Video Generation Models are General-Purpose Vision Learners

Letian Wang[1,2], Chuhan Zhang[1], Rishabh Kabra[1,3], Jasper Uijlings[1], Steven Waslander[2], Andrew Zisserman[1,4], Joao Carreira[1], Kaiming He[1,5], Misha Andriluka[1], Eduard Gabriel Bazavan[1], Andrei Zanfir[1], and Cristian Sminchisescu[1,6*]

[1]Google DeepMind [2]University of Toronto [3]University College London
[4]University of Oxford [5]MIT [6]Lund University

Abstract. Driven by next-token prediction, NLP shifted from task-specific models into powerful generalist foundation models. What, then, is the equivalent catalyst needed to achieve a general-purpose model in computer vision? In this paper, we contend that large-scale text-to-video generation serves as a strong pre-training paradigm for computer vision, providing the necessary spatiotemporal priors, vision-language alignment, and scalability required for general visual intelligence. We introduce GenCeption, which leverages a pre-trained video generative diffusion backbone to define a feed-forward perception model, capable of performing various vision tasks steered by text instructions. Empirical results demonstrate that GenCeption achieves state-of-the-art performance across a diverse suite of tasks, including depth, surface normal, and camera pose estimation, expression-referring segmentation, and 3D keypoint prediction, often matching or surpassing specialized models (e.g., DepthAnything V3, SegmentAnything V3). Furthermore, the video generative pretrained backbone outperforms alternative pretraining paradigms (e.g., V-JEPA, and Video MAE) under comparable settings. Importantly, GenCeption exhibits preliminary data and model scaling properties along with exceptional data efficiency where it achieves comparable performance with leading models like D4RT and VGGT-Ω with 7× to 500× less training data. Finally, GenCeption also exhibits intriguing emergent behaviors: a model trained exclusively on synthetic human videos generalizes to real-world footage and out-of-distribution object categories (e.g., animals and robots). These findings suggest that video generation is not merely a synthesis tool, but a foundational path toward generalist vision intelligence for the physical world.

1 Introduction

Natural language processing (NLP) evolved from an era of specialized models—where separate models were developed for each language task (e.g. translation, summarization)—to a singular, unified foundation model paradigm. This paradigm, supported by large-scale next-token prediction pre-training followed by task-aligned post-training, has effectively collapsed thousands of disparate linguistic challenges into a single generalist intelligence, ultimately unlocking emerging behaviors such as chain-of-thought and in-context learning.

*Work done while at Google DeepMind.

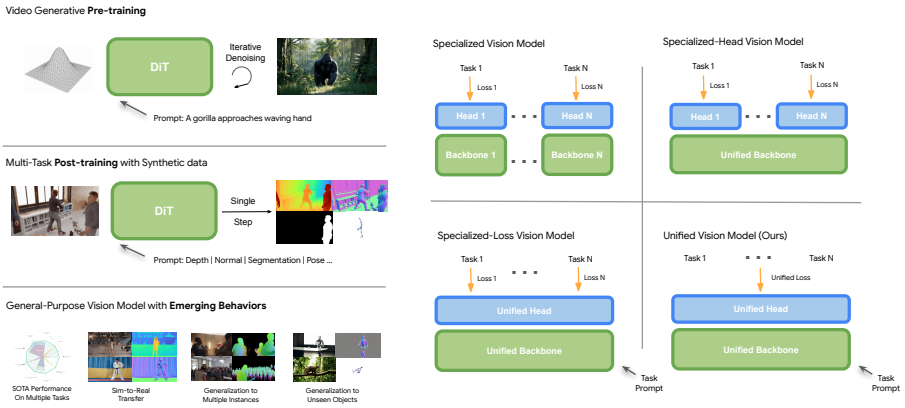


Fig. 1: Methdology (Left): GenCeption treats a video generative diffusion model as a *pre-training* base to capture rich spatio-temporal world priors and native vision-language alignment at scale. During multi-task *post-training*, the model is adapted to feed-forward model fine-tuned on predominantly synthetic data to handle diverse perception tasks. GenCeption shows strong performance with intriguing *Emerging Behaviors*, enabling seamless sim-to-real transfer and generalization to out-of-distribution object categories. **Paradigm Shift** (Right): This highlights a paradigm shift from specialized task-specific computer vision models, to fully unified generalist vision models.

In stark contrast, computer vision is still lingering in its "specialized model" stage. While recent years have yielded powerful foundation models such as Segment Anything series [36, 54] for localization or Depth Anything series [74] for geometry, these remain fundamentally task-specific models that need customized architectures for each task. We have mastered specialized perception, but we have yet to realize a unified vision foundation model—a unified task-agnostic architecture that mirrors LLMs from general pre-training to versatile, emerging intelligence. We posit that the quest for a generalist vision model is essentially a search for a universal pre-training objective—a visual analog to next-token prediction. Such a pre-training paradigm should satisfy three core imperatives:

1) *Spatio-Temporal Evolution*: The world is a 4D continuum. The pre-training objective should force the model to internalize the 4D temporal causality and physics of a world in motion.
2) *Vision-Language Alignment*: To inherit the instruction-following capabilities and common-sense knowledge of language models, visual features should be natively aligned with linguistic semantics during both pre-training and post-training stages.
3) *Scaled up*: The paradigm should be scaled in both data and compute, ultimately enabling the emergence of vision intelligence.

