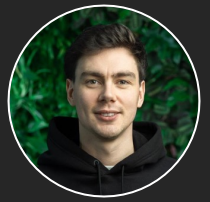
HyperHuman: Hyper-Realistic Human Generation with Latent Structural Diffusion


Xian Liu


Jian Ren


Aliaksandr Siarohin


Ivan Skorokhodov


Yanyu Li


Dahua Lin


Xihui Liu


Ziwei Liu


Sergey Tulyakov

Snap Inc.

香港中文大學 CUHK

香 港 大 學 THE UNIVERSITY OF HONG KONG

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

What is text-to-image (T2I) synthesis?

"Amazon Registry Best Wedding Gift Items"

"Students on the Walk of the Gods located on the Amalfi Coast"

...

"Bespoke cabinets and shelving with oak worktop"

"Metal Bed Frame Queen Instructions"

A dataset of RGB images with text descriptions

training a generator

"Astronaut riding a horse on the moon"

Synthesize **realistic** images that **align with text prompts**

Problems with modern T2I generators on human domain

Modern T2I generators lack structural information for articulated humans with non-rigid deformations, which can hardly be depicted by text prompts.

This makes them struggle to create human images with coherent anatomy like correct number of arms and legs, even with the help of highly-detailed prompts:

Text Prompt:
A realistic digital photograph capturing the dynamic energy of a skilled basketball player in action, as he dribbles the ball with precision and leaps high into the air for a powerful slam dunk, the intensity in his eyes reflecting his passion for the game, 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3.

Text Prompt:
A mesmerizing digital photograph of a ballet dancer in mid-air, captured with impeccable timing and precision, showcasing the grace, strength, and elegance of their movements, 64K resolution, an artistic tribute to the art of dance and the human body in motion.

Problems with modern T2I generators on human domain

Though recent studies like ControlNet [1], T2I-Adapter [2] and HumanSD [3] incorporate structural guidance for controllable generation, they either suffer from inadequate pose control, or are confined to artistic styles of limited diversity.

A pedestrian walks down the snowy street with an umbrella.

pose condition | ControlNet [1] | T2I-Adapter [2] | HumanSD [3]

[1] Zhang et al., "Adding conditional control to text-to-image diffusion models", ICCV 2023
[2] Mou et al., "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models", arXiv preprint arXiv:2302.08453
[3] Ju et al., “Humansd: A native skeleton-guided diffusion model for human image generation”, ICCV 2023

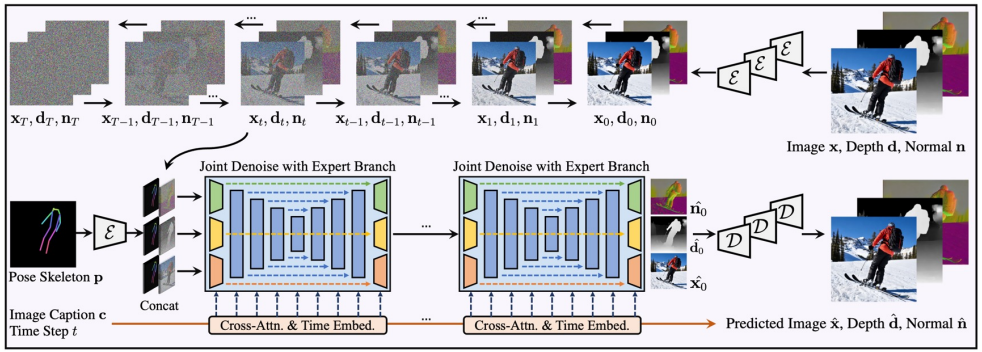
HyperHuman: jointly capture explicit appearance and latent structure

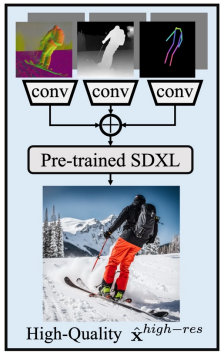
To enable in-the-wild human generation, our key insight:

Human image is inherently structural over multiple granularities, from the coarse-level body skeleton to the fine-grained spatial geometry. Capturing such correlations between the explicit appearance and latent structure in one model is essential to generate coherent human images.

