ICLR

Phidias: A Generative Model for Creating 3D Content from Text, Image, and 3D Conditions with Reference-Augmented Diffusion

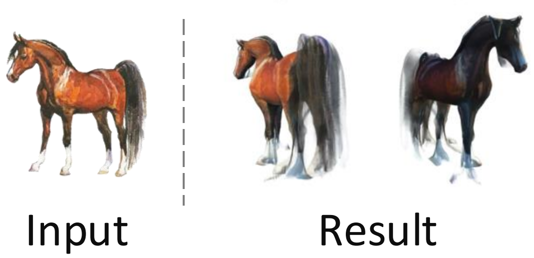
Zhenwei Wang[*1], Tengfei Wang*[*2], Zexin He[3], Gerhard Hancke[1], Ziwei Liu[4], Rynson W.H. Lau[1]*

[*]Joint First Author, [1]City University of Hong Kong, [2]Shanghai AI Lab,
[3]Chinese University of Hong Kong, [4]S-Lab, Nanyang Technological University

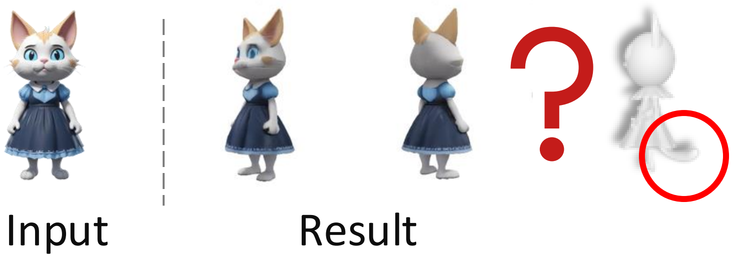
MOTIVATION

Problems of existing image to 3D generative models:

- Generation quality
- Generalization Ability
- Controllability



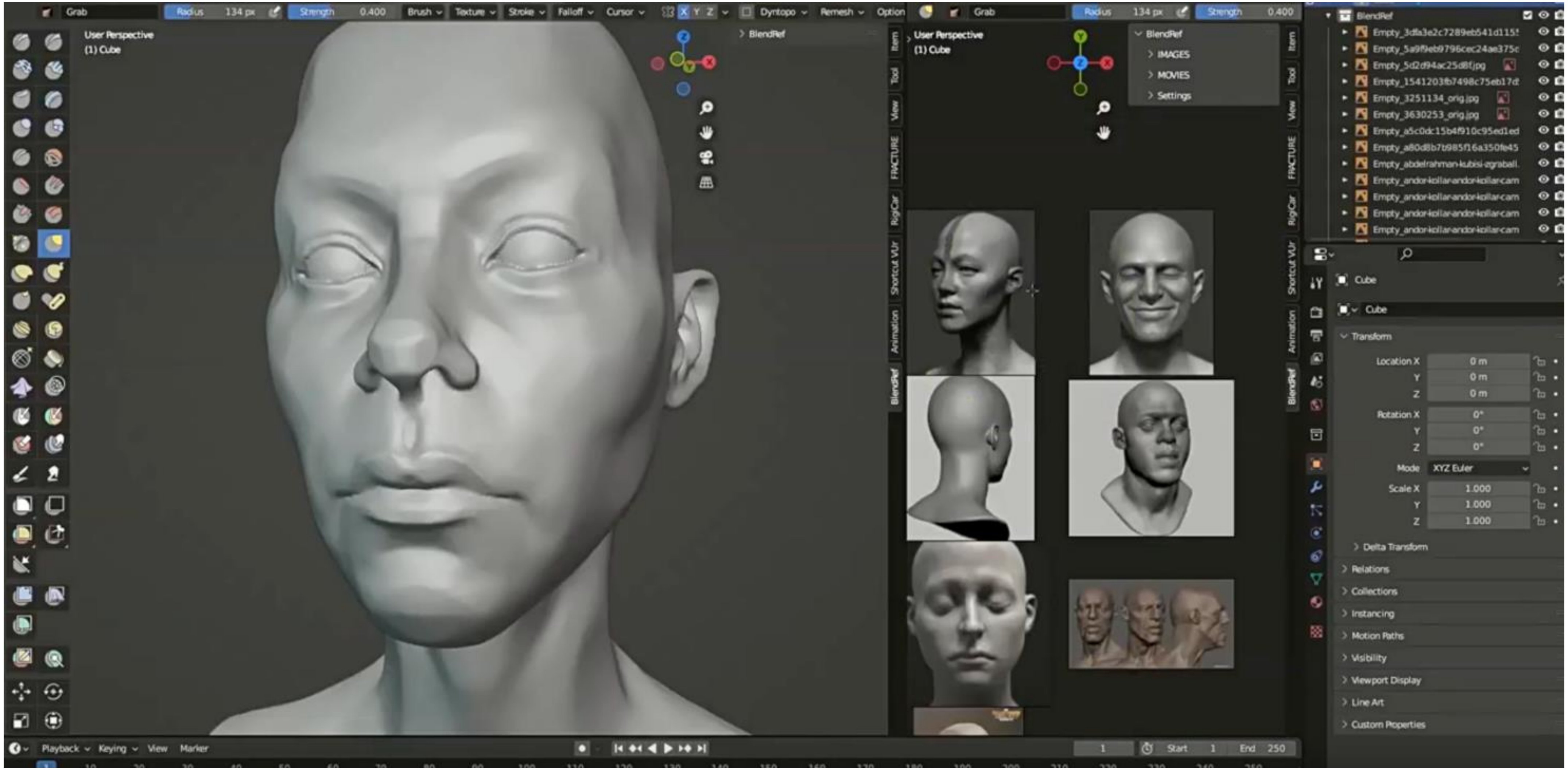
Input
Result



Success of Retrieval augmented generation (RAG) for language and image

What about RAG for 3D?