In this paper, we demonstrate that large-scale text-to-video generation serves as this pivotal pre-training paradigm, uniquely satisfying all three requirements, compared to prior vision pretraining methods (e.g. contrastive learning, masked autoencoder). First, the objective of generating high-fidelity video sequences

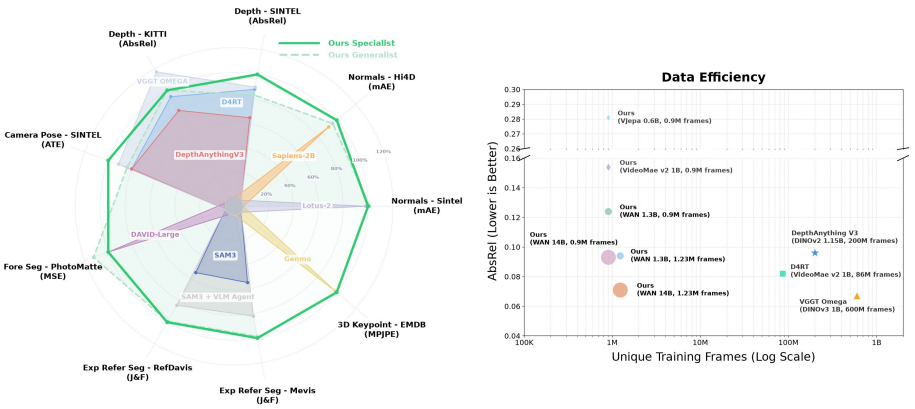


Fig. 2: Generalist Capability (Left): GenCeption achieves competitive performance, matching or outperforming state-of-the-art models dedicated to individual tasks (e.g. DepthAnything3 [41], SAM3 [11], Sapiens [33], David [57], Genmo [38], Lotus-2 [22]). **Data Efficiency in Finetuning** (Right): Validated on depth estimation, the video generative pretrained backbone (i) outperforms the largest available variants alternative pretraining paradigms (e.g., V-JEPA, and VideoMAE V2) under the same finetuning data. (ii) exhibits preliminary scaling properties, where the performance improves with more data and large model size; (iii) shows exceptional data efficiency, achieving comparable performance with leading models like D4RT [79] and VGGT-Ω [65] with 7× to 500× less training data.

forces a model to internalize the spatio-temporal world priors, including 3D geometry, object permanence, and physical interactions. Second, as these models are natively conditioned on text, they provide the necessary vision-language alignment. Finally, modern text-to-video generation models have been scaled using massive datasets and compute due to their commercial value and relatively low annotation cost, providing a direct path toward emergent behaviors.

We introduce GenCeption, a generalist video perception model that instantiate this paradigm. As illustrated in Figure 1, we treat the pre-trained video diffusion backbone as our "base model", leveraging the rich priors learned during generative pre-training. We then apply the post-training phase, where we fine-tune the model across diverse visual tasks primarily with synthetic data. This enables GenCeption to execute heterogeneous perception tasks within a single task-agnostic architecture, ranging from pixel-level segmentation to more complex 3D pose estimation and tracking. Specifically, by repurposing iterative diffusion into a feed-forward architecture, GenCeption dynamically infers the intended modality in a single forward pass, steered by the text instruction. Notably, GenCeption promotes a paradigm shift from specialized vision models to unified generalist intelligence that adopt unified backbone, head, and loss function across different vision tasks. Under this paradigm, task specifications are shifted from architectural modifications to data format designs, which enables scaling up with an ever-growing spectrum of visual tasks.

As shown in Figure 2, our empirical results demonstrate that GenCeption achieves state-of-the-art performance across multiple perception tasks, perform-

ing on par with or even surpassing specialized models designed for single modalities, including DepthAnything V3 [41], SegmentAnything V3 (SAM3) [11], D4RT [79], VGGT-Ω [65], Sapiens [33], David [57], and Genmo [38]. Under comparable model size and same fine-tuning data, the video generative pretrained backbone outperforms the largest available variants of alternative pretraining paradigms (e.g., V-JEPA [4], VideoMAE V2 [61]). Importantly, GenCeption exhibits preliminary scaling properties where the performance improves with more data and larger model size, along with exceptional data efficiency where it achieves comparable performance with leading models like D4RT and VGGT-Ω with 7$\times$ to 500$\times$ less training data. Finally, this paradigm triggers emergent behaviors, including seamless sim-to-real transfer and out-of-distribution generalization to unseen object categories, as shown in Figure 8b and Figure 8a. This evidence suggests that video generation is not merely a synthesis tool, but the foundational path toward generalist vision intelligence for the physical world.