We develop two modules to achieve high realism under diverse scenarios:

1. *Latent Structural Diffusion Model* that jointly learns **appearance, spatial relationship and geometry** in a unified framework
2. *Structure-Guided Refiner* that composes the predicted conditions for **detailed generation**





Latent Structural Diffusion Model: Challenges

- To incorporate pose control, the simplest way is by feature residual or input concatenation.
- However, it's non-trivial to equip the model with structure awareness with several challenges:
  - 1) Sparse keypoints only depict coarse human structure, while the fine-grained geometry and foreground-background relationship are ignored. Besides, naive DM is only supervised by RGB signals, which fails to capture **inherent structural** information.
  - 2) The image RGB and structure representations are spatially aligned but substantially different in latent space. How to jointly model them remains challenging.
  - 3) In contrast to the colorful RGB images, the structure maps are mostly monotonous with similar values in local regions, which are hard to learn by DMs [5].

Unified Model for Simultaneous Denoising

- Simultaneously denoise the depth, surface-normal, and RGB image.
- Choose them as additional learning targets because:
- 1) Depth and normal can be easily annotated for large-scale dataset, which are also used in recent controllable T2I models (*e.g.*, ControlNet and T2I-Adapter).
- 2) As commonly-used guidance, they complement the spatial and geometry information, which are proven beneficial in recent 3D studies (e.g., MonoSDF [4]).

$$\mathcal{L}^{\epsilon\text{-}pred} = \mathbb{E}_{\mathbf{x},\mathbf{d},\mathbf{n},\mathbf{c},\mathbf{p},\boldsymbol{\epsilon},t}\left[\underbrace{||\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_{t_\mathbf{x}};\mathbf{c},\mathbf{p})-\boldsymbol{\epsilon}_\mathbf{x}||_2^2}_{\textit{denoise image } \mathbf{x}} + \underbrace{||\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{d}_{t_\mathbf{d}};\mathbf{c},\mathbf{p})-\boldsymbol{\epsilon}_\mathbf{d}||_2^2}_{\textit{denoise depth } \mathbf{d}} + \underbrace{||\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{n}_{t_\mathbf{n}};\mathbf{c},\mathbf{p})-\boldsymbol{\epsilon}_\mathbf{n}||_2^2}_{\textit{denoise normal } \mathbf{n}}\right]$$

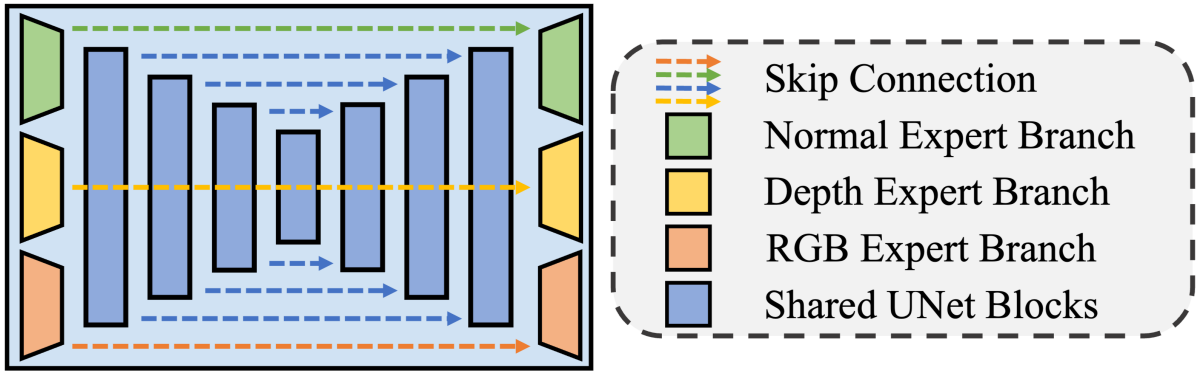
[4] Yu et al., “Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction”, NeurIPS 2022

Latent Structural Diffusion Model: Challenges

- To incorporate pose control, the simplest way is by feature residual or input concatenation.
- However, it's non-trivial to equip the model with structure awareness with several challenges:
  - 1) Sparse keypoints only depict coarse human structure, while the fine-grained geometry and foreground-background relationship are ignored. Besides, naive DM is only supervised by RGB signals, which fails to capture inherent structural information.
  - 2) The image RGB and structure representations are spatially aligned but substantially different in latent space. How to jointly model them remains challenging.
  - 3) In contrast to the colorful RGB images, the structure maps are mostly monotonous with similar values in local regions, which are hard to learn by DMs [5].

Structural Expert Branches with Shared Backbone

- Replicate the first and last several UNet layers as each expert branch.
- # of shared layers trade-off between the spatial alignment and distribution learning:
- 1) More shared layers guarantee the more similar features of final output, leading to the paired texture and structure corresponding to the same image.
- 2) The RGB, depth, and normal can be treated as different views of the same image, where predicting them from the same feature resembles an image-to-image translation task in essence. We need enough separate parameters to do this.