PRACTICAL 3D MODELING WORKFLOW



Grab

Radius 134 px

Strength 0.400

User Perspective
(1) Cube

Transform

Location X 0 m
Y 0 m
Z 0 m

Rotation X 0°
Y 0°
Z 0°

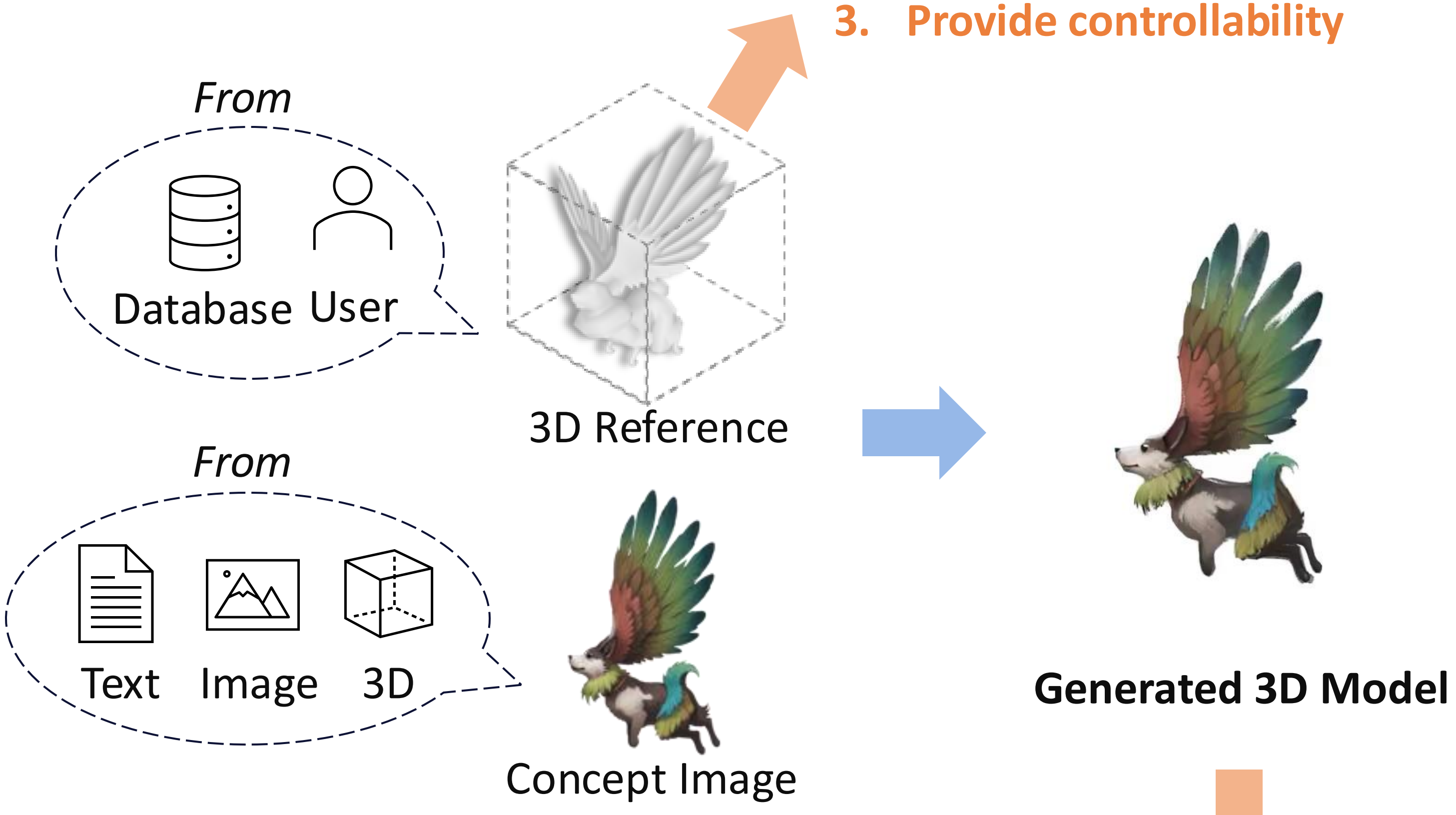
Mode XYZ Euler

Scale X 1.000
Y 1.000
Z 1.000

1 Start 1 End 250

OUR GOAL

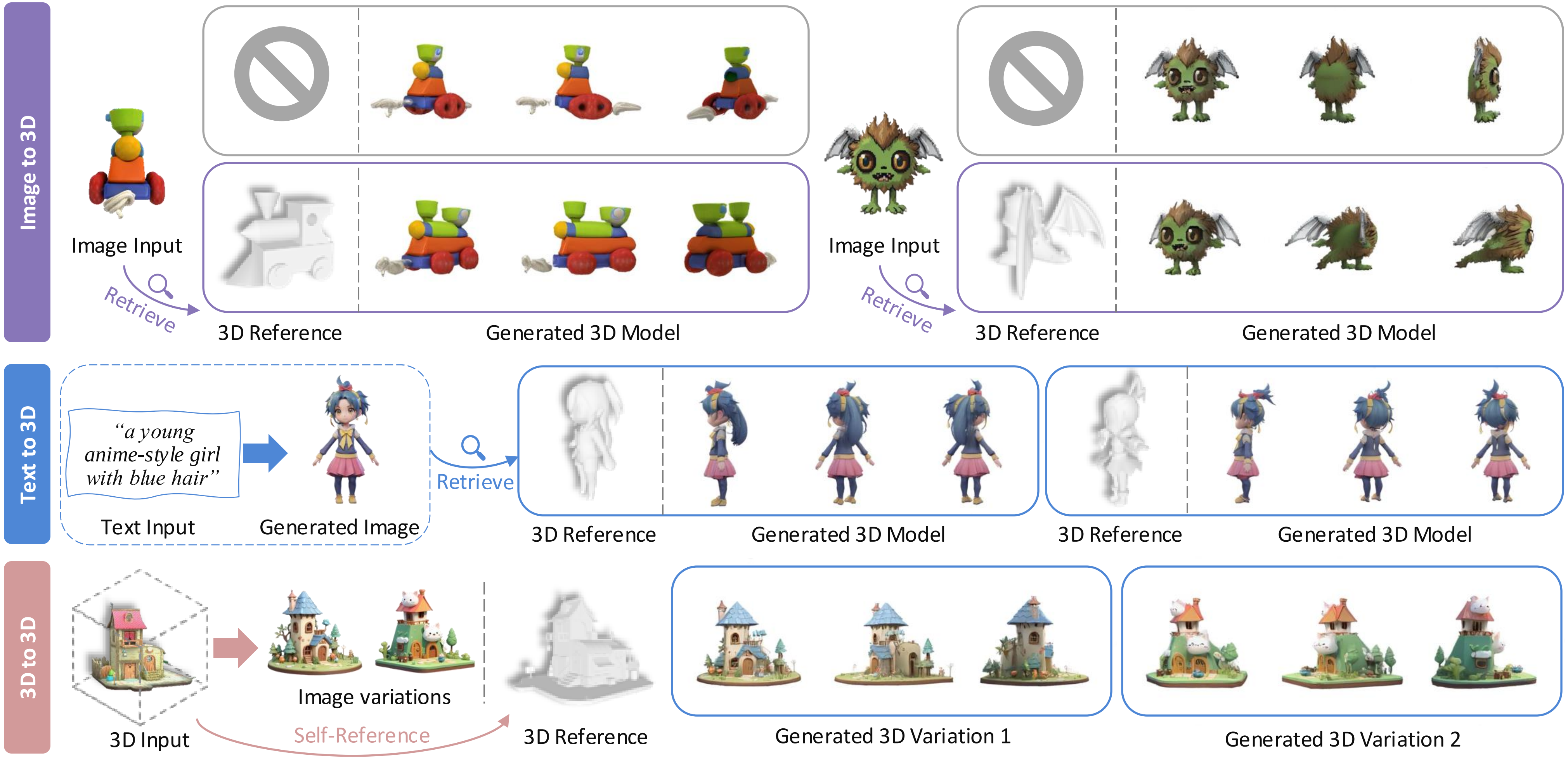
1. Improve quality
2. Enhance generalization ability
3. Provide controllability



