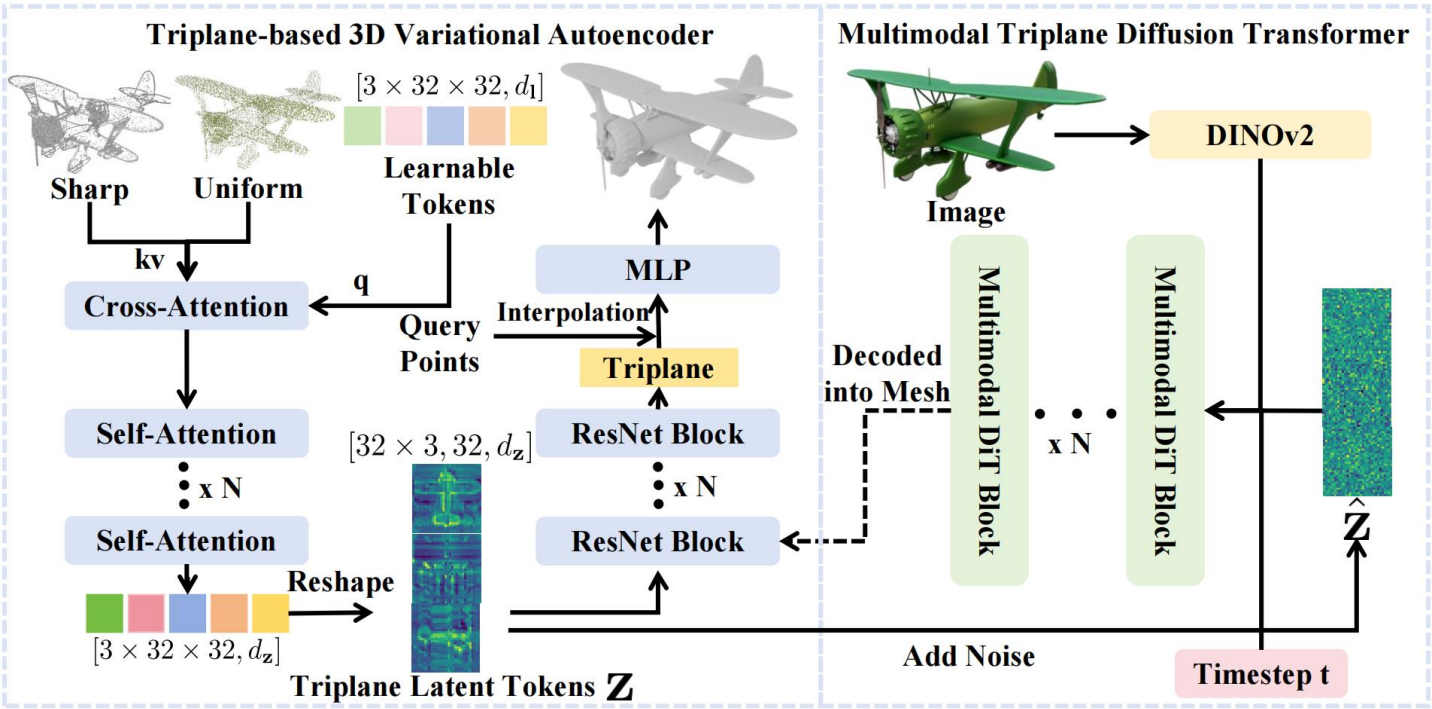
Multimodal Triplane Diffusion Transformer for High Fidelity 3D Shape Generation

Shuang Wu, Youtian Lin, Yifei Zeng, Feihu Zhang, Hao Zhu, Xun Cao, Yao Yao

Frontiers of Computer Science, DOI: 10.1007/s11704-026-50361-3

Problems & Ideas

- Problems of previous 3D generation approaches:
  - Insufficient alignment between generated meshes and input images.
- Ideas: Employ multimodal triplane diffusion transformer to ensure cross-modal alignment between 3D shapes and images, guaranteeing that the generated meshes match the input images in terms of detail.



Main Contributions

- Contributions:
  - A novel native 3D generation framework, comprising a triplane-based 3D VAE and an image-conditioned 3D DiT;
  - A multimodal triplane diffusion transformer, which ensures cross-modal alignment between 3D shapes and images, guaranteeing that the generated meshes match the images in terms of detail.

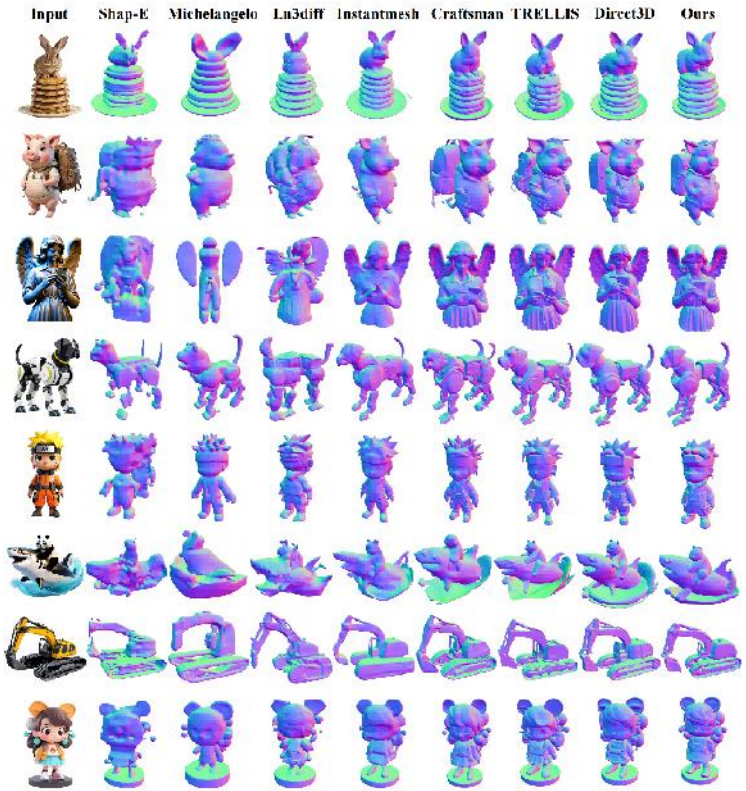


Table 1 The qualitative comparisons of geometry generated by our method and other methods in the image-to-3D task. The best result is denoted in bold, while the second best is underlined.

Methods	ULIP2 ↑	Uni3D ↑
Shap-E [27]	0.0909	0.2212
Michelangelo [20]	0.2177	0.3453
Ln3diff [29]	0.2285	0.3496
InstantMesh [16]	0.2334	0.3823
Craftsman [46]	0.2610	<u>0.4163</u>
Direct3D [24]	0.2746	0.4149
TRELLIS [39]	**0.2894**	0.4142
Ours	<u>0.2826</u>	**0.4225**