Latent Structural Diffusion Model: Challenges

- To incorporate pose control, the simplest way is by feature residual or input concatenation.
- However, it's non-trivial to equip the model with structure awareness with several challenges:
  - 1) Sparse keypoints only depict coarse human structure, while the fine-grained geometry and foreground-background relationship are ignored. Besides, naive DM is only supervised by RGB signals, which fails to capture inherent structural information.
  - 2) The image RGB and structure representations are spatially aligned but substantially different in latent space. How to jointly model them remains challenging.
  - 3) In contrast to the colorful RGB images, the structure maps are mostly monotonous with similar values in local regions, which are hard to learn by DMs [5].

[5] Lin et al., "Common diffusion noise schedules and sample steps are flawed", arXiv preprint arXiv:2305.08891

Noise Schedule for Joint Learning

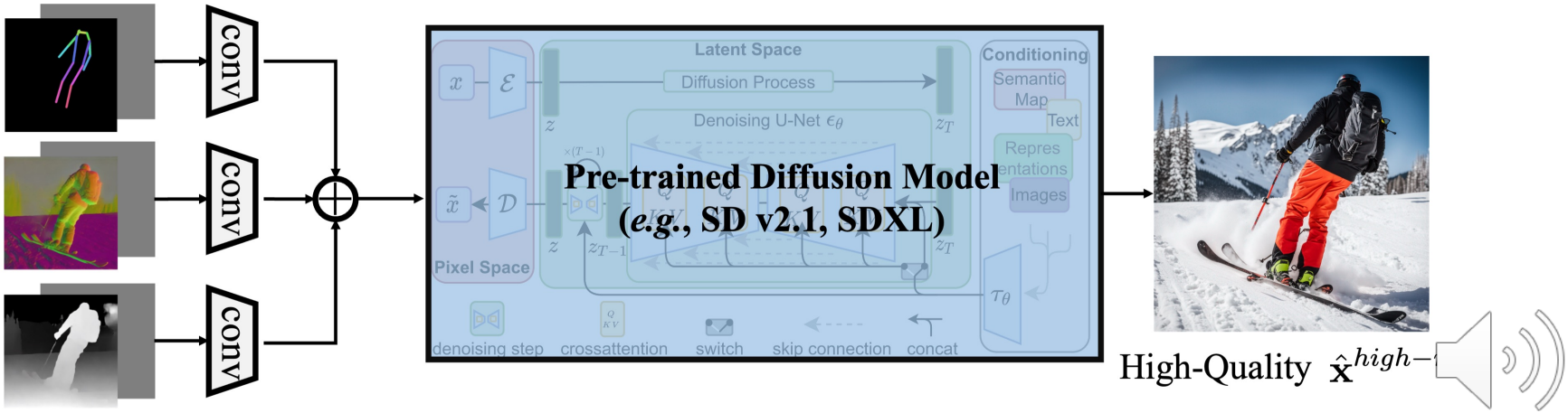
- Normalize the depth and normal latent features to the similar distribution of RGB latent, so that the pre-trained denoising knowledge can be adaptively used.
- Enforce zero-terminal SNR to eliminate structure map's low-frequency information.
- Sample the same noise level for each structural expert branch, so that intermediate features follow the similar distribution when they fuse in the shared backbone, which could better complement to each others
- ***v***-prediction [6] learning target as network objective:

$$\mathcal{L}^{\mathbf{v}\text{-}pred} = \mathbb{E}_{\mathbf{x},\mathbf{d},\mathbf{n},\mathbf{c},\mathbf{p},\mathbf{v},t}\left[||\hat{\mathbf{v}}_{\boldsymbol{\theta}}(\mathbf{x}_t;\mathbf{c},\mathbf{p}) - \mathbf{v}_t^{\mathbf{x}}||_2^2 + ||\hat{\mathbf{v}}_{\boldsymbol{\theta}}(\mathbf{d}_t;\mathbf{c},\mathbf{p}) - \mathbf{v}_t^{\mathbf{d}}||_2^2 + ||\hat{\mathbf{v}}_{\boldsymbol{\theta}}(\mathbf{n}_t;\mathbf{c},\mathbf{p}) - \mathbf{v}_t^{\mathbf{n}}||_2^2\right]$$

[6] Salimans et al., "Progressive distillation for fast sampling of diffusion models", ICLR 2022

Structure-Guided Refiner

- In contrast to previous studies that handle a singular condition per run, we unify multiple control signals at the training phase.
- To bridge the train-test gap caused by error accumulation from the first-stage structure map prediction, we randomly dropout control signals, such as replace text prompt with empty string, or substitute the structural maps with zero-value images.