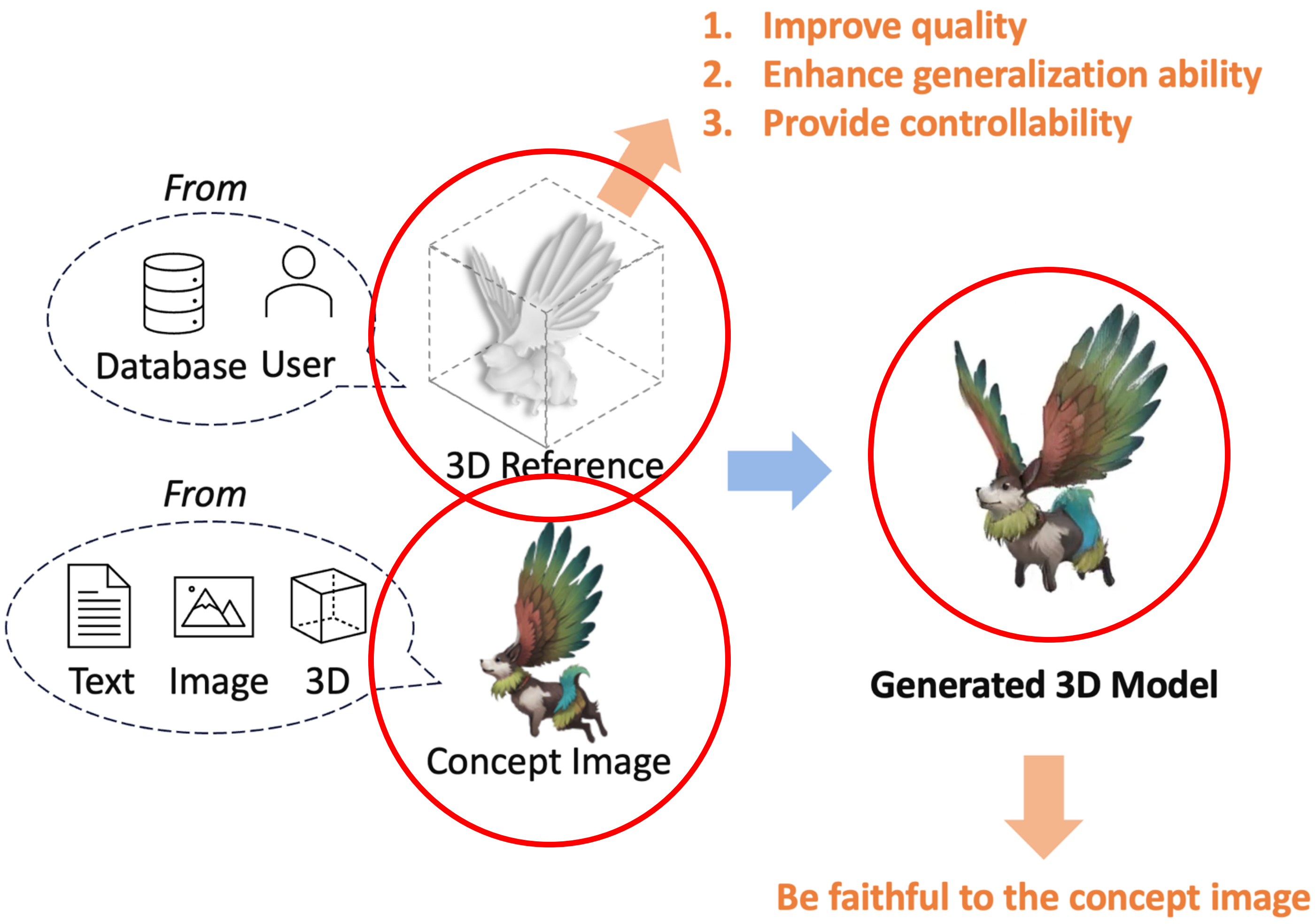
Be faithful to the concept image

OUR GOAL

Phidias



CHALLENGES

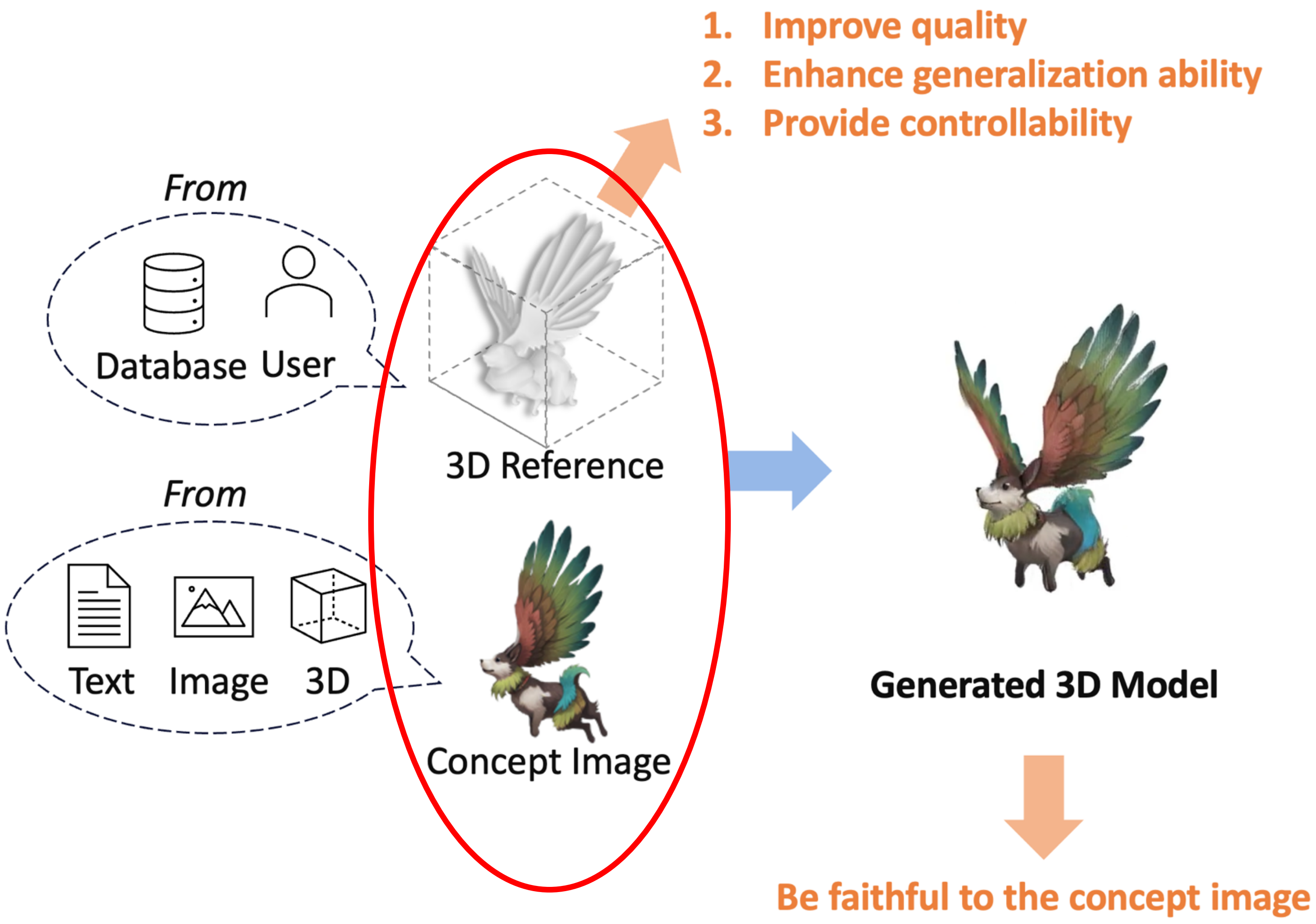


Data shortage

1. **Self-reference augmentations**

➔ self-supervised training with progressive curriculum

CHALLENGES



1. Improve quality
2. Enhance generalization ability
3. Provide controllability

Be faithful to the concept image

Data shortage

1. Self-reference augmentations

➔ self-supervised training with progressive curriculum

Misalignment dilemma

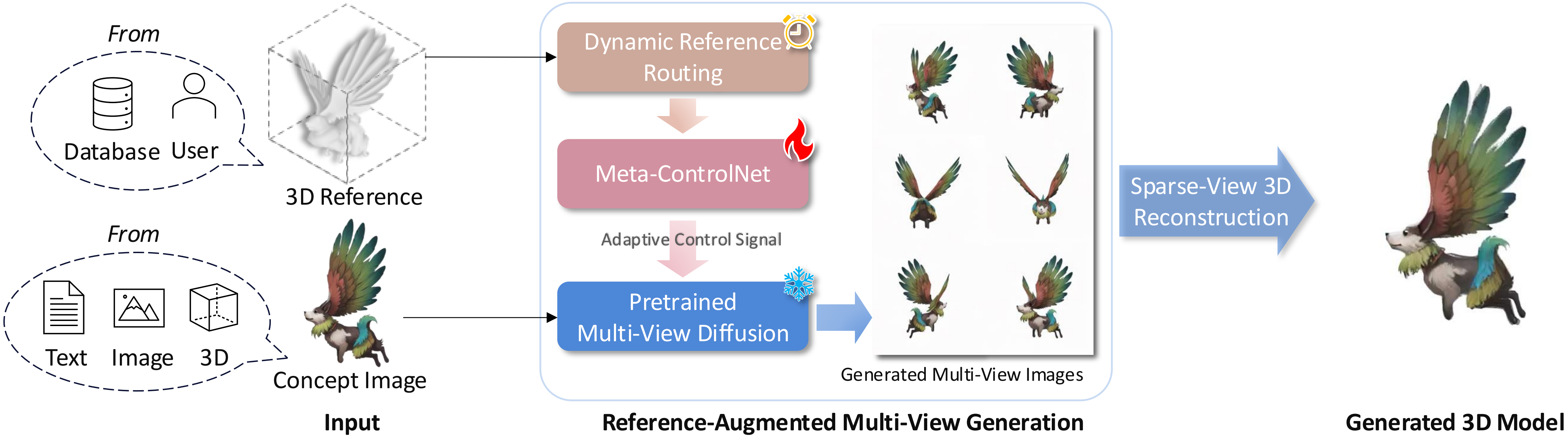
2. Meta-ControlNet
3. Dynamic reference routing