2 Related Work

2.1 Perception Foundation Models.

Perception foundation models have recently attracted significant attention, with frameworks such as Segment Anything series [11,36,54] and Depth Anything series [41,73,74] leveraging massive datasets to achieve robust performance across diverse visual scenarios. Parallel efforts have sought to develop unified models capable of handling multiple tasks simultaneously. However, these works operate primarily in the image domain [2, 48, 57], lacking an inherent understanding of temporal dynamics. Furthermore, multitasking in these models is often implemented at a rigid architectural level—relying on task-specific encoder or decoder [57]—which lacks the flexibility required for truly general-purpose systems. Even among models that attempt to process video and flexibly handle varied tasks [26, 81], they have not yet reached the level of efficacy seen in specialized, task-specific models when evaluated on a wide range of video tasks. This stands in stark contrast to Large Language Models (LLMs), which have demonstrated that a single unified architecture can flexibly master a vast array of tasks with human-level or even superhuman proficiency. A truly unified video perception model that matches the general-purpose proficiency of LLMs has yet to be realized.

2.2 Visual Representation Learning

At the heart of this challenge lies the long-standing task of representation learning: learning rich and robust visual representations to enable a wide range of vision tasks. In the literature, unsupervised and self-supervised learning have become core paradigms for representation learning, as they scale efficiently with large amounts of unlabeled data. Among representative approaches, masked autoencoders [23,50] (MAE), inspired by masked language modeling [9,14], learn to reconstruct missing image regions and have established a foundational framework

for visual pretraining, where the learned representations can be effectively fine-tuned to various downstream vision tasks. While extensions like VideoMAE [61] and RVM [82] have translated this to the video domain, these methods often lack explicit multimodal vision-language alignment, lacking the semantic grounding required to link visual patterns to high-level concepts. Another prominent paradigm, contrastive learning, aims to align semantically similar samples while separating dissimilar ones. By performing contrastive learning across vision and language modalities, models such as CLIP [52] and SigLip [78] learn powerful multimodal representations that unlock a wide range of tasks such as open-vocabulary classification and segmentation. Despite the vibrancy of this field, multimodal representation learning at the video level remains a crucial problem, presenting numerous challenges, including, among other reasons, the high computational cost of training and the added complexity of modeling temporal dynamics.

2.3 Re-purposing Diffusion Models.

Recently, an emerging direction is to leverage the rich features learned in pre-trained diffusion models, a path which our work follows. Early work in this area focused mainly on single-image prediction. *Marigold* [30] demonstrated how a pre-trained image diffusion model like Stable Diffusion [56] can be repurposed through fine-tuning on synthetic data to act as a monocular depth estimator. [46] improves the efficiency of prior diffusion-based depth estimators [17, 30] by aligning the time step with the noise level, showing that a simple end-to-end fine-tuning approach can create a single-step, deterministic model that is both fast and highly accurate. *GenPercept* [70] conducts a systematic study and identifies key components such as feature injection, decoding mechanisms, and training objectives, that are critical for successfully adapting pre-trained diffusion models to dense perception tasks. *Diception* [80] demonstrates strong vision generalists capable of addressing multiple tasks, steering perception tasks using text prompts.

An inherent limitation of these single-image models, however, is their inability to produce temporally consistent predictions when applied to video sequences. To address this, *BufferAnytime* [37] proposes to adapt the existing image diffusion model to the video domain by incorporating a temporal layer for depth and normal estimation problems. Going further, various works have explored directly employing native video diffusion models to address the temporal consistency issues. *DepthCrafter* [25] and *NormalCrafter* [6] incorporate alignment objectives with large-scale Vision Foundation Models, such as CLIP [52] or DINO [49], to facilitate training. [72] explores training with scalable synthetic data to overcome the scarcity of video ground truth data. *Geo4D* [28] adapts video diffusion models for 4D scene reconstruction. *DiffusionRenderer* [40] leverage video diffusion priors to tackle the dual problems of inverse rendering and forward rendering. *ReferEverything* [3] repurpose video diffusion models for expression-referring open-world segmentation. [1] provides related evidence that video diffusion models encode reusable visual priors, demonstrating few-shot LoRA adaptation to diverse tasks via input-output video transition. However, these advanced meth-

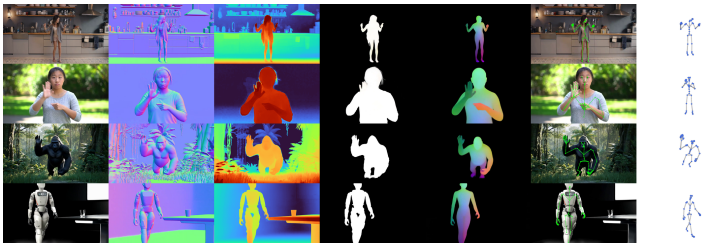


Fig. 3: Capabilities Illustrations. GenCeption is a versatile video perception model with SOTA performance on a multitude of output modalities including surface normal estimation, depth estimation, foreground segmentation, expression-referring open-world segmentation, dense human pose estimation, 2D/3D keypoints estimation, and camera pose estimation. Our approach has been trained on human-centric synthetic videos, yet it generalizes to real videos both for people as well as other categories such as animals and anthropomorphic characters.

ods primarily concentrate on a single, often dense, vision task. We present a unified architecture for both dense and sparse video perception, leveraging the rich prior knowledge and instruction-following strengths of text-to-video models to steer diverse tasks via text prompts. Additionally, we replace slow iterative sampling with an efficient, direct feed-forward architecture, with superior inference efficiency and performance that matches or exceeds SOTA specialists. This work advances our previous human-centric model, THFM [66], into a unified general-purpose model, with broader task and benchmark evaluations, and more exhaustive analysis of model behaviors.