Results from HyperHuman framework

A man sitting down with a brown teddy bear on his shoulders.

Young man standing near a lake with a snow capped mountain behind.

An elderly woman looks to the side as she sits in front of a cheese pizza in a restaurant.

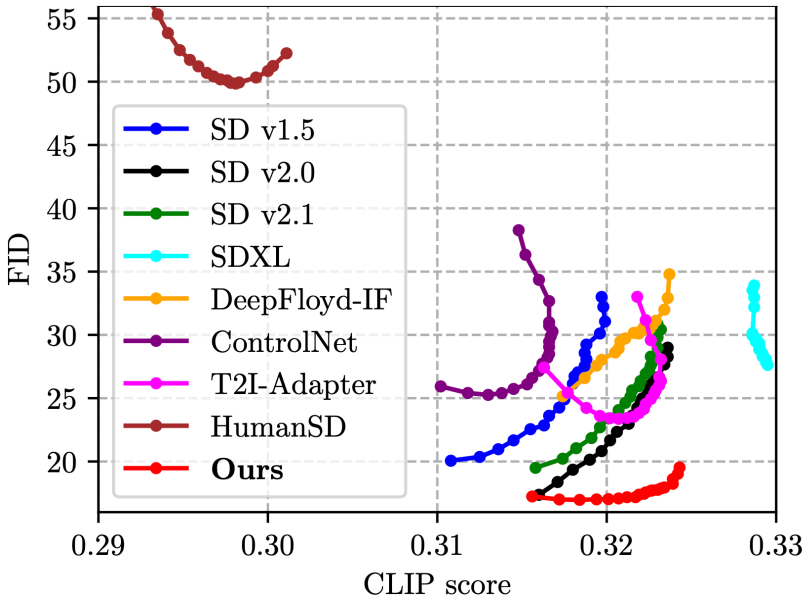
Two women holding surfboards while smiling at the camera.

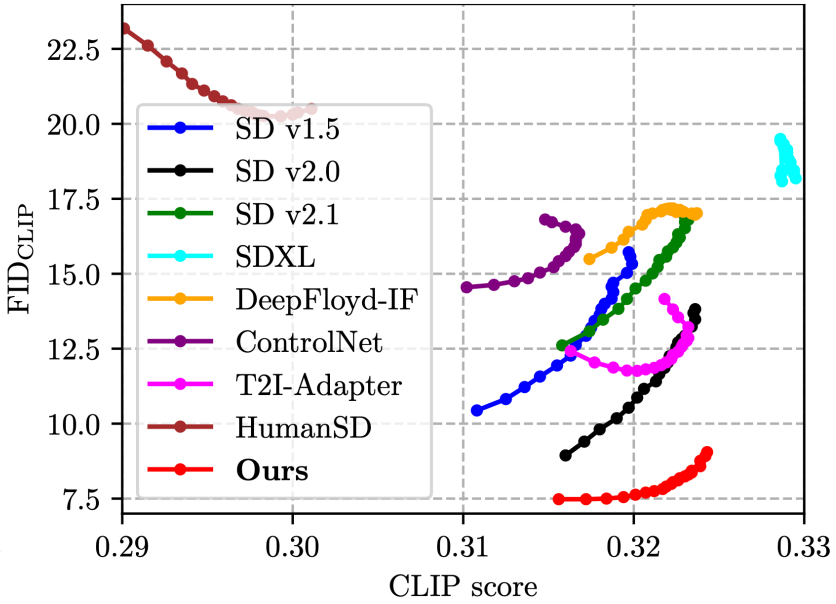
Quantitative Results on Zero-Shot MS-COCO 2014 Val Human

Table 1: **Zero-Shot Evaluation on MS-COCO 2014 Validation Human**. We compare our model with recent SOTA general T2I models (Rombach et al., 2022; Podell et al., 2023; DeepFloyd, 2023) and controllable methods (Zhang & Agrawala, 2023; Mou et al., 2023; Ju et al., 2023b). Note that †SDXL generates artistic style in 512, and ‡IF only creates fixed-size images, we first generate 1024×1024 results, then resize back to 512×512 for these two methods. We bold the **best** and underline the second results for clarity. Our improvements over the second method are shown in red.