➔ dynamically modulates conditioning strength

➔ adjust reference resolution across denoise timestep

METHOD OVERVIEW

Phidias



From

Database User

3D Reference

From

Text Image 3D

Concept Image

Input

Dynamic Reference Routing

Meta-ControlNet

Adaptive Control Signal

Pretrained Multi-View Diffusion

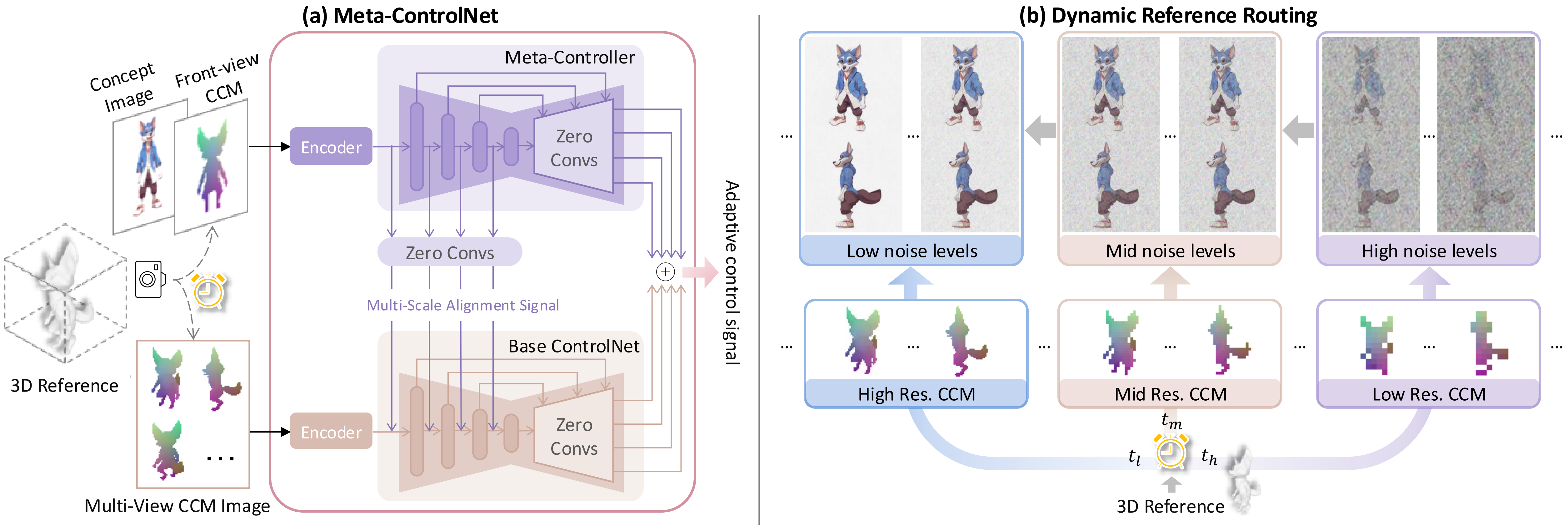
Generated Multi-View Images

Reference-Augmented Multi-View Generation

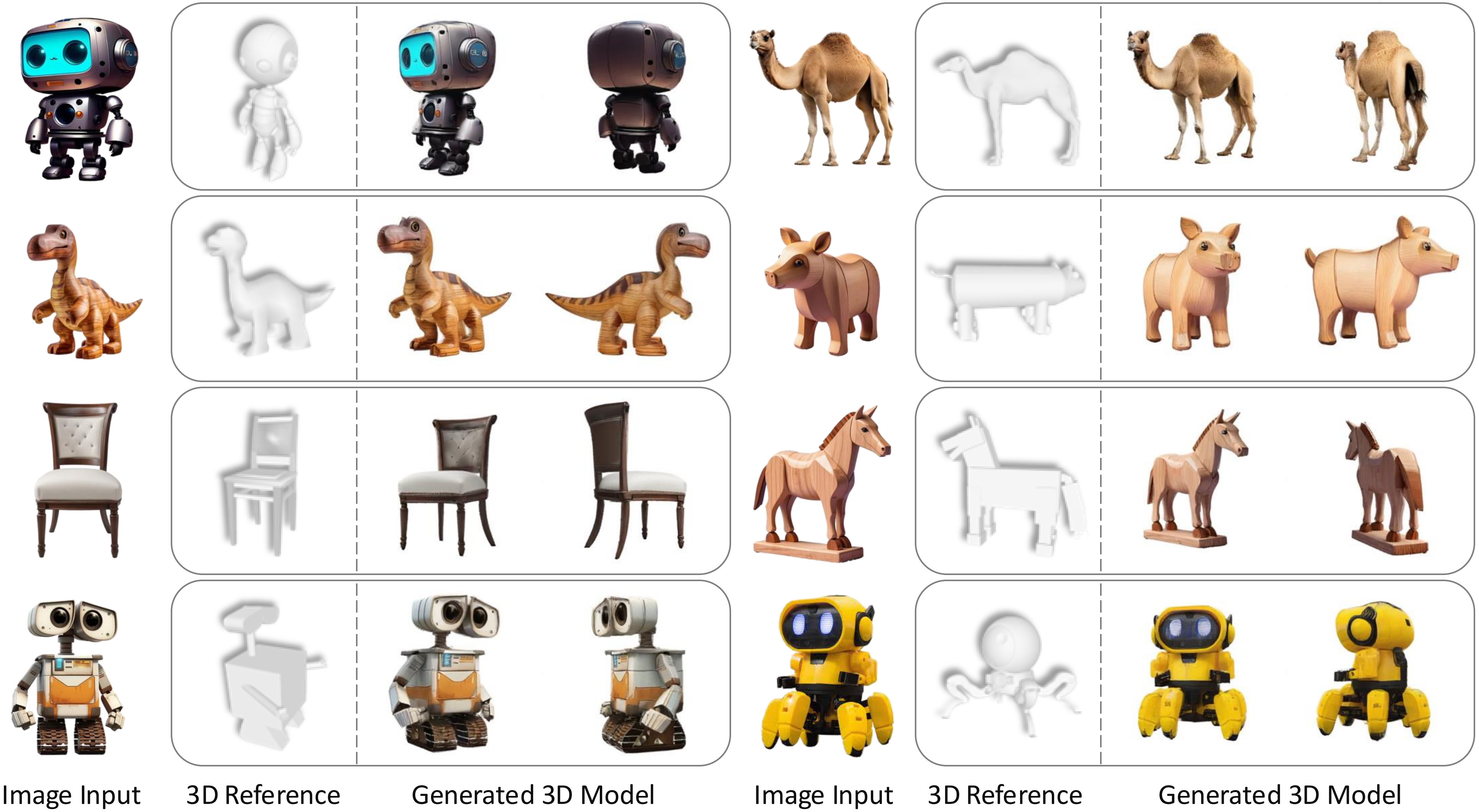
Sparse-View 3D Reconstruction

Generated 3D Model

MISALIGNMENT SOLUTIONS