Notably, our work shares close conceptual ties with two concurrent works, while differing in both architecture and evaluation. First, both our work and [69] share the hypothesis that large-scale video generative models learn powerful visual priors. If [69] demonstrates the existence of these priors, we provide the architecture to efficiently harness them. Specifically, while [69] investigates a training-free prompting approach via multi-step video generation with a focus primarily on qualitative validation, we introduces a dedicated post-training strategy and conduct rigorous quantitative evaluation across standardized benchmarks. Second, while *Vision Banana* [18] operates within the 2D image space, GenCeption extends the paradigm to the native video domain to naturally capture temporal consistency. Additionally, instead of focusing on multi-step generation, our method adopts a efficient feed-forward architecture with improved inference efficiency and strong performance across a broader task spectrum. See Figure 3 for a summary of capabilities.

3 GenCeption

3.1 Methodology

The core premise of GenCeption is that large-scale video generation models are not merely synthesis engines, but powerful, universal visual representation learners. Inspired by the trajectory of Large Language Models (LLMs)—where generative pre-training (next-token prediction) inherently builds rich features that

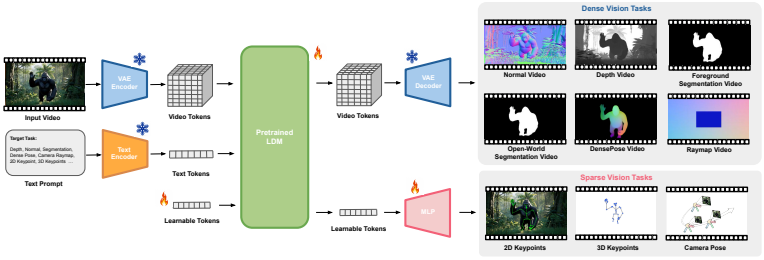


Fig. 4: Architecture overview of GenCeption, a simple yet powerful architecture adapted from text-to-video diffusion models. Given an input video and a text prompt specifying the desired output, our unified model, trained majorly on synthetic data, is capable of performing a wide range of dense and sparse perception tasks, with a single forward-pass of the model. The dense vision tasks are unified in the RGB ambient space where supervision can be applied in latent space efficiently, and the sparse vision tasks are realized by adding learnable tokens as additional inputs to the diffusion transformer (DiT).

can be harnessed for diverse downstream tasks—we reformulate video perception as a post-training endeavor built atop a pretrained video generative model. Our methodology is driven by three foundational principles:

1) *Generative Pre-training as Representation Learning*: We utilize a pre-trained text-to-video generative diffusion model as our universal backbone. The objective of generating high-fidelity video inherently forces the model to internalize robust spatiotemporal priors, 3D geometry, and physical laws, serving as a powerful alternative to traditional contrastive or masked autoencoder pre-training. To fully preserve these foundational representations, one core design principle is to maximize compliance with the original pre-training regime with minimal modifications.
2) *Task-Agnostic Post-Training in Unified Architecture*: To transition from a specialized generator to a generalist perceiver, we treat diverse vision tasks as unified sequence-to-sequence mapping problems, rather than engineering task-specific architectures or relying on specialized encoders and decoders. By formatting various target modalities into a shared representational space (e.g., standard RGB ambient space for dense tasks), we allow a single, unified architecture to master multiple tasks guided seamlessly by text instructions. Under this paradigm, integrating new visual capabilities no longer necessitates altering the architecture or training pipelines; instead, task specifications are directly reflected in the data representation—a data-driven approach that enables scaling up efficiently across a continually expanding range of visual tasks.
3) *Feed-Forward Reformulation*: While diffusion models typically rely on slow, iterative sampling processes for synthesizing diverse outputs, perception tasks demand highly accurate and efficient predictions. Therefore, a critical design decision in GenCeption is the transformation of the multi-step generative backbone into a single-step, feed-forward perception engine.

Guided by these principles, we instantiate GenCeption, a model capable of various video perception tasks. As illustrated in Figure 4, we repurpose text-to-video generation models to enable both dense vision tasks (normal estimation, depth estimation, foreground segmentation, expression referring open-world segmentation, dense human semantic estimation, raymap estimation) and sparse vision tasks (2D and 3D keypoint prediction, camera pose estimation) within a unified architecture. Given the input RGB video and the text prompt specifying the target modality, GenCeption is able to produce target video in a single forward pass (Sec 3.2), with a unified representation for various perception tasks (Sec 3.3); A scalable data synthesis strategy is proposed to enable efficient low-cost collection of diverse high-quality data, which is described in Sec 3.4. We then discuss our training recipe in Sec 3.5.