	Image Quality			Alignment	Pose Accuracy			
Methods	FID ↓	$\text{KID}_{\times 1k}$ ↓	FID_{CLIP} ↓	CLIP ↑	AP ↑	AR ↑	AP_{clean} ↑	AR_{clean} ↑
SD 1.5	24.26	8.69	12.93	31.72	-	-	-	-
SD 2.0	22.98	9.45	11.41	32.13	-	-	-	-
SD 2.1	24.63	9.52	15.01	32.11	-	-	-	-
SDXL†	29.08	12.16	19.00	**32.90**	-	-	-	-
DeepFloyd-IF‡	29.72	15.27	17.01	32.11	-	-	-	-
ControlNet	27.16	10.29	15.59	31.60	20.46	30.23	25.92	38.67
T2I-Adapter	23.54	7.98	11.95	32.16	27.54	36.62	34.86	46.53
HumanSD	52.49	33.96	21.11	29.48	26.71	36.85	32.84	45.87
HyperHuman	**17.18** 25.2%↓	**4.11** 48.5%↓	**7.82** 31.5%↓	32.17	**30.38**	**37.84**	**38.84**	**48.70**

FID-CLIP and FID_{CLIP}-CLIP Curves w/ CFG scales 4.0-20.0





Qualitative comparisons: user preference result

Table 8: **Detailed Comparion Statistics in User Study.** We conduct a comprehensive user study on zero-shot MS-COCO 2014 validation human subset with well-trained participants.

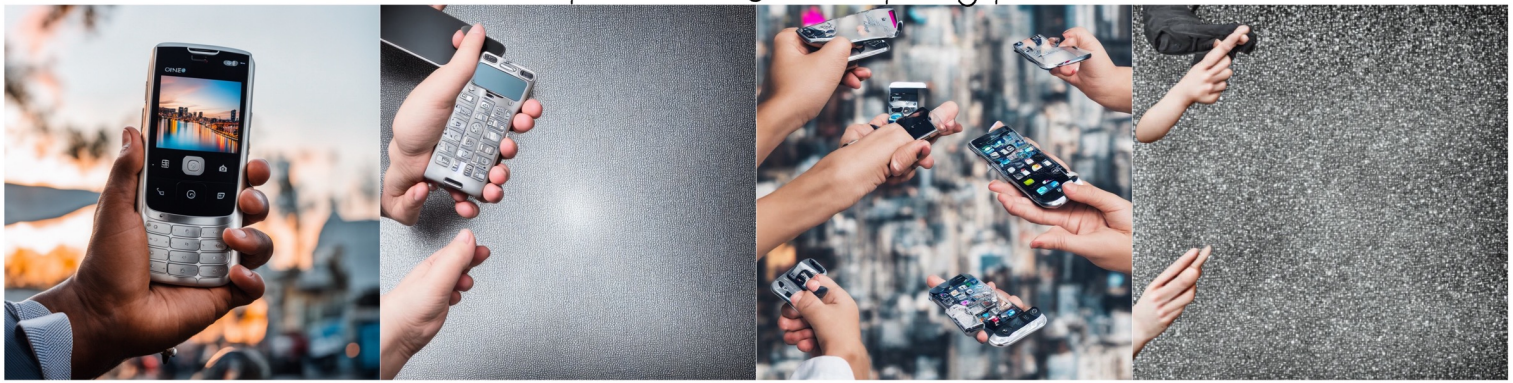
Methods	SD 2.1	SDXL	IF
HyperHuman	**7350** *vs.* 886	**4978.5** *vs.* 3257.5	**6787.5** *vs.* 1444.5

Methods	ControlNet	T2I-Adapter	HumanSD
HyperHuman	**7604** *vs.* 632	**8076** *vs.* 160	**8160** *vs.* 76

Methods	SD 2.1	SDXL	IF	ControlNet	T2I-Adapter	HumanSD
HyperHuman	89.24%	60.45%	82.45%	92.33%	98.06%	99.08%

More qualitative comparison results (1024x1024)

Small silver cell phone being held up any person's hand.