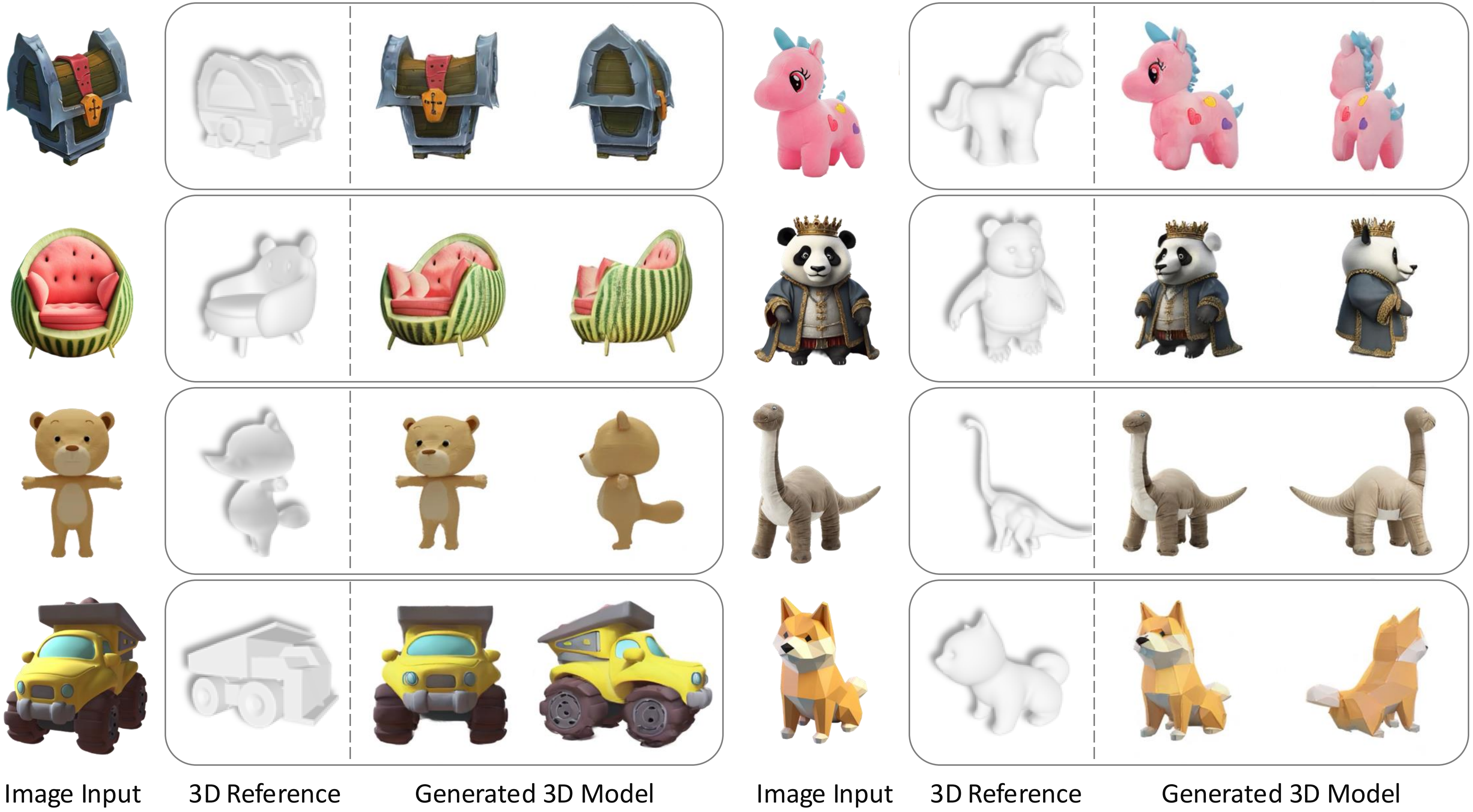


RAG IMAGE TO 3D

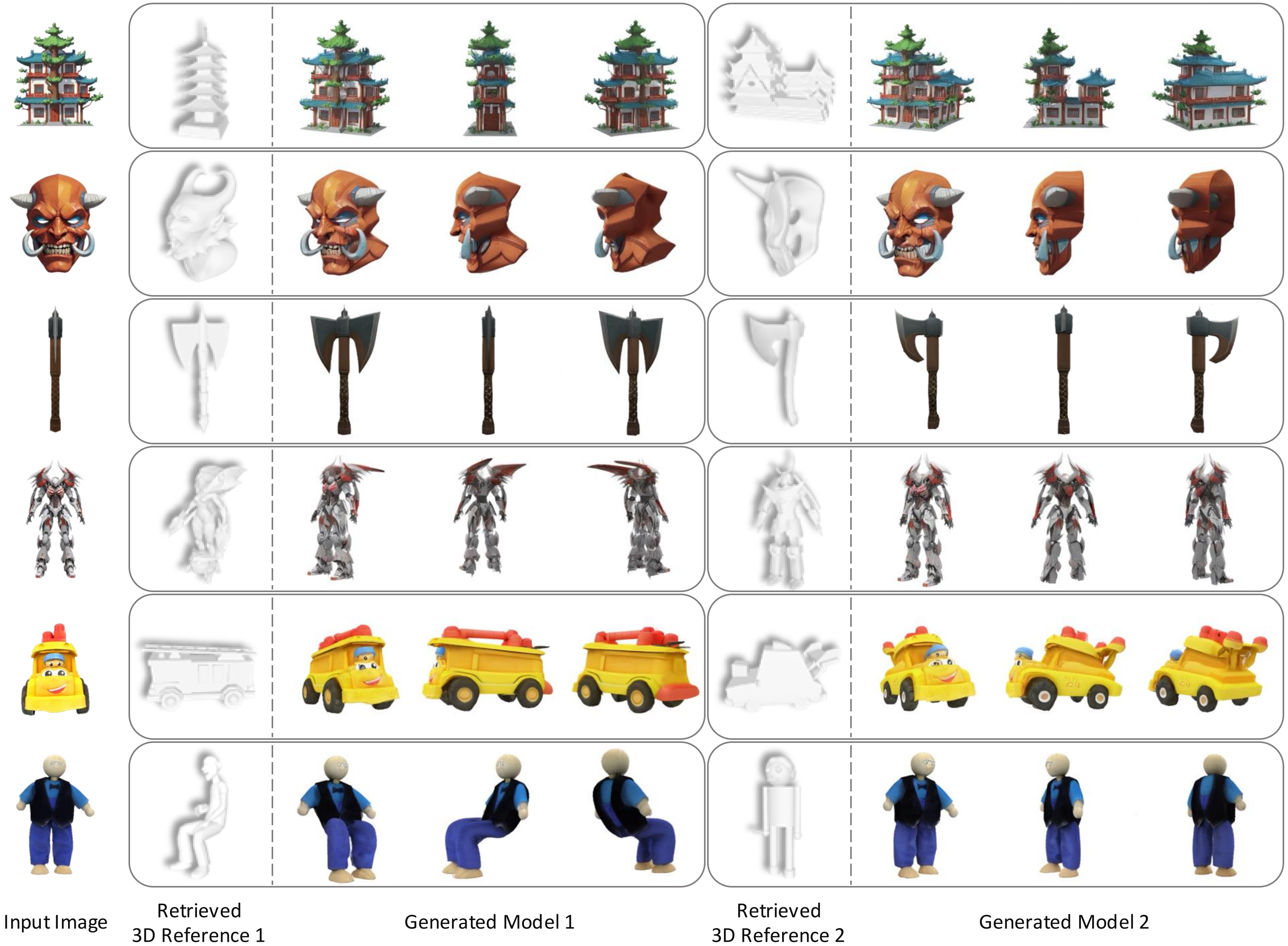


RAG IMAGE TO 3D

Phidias



DIFFERENT 3D REFERENCES

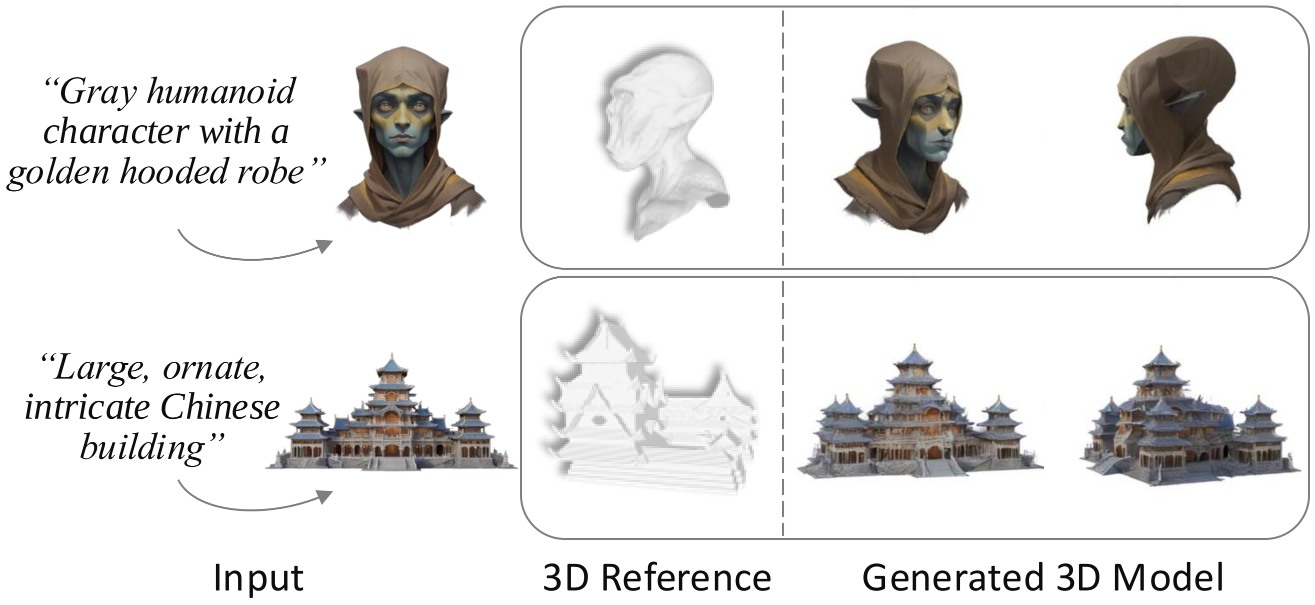


APPLICATIONS

- **Retrieval-Augmented Text-to-3D Generation**
- **Theme-Aware 3D-to-3D Generation**
- **Interactive 3D Generation with Coarse Guidance**
- **High-Fidelity 3D Completion**

APPLICATIONS

- **Retrieval-Augmented Text-to-3D Generation**



“Gray humanoid character with a golden hooded robe”

“Large, ornate, intricate Chinese building”