3.2 From Diffusion Pre-training to Perception Finetuning

Preliminary. Our model is based on text-to-video diffusion model, with its core components being a VAE encoder-decoder pair, a text encoder, and a transformer-based latent diffusion model (DiT). The original DiT model takes a latent token initialized with Gaussian noise and text prompt as inputs and conducts a number of denoising steps, gradually transforming it into latent tokens, which can be decoded to the generated video. The index of the denoising step is provided to DiT as additional input to modulate the predictions with respect to amount of input noise. Let us denote the video token inputs to DiT at step t as x_t, with x_0 being the noise-free latents of the original video. Depending on the particular formulation, DiT module is trained to predict the noise component ϵ as in [24, 60], the velocity $v = \epsilon - x_0$ between the noise and original input as in the "Rectified flow" variant [16, 44, 45], or the target x_0 directly [39].

Feedforward Perception Formulation. In this work, we adapt the multi-step diffusion model to a feed-forward prediction model. To that end we directly use the latent tokens encoded from original video as the inputs to DiT and conduct only a single forward-pass of the model. The input time-step is set to the time step used at the end of multi-step diffusion process, to indicate that the input is noise free. We rely on DiT trained with "Rectified flow" objective, and negate the output of the DiT prior to passing it to the loss functions or decoder. The negated output $-v = x_0 - \epsilon$ is then close to the latent encoding of the RGB video, which, in our experiments, helps the model to converge faster.

3.3 Unified Task Representation

Unify Dense Tasks in RGB Space. Unlike prior practices that often adopt separate backbones and decoders for different tasks, our framework promotes a unified approach, where all tasks share the same architecture and weights for both backbone and decoder, and the target task is simply modulated via the text prompt. To achieve this, the outputs for all dense perception tasks are represented in the 3-channel RGB space within the range $[0, 1]$. Specifically,

the three RGB channels are replicated for single-dimensional outputs like depth and segmentation, whereas they represent distinct spatial dimensions for three-dimensional tasks including surface normals and DensePose estimation. Importantly, this representation is highly generalizable, and many high-dimensional vision modalities can be elegantly adapted to fit within this 3-channel constraint. For instance, to estimate camera poses—which are often represented as extrinsics and intrinsics matrices, we utilize a pixel-space raymap to closely align with the format of video generative pre-training, see Figure 5. To further accommodate the 6-channel nature of the camera raymap, we spatially arrange the raymap by allocating the ray origin to the central region and the ray direction to the outer boundaries of the frame. This design preserves the integrity of our single-decoder framework while maximizing the leverage of powerful pre-trained visual priors. We believe this design philosophy scales naturally to a wider spectrum of vision tasks. Analogous to the prevailing practice of reformatting non-text data—such as bounding boxes or robot control tokens—into text strings to leverage the power of LLMs, we advocate for reformatting vision-friendly tasks directly into the continuous pixel space. Instead of forcing inherently visual modalities into text to fit language models, we promote a return to the pixel space—especially for dense, high-dimensional tasks that are fundamentally visual.

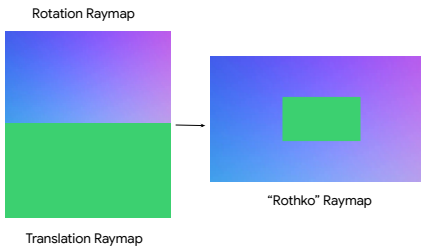


Fig. 5: The "Rothko" Raymap as an example of adapting high-dimensional modal data into standard 3 RGB channels.

Enable Sparse Tasks with Learnable Tokens. While RGB is versatile, downstream tasks like robotic control often require directly manipulatable, structured outputs such as explicit 2D/3D coordinates. To enable these sparse predictions, we introduce a lightweight token-based extension. We append T learnable tokens—one for each of the T video frames—to the video latents. After passing through the model, an MLP decodes each token to predict a K-dimensional target per frame. To maintain compliance with the base DiT model, we utilize its native 3D RoPE. Because the spatial positions of these additional tokens are unknown, we make them learnable. Although the temporal positions are known, the video frame count T exceeds the compressed latent length T'. To stay within the temporal bounds seen during training, we down-scale these positions using position interpolation [13]. Empirically, this additional-query approach outperforms methods that rely on adding extra attention layers.

3.4 Scalable Synthetic Data Generation

Training of our model requires high-quality and diverse video data with ground-truth for depth, normal, segmentation mask, dense pose, 2D keypoints, 3D keypoints, and camera poses. Since most real-world datasets may contain only a

subset of the modalities in a limited scale, we resort to synthetic data as a scalable approach to address the data scarcity problem.