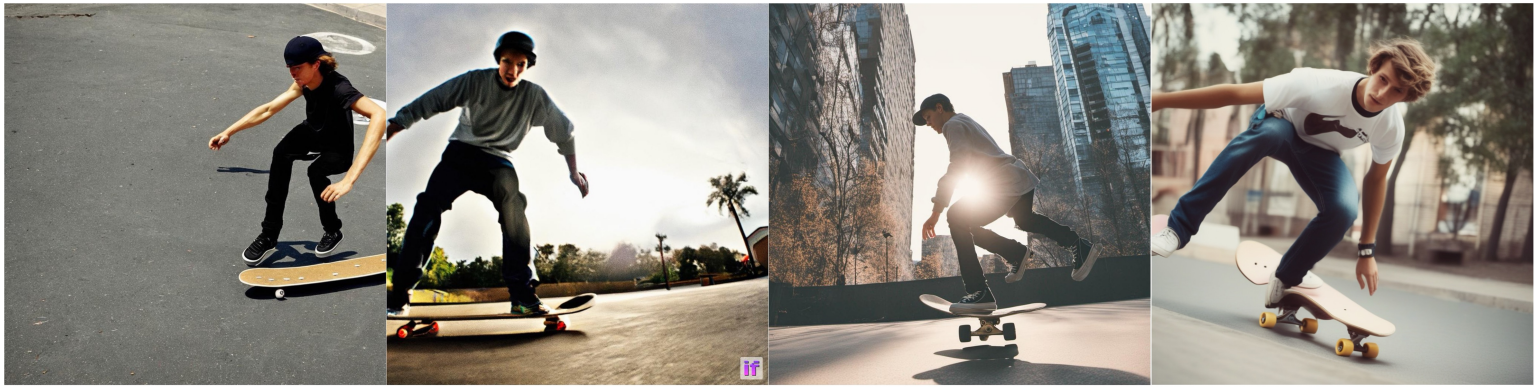
(a) HyperHuman (Ours) (b) ControlNet (c) T2I-Adapter (d) HumanSD

(e) SD v2.1 (f) DeepFloyd-IF (g) SDXL (h) T2I-Adapter+SDXL

More qualitative comparison results (1024x1024)

Mastering the art of skateboarding is profoundly beneficial.

(a) HyperHuman (Ours) (b) ControlNet (c) T2I-Adapter (d) HumanSD



(e) SD v2.1 (f) DeepFloyd-IF (g) SDXL (h) T2I-Adapter+SDXL

More qualitative comparison results (1024x1024)

A group of men who are standing behind a banner that has various flags on it.

(a) HyperHuman (Ours) (b) ControlNet (c) T2I-Adapter (d) HumanSD

(e) SD v2.1 (f) DeepFloyd-IF (g) SDXL (h) T2I-Adapter+SDXL

More qualitative results of HyperHuman (1024x1024)

A baby girl with beautiful blue eyes standing next to a brown teddy bear.

A little girl with wavy hair and smile holding a teddy bear.

More qualitative results of HyperHuman (1024x1024)

A man and woman seated at a table in a restaurant.

A cow laying on the grass behind a man holding a cup of coffee.

More qualitative results of HyperHuman (1024x1024)

A man in a red shirt is holding a skate board up over his head.

Two men who are sitting next to each other with a large pizza in front of them.

More qualitative results of HyperHuman (1024x1024)

A man standing on grassy area next to trees.

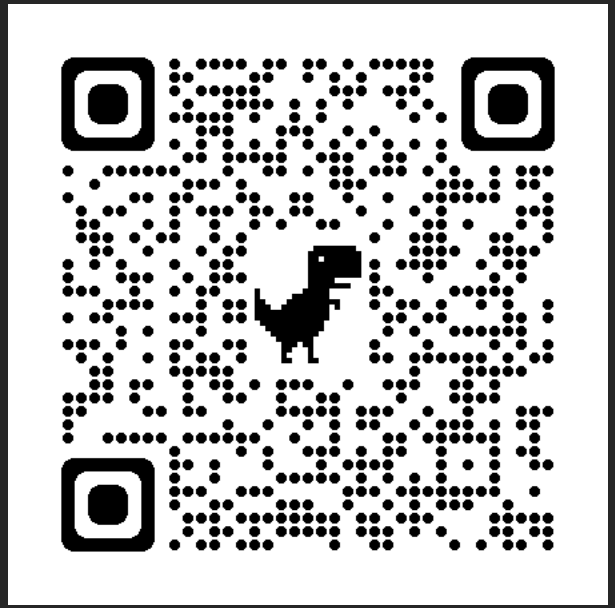
A girl with blue hair is taking a self portrait.

More qualitative results of HyperHuman (1024x1024)

A man wearing a helmet is sitting on his blue motorcycle.

A person dressed up taking a picture at a street with his fist up.

Thank you for watching!



https://snap-research.github.io/HyperHuman/