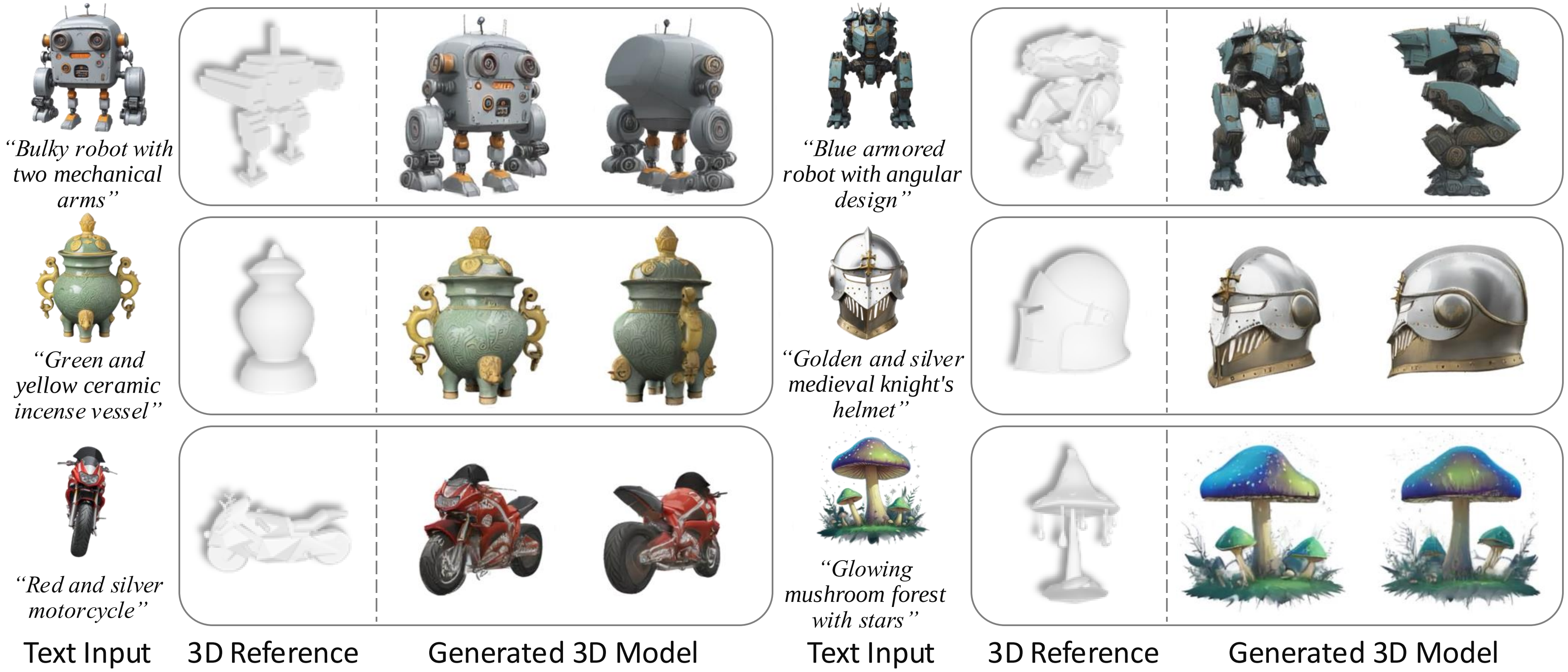
Input

3D Reference

Generated 3D Model

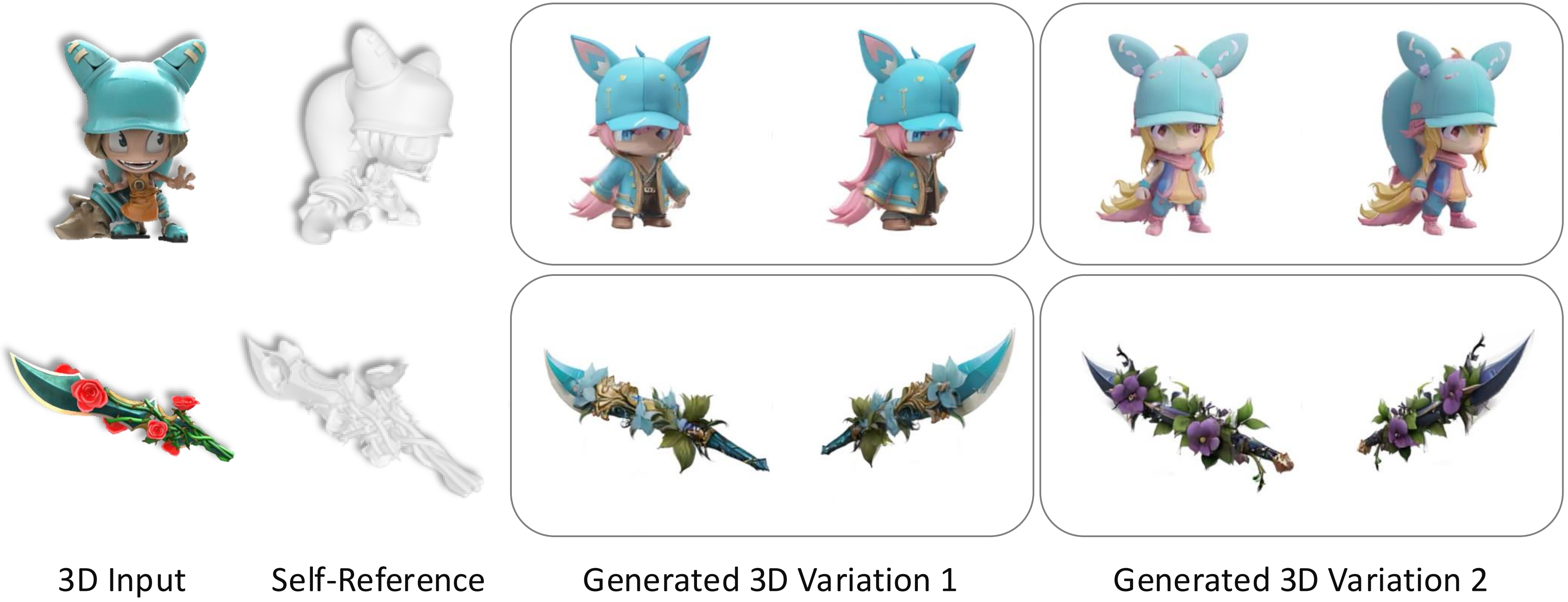
APPLICATIONS

- **Retrieval-Augmented Text-to-3D Generation**


"Bulky robot with two mechanical arms"
"Blue armored robot with angular design"
"Green and yellow ceramic incense vessel"
"Golden and silver medieval knight's helmet"
"Red and silver motorcycle"
"Glowing mushroom forest with stars"
Text Input
3D Reference
Generated 3D Model
Text Input
3D Reference
Generated 3D Model