Specifically, we design a synthetic data generation workflow to produce high-quality human-centric video data, covering diverse human entities and motions. We use 800 RenderPeople assets [55], and animate them with 200 motions from the CMU motion capture dataset [12]. We use various 3D full scenes or HDRI backdrops similar to [5] for enhanced background variability, and vary focal lengths, camera positioning and trajectory for enriched camera conditions. In total, we generate a synthetic dataset of 7,500 videos with various identities, motions and backgrounds. To support dense vision tasks, video normals, depth and segmentation masks are generated with separate render passes via Blender [8]. We recover the human joint positions from the rigged RenderPeople [55] assets and use them as ground truth for our 2D and 3D keypoint regression task. Each of the generated videos is trimmed to the target number of frames in the video model. We apply a offline data preprocessing stage, where we precompute and cache both the input RGB video and target-modality video into video latents, and the text conditioning prompt into text embeddings.

3.5 Training Recipe

With the model architecture and data pipeline established, we now describe the training objectives employed in our framework. Conventionally, joint multi-task learning presents an inherent optimization challenge: distinct vision tasks demand specialized loss functions tailored to their unique modalities. In practice, this heterogeneity introduces substantial engineering overhead to manually balance loss weights across different objectives, presenting a major bottleneck when scaling up to a larger variety of vision tasks. Instead, we adopt a fully unified loss design: specifically, our model is trained solely using a standard L_2 loss, applied in the latent space for dense tasks and in the output space for sparse tasks.

To achieve this unified optimization without compromising performance, we shift the responsibility of task-specific customization from the loss formulation to the data representation designs. Taking monocular depth estimation as a primary example—where scale ambiguity typically necessitates customized, scale-invariant losses—we resolve this constraint by normalizing the depth map of each video frame using the median depth of the scene, thereby inherently eliminating scale variance. To further align the normalized depth d into the standard $[0, 1]$ RGB range, we apply a nonlinear mapping function: $d' = \text{clip}(\alpha \log(d+1), 0, 1)$, where α can dynamically adjusts the model's focus between near-field details and far-field structures. This data-level customization effectively circumvents the need for any task-specific objective functions, where balancing between different tasks only needs to be managed via the data mixture ratio, analogous to the prevailing practice in LLMs. We believe this data-centric harmonization principle can scale naturally across a wider spectrum of vision tasks, ensuring a streamlined and robust multi-task training paradigm.

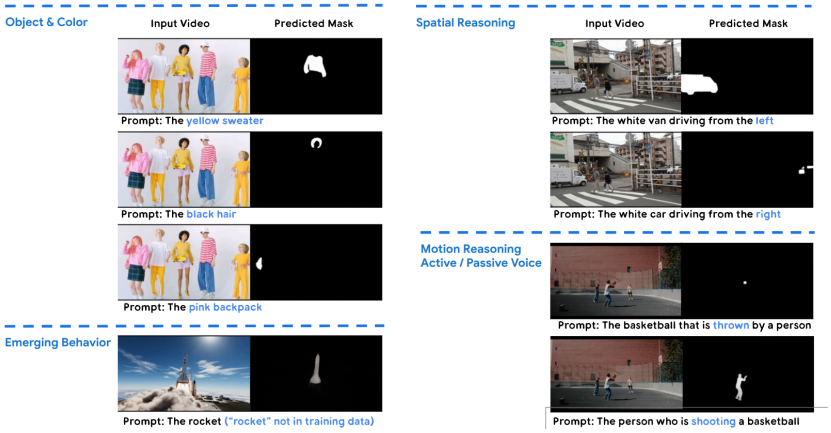


Fig. 6: Qualitative results of our approach on referring expression segmentation. Our model accurately recognizes objects, colors, spatial relationships, and motion. Furthermore, it demonstrates strong generalization by identifying unseen objects, effectively leveraging the knowledge acquired during text-to-video pretraining.

4 Experiments

We comprehensively evaluate our model on various challenging real-world datasets across vision tasks: depth estimation, normal estimation, segmentation, 3D keypoint prediction, and camera pose estimation. Please also see the supplementary for more results and analysis.

4.1 Implementation Details

Our approach is implemented based on an open-source and open-weights text-to-video diffusion model WAN [63]. Following its original architecture, we train our model on video samples at a resolution of 480x832, comprising 81 frames captured at 24 frames per second (FPS). The model's encoder downsamples the input dimensionality with a temporal factor of 4 and a spatial factor of 8 for both width and height. We train our model with a batch size of 64 on 256 v6e TPUs. We optimize the model using the Adam optimizer [35], with a learning rate of $5e^{-5}$ for 15000 training steps, implementing a linear warmup over the first 250 training steps. Crucially, for training stability, we apply both gradient clipping, constraining the gradient norm within a certain threshold, and gradient dropping, which discards any batch whose gradient norm exceeds a higher threshold. These techniques proved vital for stable training and high-quality performance. To enhance training data diversity, we augment our proposed synthetic data with several publicly available synthetic datasets. Specifically, we incorporate TartanAir [67], Virtual KITTI [19], and MVS Synth [27] to provide additional depth annotations, while utilizing the camera trajectory data provided by TartanAir [67]. For expression-referring segmentation tasks, we leverage real-world datasets owing to their immediate availability, integrating MeViS [15], Ref-COCO [77], and YouTube-VOS [71].

