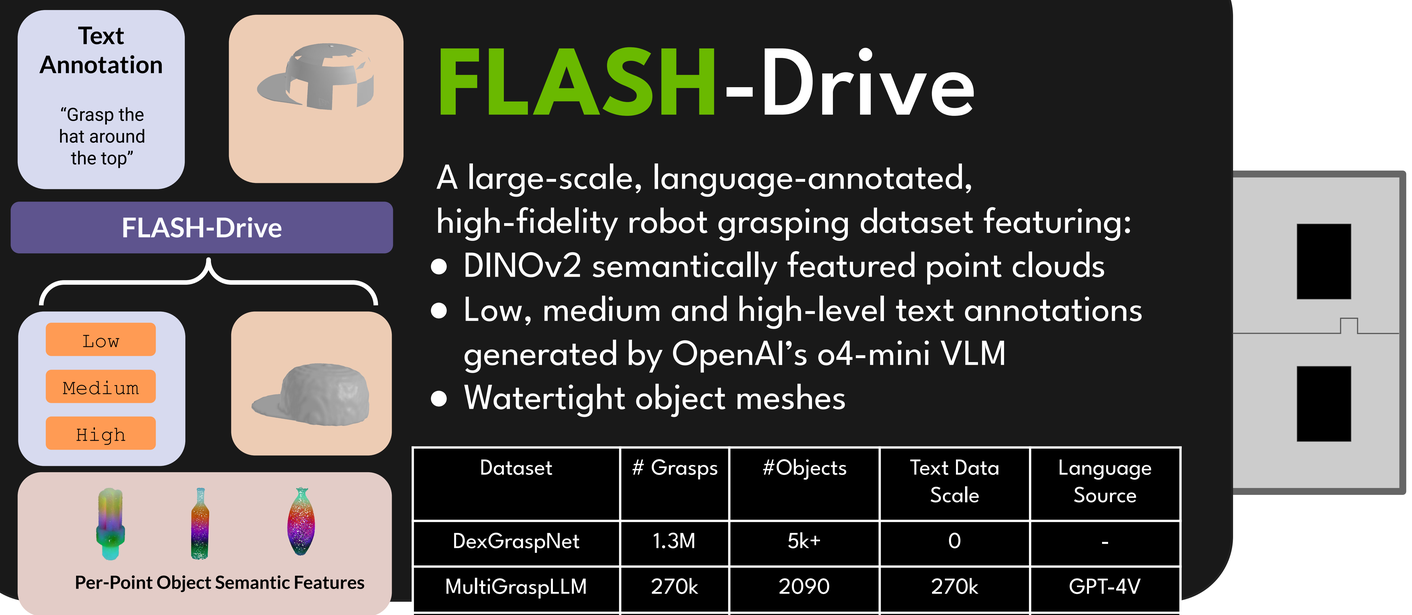
Problem Definition and Motivation

- Previous grasping methods decouple semantic intent from physical plausibility
- Many public grasp datasets have non-watertight meshes, weak language grounding, and unreliable SDFs
- Hand-pose vectors are a narrow information bottleneck; they ignore evolving geometry making models memorize joint statistics without considering contact



We present **FLASH**, a **conditional flow-matching** model that couples an **LLM backbone** with **live-updated hand meshes**, trained on our dataset's **richly annotated assets** to **synthesize dexterous grasps**

FLASH-Drive

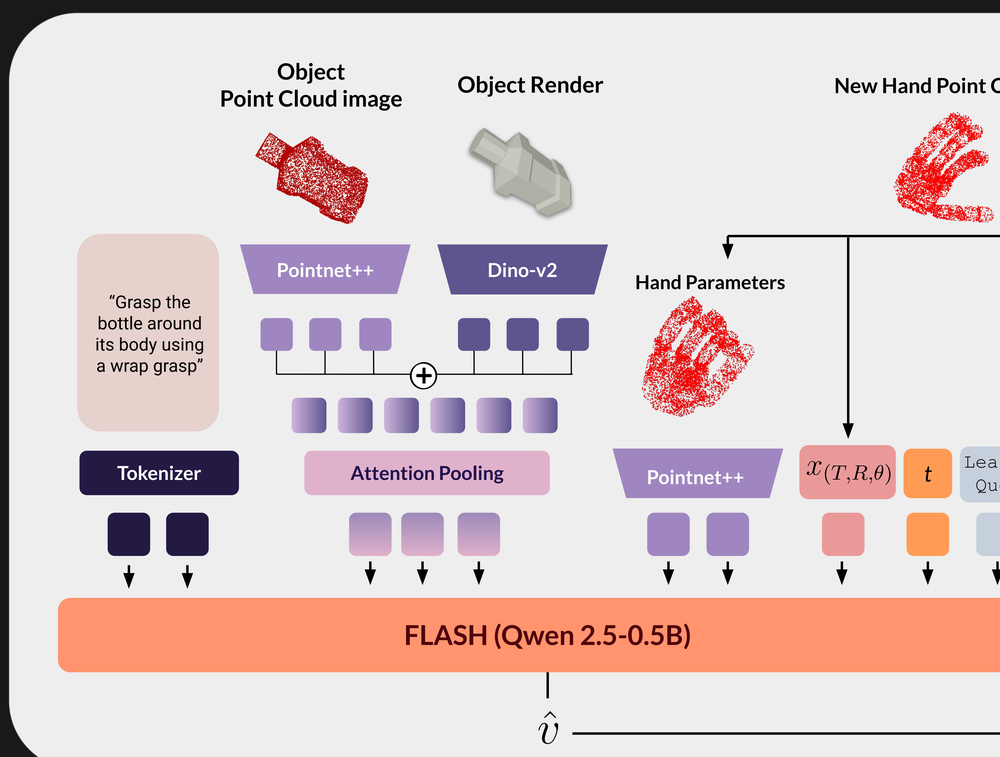


A large-scale, language-annotated, high-fidelity robot grasping dataset featuring:

- DINOv2 semantically featured point clouds
- Low, medium and high-level text annotations generated by OpenAI's o4-mini VLM
- Watertight object meshes

Dataset	# Grasps	#Objects	Text Data Scale	Language Source
DexGraspNet	1.3M	5k+	0	-
MultiGraspLLM	270k	2090	270k	GPT-4V

FLASH Architectur



FLASH is a **conditional flow-matching architecture** trained on **FLASH**-Drive and contains:

- A pre-trained Qwen2.5 LLM backbone
- Pointnet++ for point cloud processing
- Efficient Mesh Batching that feeds the live hand point cloud during flow inference

FLASH inference:

1. Encode object point clou
2. Sample x_0 from a gaussi around the mean of the
3. Numerically integrate th velocity field $\hat{v}(x$□, t, c), point cloud of hand para
4. Return x_1 as the final gr