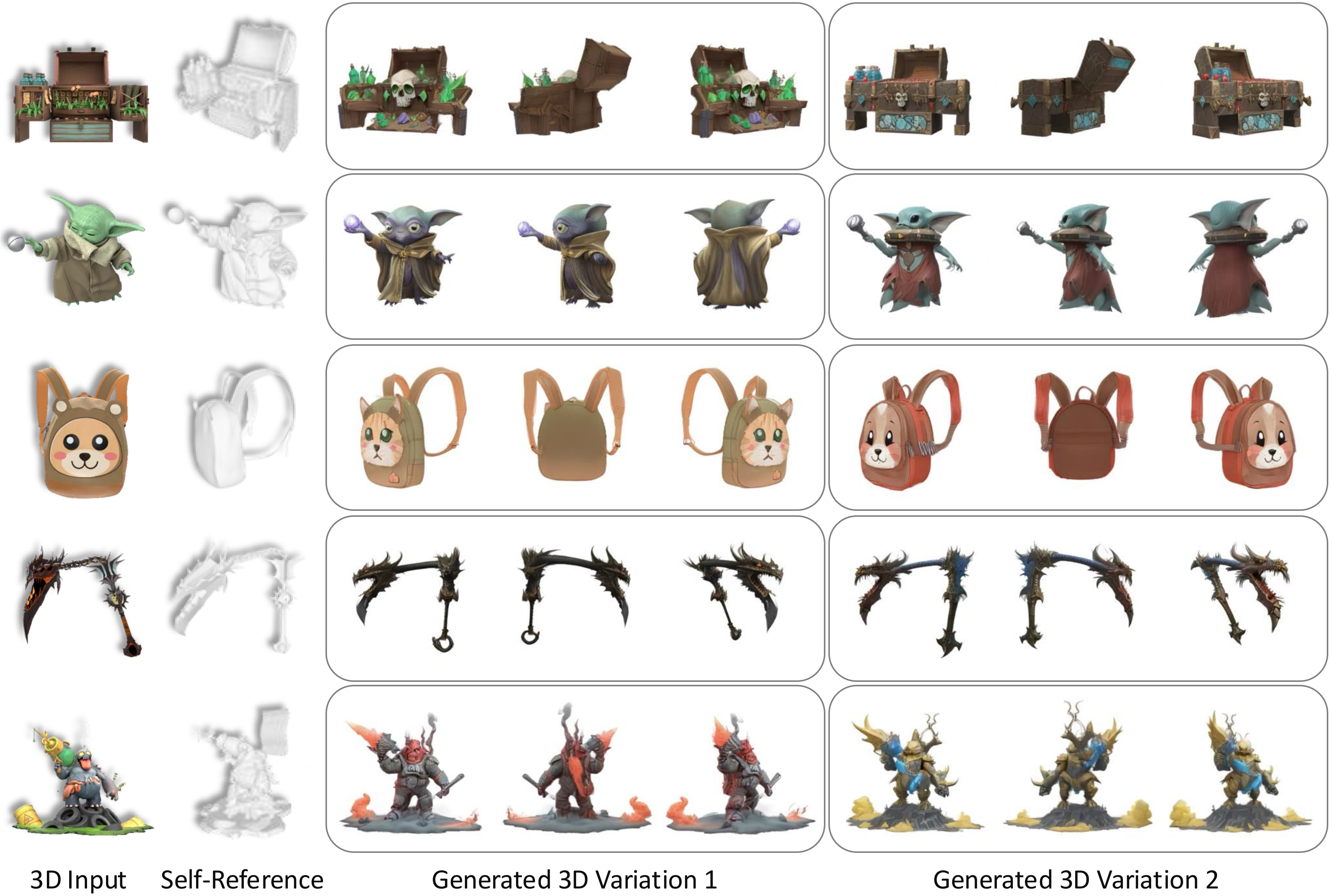
APPLICATIONS

- **Theme-Aware 3D-to-3D Generation**



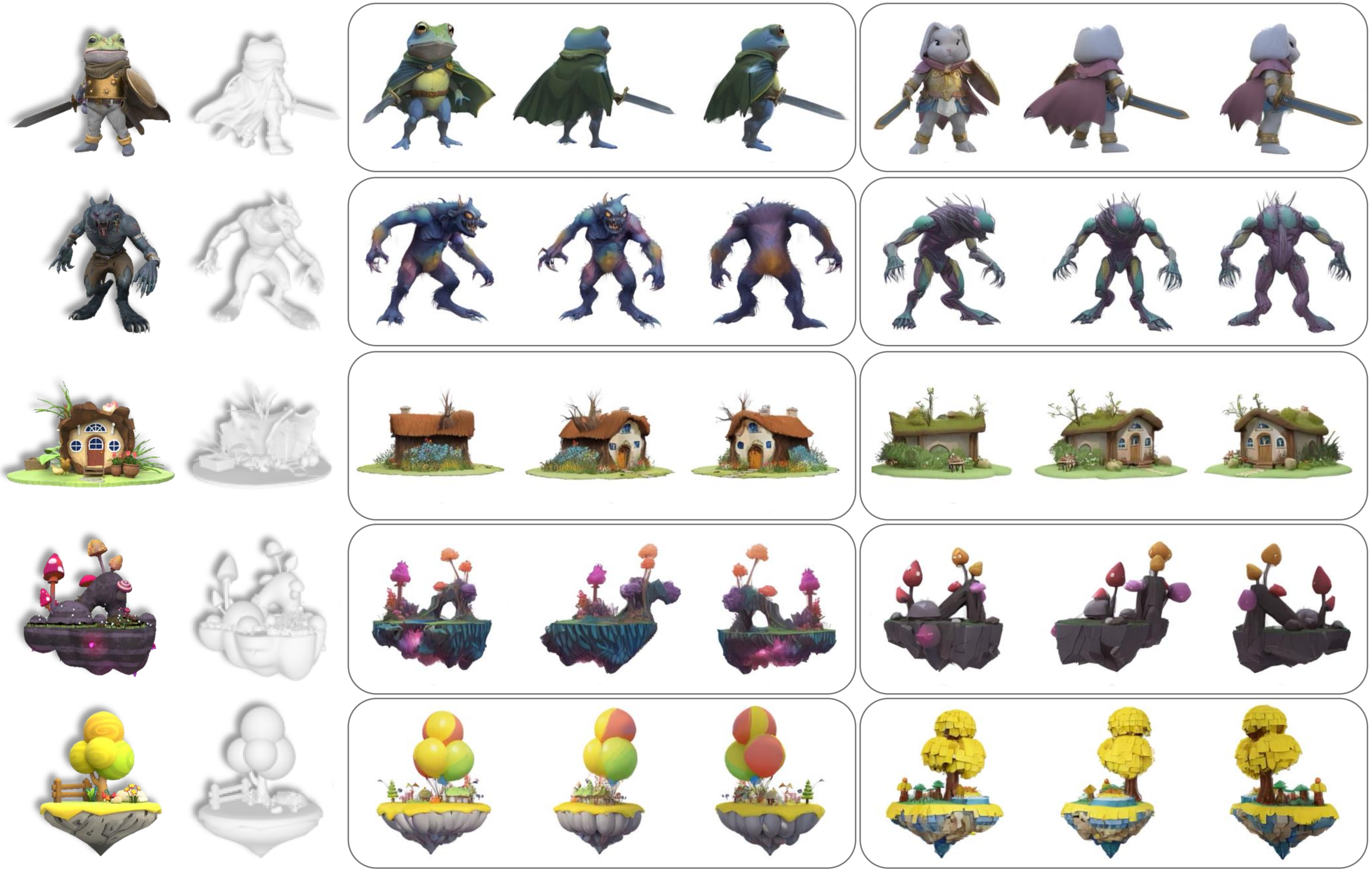
APPLICATIONS

- **Theme-Aware 3D-to-3D Generation**



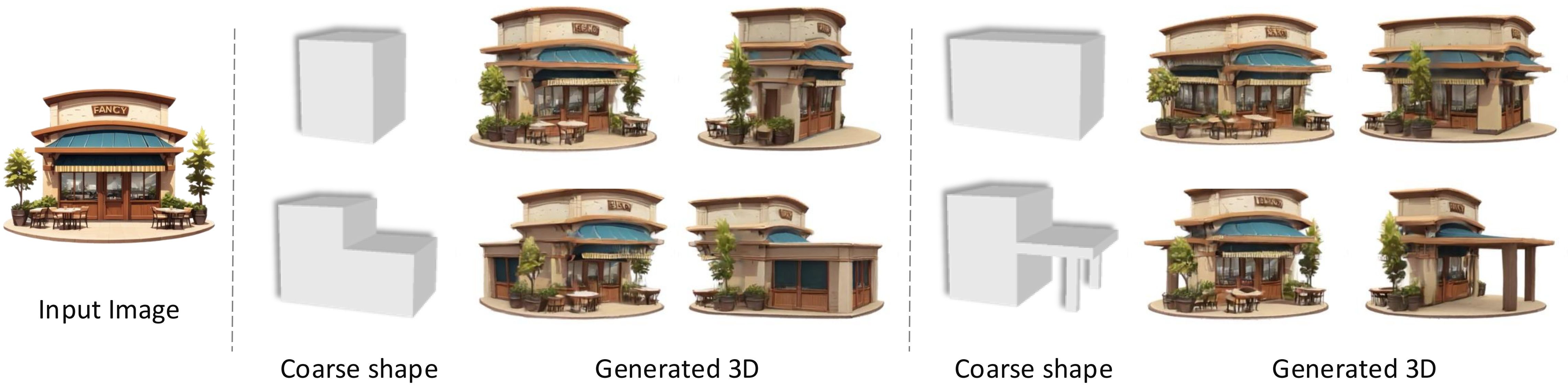
APPLICATIONS

- **Theme-Aware 3D-to-3D Generation**



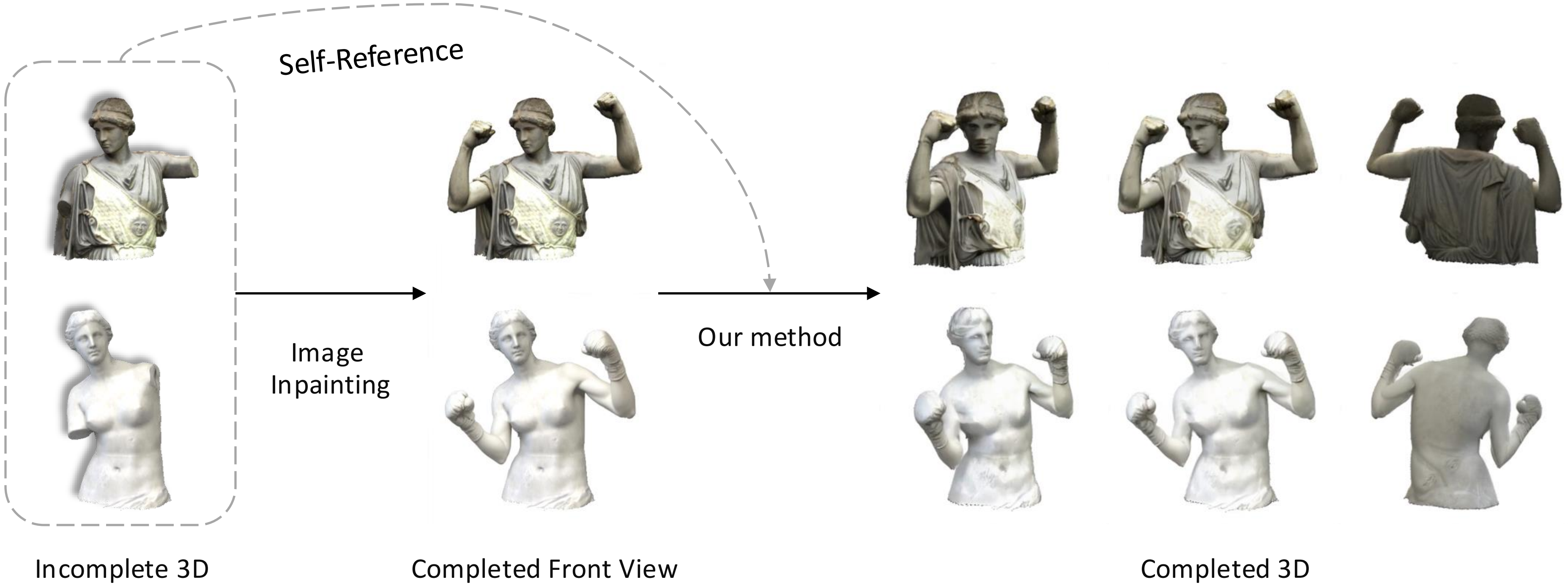
APPLICATIONS

- **Interactive 3D Generation with Coarse Guidance**



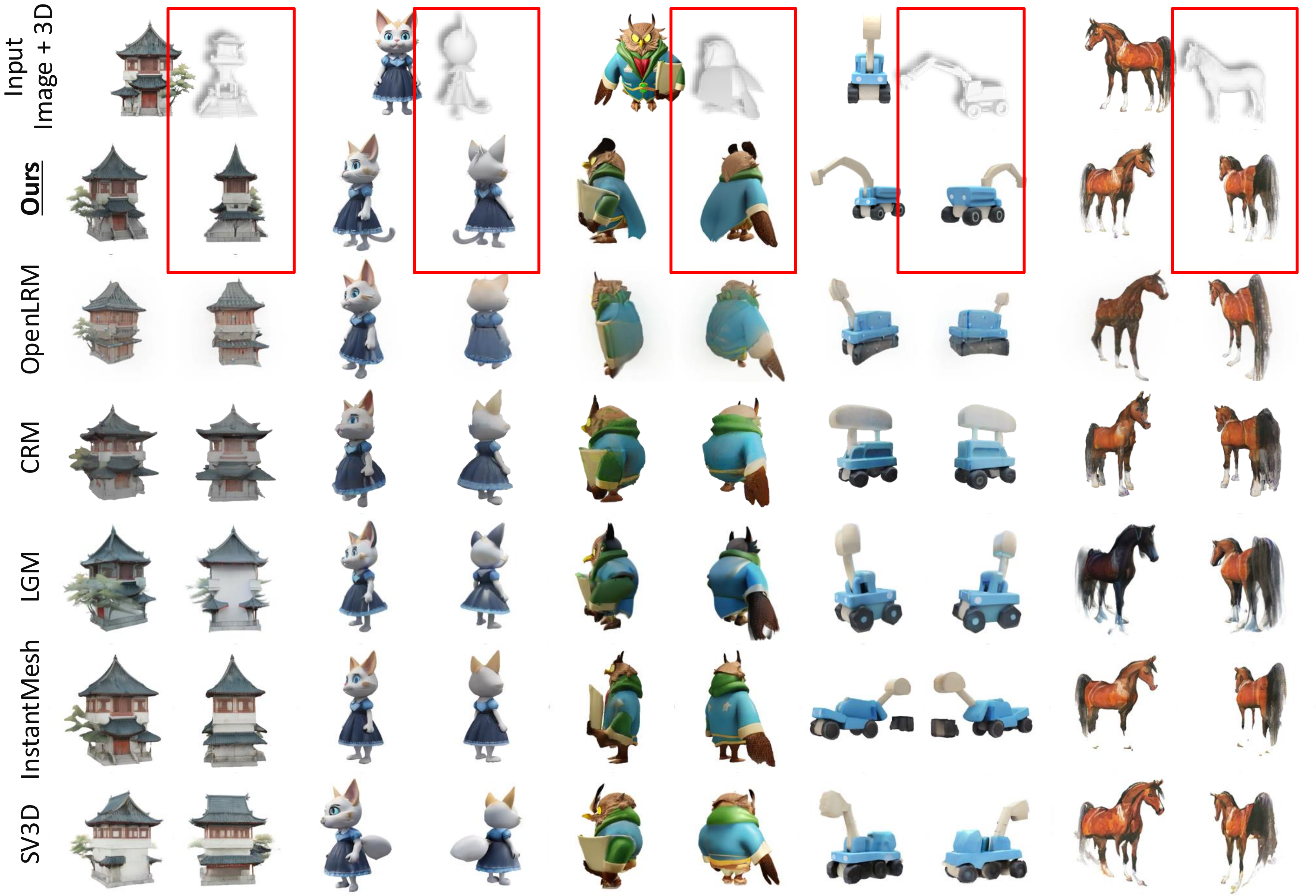
APPLICATIONS

- **High-Fidelity 3D Completion**