Table 1: Comparison to state-of-the-art specialized models. We show the performance for a common subset of tasks, including surface normal, depth estimation, camera pose estimation, foreground segmentation, expression-referring open-world segmentation, and 3D human keypoint estimation. "-" denotes that a given model cannot solve the task out-of-the-box, ∼ denotes results are not obtained, * denotes results when trained with the same data as our model. Best results are **bold**, second best underlined. Note that on all tasks except for expression-referring segmentation, our method is trained purely on synthetic data.

Task		Normals		Depth				Cam. Pose			Foreground Seg.		Expression-Referring Seg.		3D Human
Benchmark		Sintel [10]	Hi4D [76]	Sintel [10]	KITTI [21]	ETH3D [58]	Goliath [47]	Sintel [10]			V.Mat. [42]	P.Mat. [42]	Ref-DAVIS [34]	MEVIS [15]	EMDB [29]
Method	Backbone	mAE ↓		AbsRel ↓				ATE ↓	RPE-T ↓	RPE-R ↓	MSE ↓		J&F ↑		MPJPE ↓
David-Large [57]	ViT-L	-	15.37	-	-	-	-	-	-	-	-	-	-	-	-
Sapiens-2B [33]	MAE	-	12.14	-	-	-	-	-	-	-	-	-	-	-	-
Marigold-E2E-FT [20]	SD V2	33.5	∼	-	-	-	-	-	-	-	-	-	-	-	-
NormalCrafter [7]	SVD	30.7	∼	-	-	-	-	-	-	-	-	-	-	-	-
Lotus-2 [22]	SVD	30.3	∼	-	-	-	-	-	-	-	-	-	-	-	-
David-Large [57]	ViT-L	-	-	-	-	-	0.012	-	-	-	-	-	-	-	-
Sapiens-2B [33]	MAE	-	-	-	-	-	0.033	-	-	-	-	-	-	-	-
MapAnything [32]	DinoV2	-	-	0.306	0.096	∼	∼	0.202	0.089	2.383	-	-	-	-	-
VGGT [64]	DinoV2	-	-	0.247	0.067	∼	∼	0.168	0.056	0.428	-	-	-	-	-
DepthAnything 3 [41]	DinoV2	-	-	0.201	0.059	0.028	∼	0.065	0.048	0.456	-	-	-	-	-
D4RT [79]	VideoMAE2	-	-	0.148	0.051	∼	∼	0.065	0.024	**0.126**	-	-	-	-	-
VGGT *Ω* [65]	DinoV3	-	-	0.145	**0.041**	**0.016**	∼	0.057	0.059	0.407	-	-	-	-	-
MODNet [31]	MNetV2	-	-	-	-	-	-	-	-	-	0.0054	0.0304	-	-	-
RVM [43]	MNetV3	-	-	-	-	-	-	-	-	-	**0.0010**	∼	-	-	-
David-Large [57]	ViT-L	-	-	-	-	-	-		-	-	∼	0.0009	-	-	-
ReferFormer [31]	Swin-L	-	-	-	-	-	-	-	-	-			61.1	31.0	-
VISA-13B [31]	Chat-Univi	-	-	-	-	-	-	-	-	-	-	-	70.4	44.5	-
Vocap [43]	SAM2	-	-	-	-	-	-	-	-	-		-	75.1	51.9	-
ReferEverything [57]	WAN 2.1	-	-	-	-	-	-		-	-	-		75.0	60.3	-
SegmentAnything 3 [11]	PE	-	-	-	-	-	-			-	-	-	41.3	38.3	-
SegmentAnything 3 + Gemini 3.5 Flash [11]	PE	-	-	-	-	-	-			-	-	-	64.5	57.5	-
GVHMR [59]	ViT	-	-	-	-	-	-	-	-	-	-	-	-	-	72.6
TRAM [68]	ViT-H	-	-	-	-	-	-	-	-	-	-	-	-	-	74.4
Genmo [38]	ViT-H	-	-	-	-	-	-	-	-	-	-	-	-	-	73.0
Diception - Generalist [80]	SD3	37.9	20.4	0.500	0.145	0.177	∼	-	-	-	-	-	-	-	-
Ours - Specialist	WAN 2.1	29.7	**10.99**	**0.130**	0.048	0.037	**0.008**	**0.050**	**0.018**	0.464	0.0027	0.0009	**76.4**	**70.0**	**71.8**
Ours - Generalist	WAN 2.1	**29.3**	11.47	0.156	0.048	0.044	0.011	0.062	**0.018**	0.485	**0.0010**	**0.0008**	75.8	69.0	∼

