APPENDIX A
ADDITIONAL RESULTS AND ANALYSIS

A. Ablation of the Nullspace Projection Method

We propose a nullspace projection method to address the potential decline in grasp quality caused by the introduction of guidance, which ultimately maintains grasp quality. The main task is defined as keeping a point on the palm fixed in the world coordinates. By projecting the guidance gradient ∇g to the main task's nullspace, the hand keeps facing the object in most cases, thus maintaining the grasp quality. The ablation results in Tab. VI show the percentage change of metrics when the nullspace projection is enabled (w/ proj.) compared to '**Ours**' in Tab. I. As shown by the results, the GSR increases by over 10% without projection, indicating improved grasp quality. However, there is minimal change in the SR, indicating that the absolute number of successful grasps in a batch remains consistent. This trend arises from the interplay between the decrease in feasible grasps (FR) and the increase in grasp quality (GSR). The decline in FR is expected, as the direction of gradient descent is modified by the projection. The OSR presents similar trends to SR.

B. Ablation of the Noised Gradient

When the noised wrist pose $\boldsymbol{x}_t$ instead of $\hat{\boldsymbol{x}}_0$ is used to compute ∇g in (13), we assume that the resulting noised gradient may result in poor convergence. The ablation results with noised gradient (w/ noised grad.) are presented in Tab. VI. The FR decreased by 11.75% and 0.85% in S1 and S2 as expected, while it increased by 8.31% in S3. The obstacles of S1 and S2 result in the robotic arm alternately collide with two walls and produces inconsistent gradients which hinder convergence. While for S3 with less obstacles, $\nabla g(\boldsymbol{x}_t)$ performs better by providing more direct and timely guidance. Therefore, in specific applications, the choice between $\boldsymbol{x}_t$ and $\hat{\boldsymbol{x}}_0$ should be determined by the metrics relevant to the particular scenario. In addition, the grasp quality marginally decreases (GSR), and the change of SR and OSR depends on the specific scene.

C. Influence of the Number of IK Solutions

Remember that solving (6) involves minimization over the IK solution set $\mathcal{Q}$, which is approximated with a finite set in practice. As shown in Tab. VII, increasing the number of IK solutions enhances all metrics, indicating that a better solution to (6) has been found. Note that using fewer than 4 IK solutions can significantly degrade performance with the UR5, as these solutions reside in disconnected sub-manifolds of the configuration space. Their varying order can result in unstable gradients due to the discontinuity of the IK solutions. In contrast, the performance with Franka is less affected, as it has a connected IK solution set. In addition, more inverse kinematics (IK) solutions reduce the number of grasps generated per second. Thus, an appropriate number—8 for UR5 and 10 for Franka—is chosen.

Table VI: Performance of the ablation groups compared to '**Ours**' in Tab. I

Scene	Method	Changes of Metrics(%)↑			
		FR	GSR	SR	OSR
S1	w/ proj.	-11.75	+13.11	+4.34	+0.68
	w/ noised grad.	-11.75	-3.66	-9.04	-2.35
S2	w/ proj.	-35.24	+16.35	-3.86	-3.19
	w/ noised grad.	-0.85	-3.05	-2.40	-1.86
S3	w/ proj.	-16.30	+11.42	+0.84	-1.22
	w/ noised grad.	+8.31	-0.50	+4.04	+0.92

D. Failure Cases of Executing Generated Grasps in the Real World

We present several failure cases in Fig. 8, highlighting potential limitations of the proposed method and possible directions for improvement. 1) In some cases, the guidance compromises the quality of a small subset of generated grasps, resulting in unstable configurations that only grasp the edge of the object (Fig. 8 (a-b)). This accounts for 7 out of 20 failed grasps. This issue often arises when the arm remains in collision until the end of the denoising process, causing conflicts between constraint satisfaction and convergence to the learned grasp distribution. Although we utilize the grasp evaluator to filter out high-quality grasps, we find that real-world point clouds are often noisier and more incomplete, leading to misestimation of grasp success probabilities. This issue can be mitigated through increased data augmentation and fine-tuning with

Table VII: Performance of grasp generation with varying number of IK solutions

Arm	IK Num.	Metrics(%)↑				Grasps Per Second↑
		FR	GSR	SR	OSR	
UR5	2	2.74	44.01	1.21	10.06	1257.86
	4	45.36	56.64	25.69	83.64	1153.40
	8	73.16	57.80	42.29	90.32	1021.45
Franka	5	35.47	48.79	17.31	71.58	676.59
	10	46.18	51.78	23.91	80.46	419.64
	20	60.70	53.65	32.57	85.72	237.25

Fig. 8: Failure cases in real-world experiments. (a–b) The quality of a small subset of generated grasps is degraded by the guidance, leading to unstable grasps holding only the object's edge. (c) The dexterous hand unintentionally contacts the object due to open-loop execution. (d) The object shifts in hand because of inaccurate mass parameter estimation or insufficient friction.

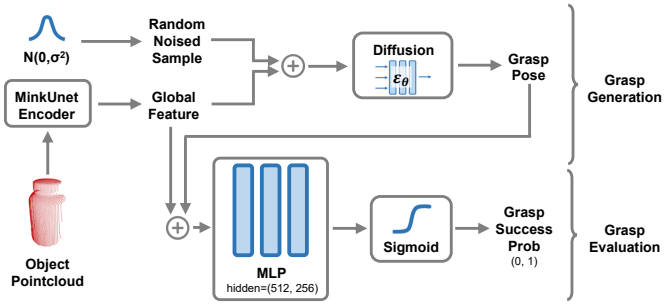

Fig. 9: Architecture of the grasp generation and evaluation networks.

real-world data. Another reason for failure is inaccuracies in motion planning and open-loop execution, which can result in unintended collisions, contributing to 9 out of 20 failed grasps (Fig. 8 (c)). This issue can be mitigated through real-time feedback and online sensing. Additionally, incorrect estimation of the object's mass parameters or insufficient friction accounts for the remaining 4 out of 20 failed grasps (Fig. 8 (d)), though this aspect lies beyond the scope of this work.

APPENDIX B
ADDITIONAL DETAILS OF THE PROPOSED METHOD

A. *Network Architecture*

Our network design is inspired by prior work [1]. For grasp pose generation, the object's partial point cloud is embedded into a 1024-dimensional global feature vector using a Minkowski Unet. This global feature, along with a randomly sampled noise from a Gaussian distribution, is input into the denoising process. The noise prediction network, conditioned on the global feature and the time step t, predicts a denoised sample from the noisy input. The grasp pose is then recovered from the fully denoised sample and the global feature. A grasp evaluator predicts the probability of successfully executing a given grasp, based on the grasp pose and the global feature. The evaluator is implemented as an MLP with hidden dimensions of $[512, 256]$, followed by a sigmoid activation function. The grasp evaluation uses the same training data as the generation network, augmented with execution success labels from performing all grasps in the MuJoCo simulator. The evaluator is trained using cross-entropy loss and successfully predicts 87% of grasps in the test set with its best checkpoint.