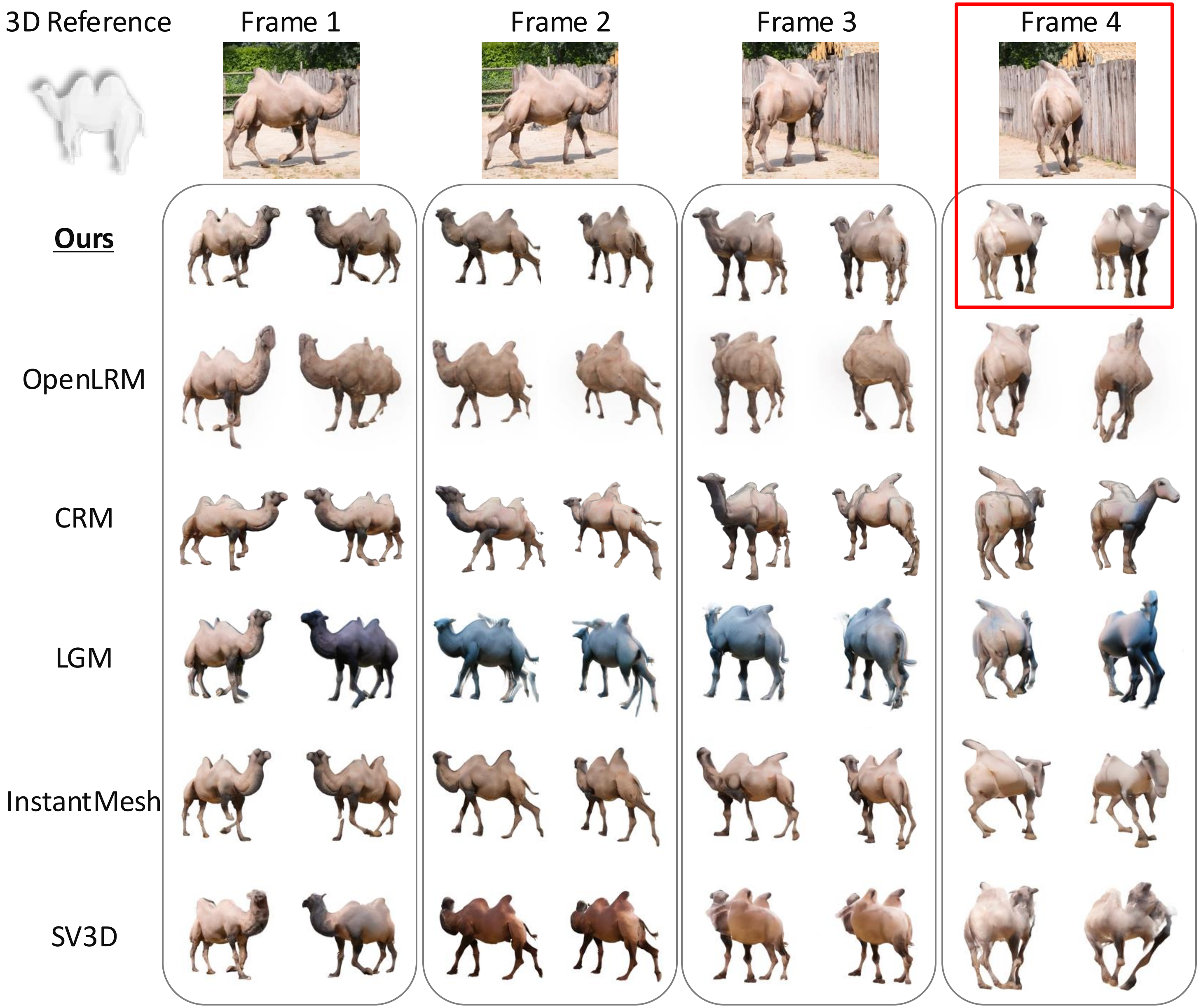


COMPARISONS – IMAGE TO 3D (RETRIEVAL)

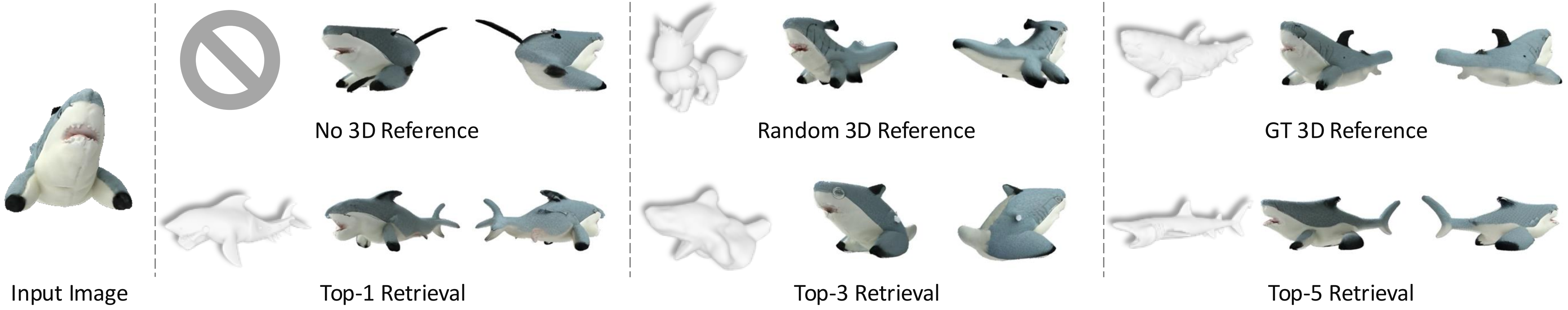
Phidias



COMPARISONS – OTYPICAL INPUT VIEWS



ANALYSIS ON SIMILARITY LEVELS



Input Image

No 3D Reference

Random 3D Reference

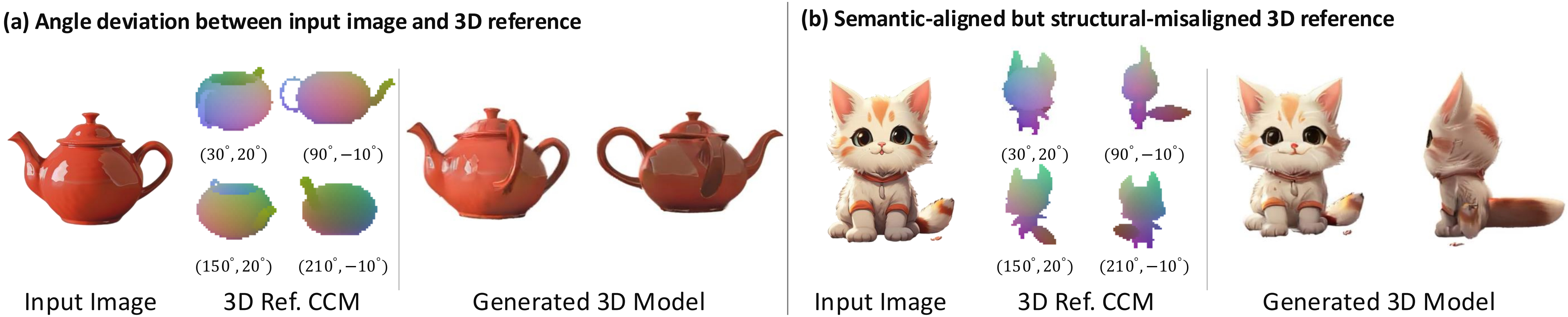
GT 3D Reference

Top-1 Retrieval

Top-3 Retrieval

Top-5 Retrieval

FAILURE CASES



(a) Angle deviation between input image and 3D reference

(b) Semantic-aligned but structural-misaligned 3D reference