4.2 Evaluation Protocol

Evaluation Datasets. We evaluate our approach on a set of challenging benchmarks that encompass both video and image data. For surface normal estimation, we evaluate on Hi4D datasets [76] and SINTEL dataset [10]. For Hi4D, we adhere to the evaluation protocol established in Sapiens [33] and DAVID [57], selecting the same sequences from pairs 28, 32, and 37. These sequences feature six unique subjects captured by camera 4, resulting in a total of 1,195 frames for testing. For depth estimation, we evaluate on KITTI [21], SINTEL [10], ETH3D [58], and Goliath [47] dataset. For Goliath [47], we follow DAVID [57] to perform evaluation to ensure fair comparison, where we utilize the identical selection of 12 cameras, 16 frames, and 4 subjects as in DAVID [57], which are categorized into data splits of face, upper and full body, with approximately 2.2k frames in total. For soft foreground segmentation, we report results on VideoMatte [42], PhotoMatte 85 [42], and PPM-100 [31]. The original VideoMatte dataset does not provide an official split, thus we evaluate on the static split composite from [43]. For camera pose estimation, we evaluate on SINTEL dataset [10]. For 3D keypoint estimation, we evaluate on EMDB dataset [29]. For expression-referring open-world segmentation, we follow VoCap [62] to reports results on RefVOS-DAVIS [34] and MeViS [15].

Evaluation Metrics. To assess our model's performance, we report standard metrics appropriate for each task. For surface normal estimation, we report the common metrics [33, 57] including mean and median angular error, alongside the percentage of pixels where the error is within the threshold $t \in \{11.25, 22.5, 30\}$. For depth estimation, we follow standard practice [33, 53, 74] by adopting the

Table 2: Validated on depth estimation, generative backbone outperform other visual pretraining (V-JEPA and Video MAE v2), and show preliminary data and model scaling properties. Note that our method was trained exclusively on synthetic data.

Method	Model Size	Number of Training			Sintel		KITTI		ETH3D		Average	
		Datasets	Videos	Frames	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
Ours - V-JEPA - H	0.6B	1	7.5K	0.9M	0.422	37.3	0.226	47.5	0.196	71.8	0.281	52.2
Ours - VideoMae V2 - H	0.6B	1	7.5K	0.9M	0.275	37.4	0.124	64.6	0.126	84.1	0.175	62.0
Ours - VideoMae V2 - G	1B	1	7.5K	0.9M	0.260	41.4	0.105	70.2	0.099	89.3	0.154	66.9
Ours - WAN 2.1 - S	1.3B	1	7.5K	0.9M	0.201	72.7	0.099	89.2	0.068	95.6	0.122	85.8
Ours - WAN 2.1 - L	14B	1	7.5K	0.9M	0.181	76.8	0.060	97.1	0.039	98.2	0.093	90.7
DepthAnything V3 - G [41]	1.15B	22+	1.23M	∼ 200M	0.201	72.1	0.059	96.6	0.028	98.8	0.096	89.1
D4RT [79]	1B	11+	∼ 1M	∼ 86M	0.148	80.3	0.055	97.9	0.045	97.0	0.082	91.7
VGGT - Ω [65]	1B	33+	∼ 3M	∼ 600M	0.145	**87.6**	0.041	**98.5**	**0.016**	**99.6**	**0.067**	**95.2**
Ours - WAN 2.1 - S	1.3B	4	8.08K	1.23M	0.167	76.9	0.060	97.3	0.056	97.6	0.094	90.6
Ours - WAN 2.1 - L	14B	4	8.08K	1.23M	**0.130**	84.5	0.048	98.3	0.037	98.8	0.071	93.8

mean absolute value of the relative depth (AbsRel) and the root mean square error (RMSE). For soft foreground segmentation, we report mean squared error (MSE) as in prior work [42, 43, 57, 75]. For expression referring open-world segmentation, we report J&F scores [51] (mean of IoU and contour accuracy) averaged on all text queries. For 3D keypoint estimation, we follow [38,68] to report mean per joint position error (MPJPE). For camera pose estimation, we report average translation error (ATE), relative pose error - translation (RPE-T), and relative pose error - rotation (RPE-R).

4.3 Comparison to the state of the art

We compare our methods with the state-of-the-art models on various challenging benchmarks. Note that our model was trained without using any of the training sets provided alongside the evaluation benchmarks. Furthermore, our training relies entirely on synthetic data, with the sole exception of the expression-referring segmentation task which incorporates real-world data. As summarized in Table 1, our approach performs favorably compared to related methods in the literature. For geometric understanding, our method exceeds specialized models like NormalCrafter [7] and Lotus-2 [22] in surface normal estimation, and outperforms or matches dedicated foundation models such as DepthAnything 3 [41], [79], and VGGT Ω [65] in depth estimation and camera pose estimation. In foreground segmentation, our approach achieves competitive results in soft foreground segmentation across both video and image datasets. On expression-referring segmentation, our work show stronger ability in understanding complex sentences compared to SegmentAnything3 [11]. Qualitative examples are show in Figure 6. Finally, for sparse tasks, it also delivers stronger 3D keypoint prediction performance compared to specialized models including Genmo [38] and TRAM [68]. Readers are referred to Appendix for more details.

4.4 Ablation Studies

Joint training vs. task-specific training: As demonstrated in Table 1, transitioning to a unified generalist yields mixed behaviors across tasks and benchmarks. For dense prediction modalities, surface normal estimation shows mixed trends on benchmarks, while depth and camera pose estimation generally undergo regression or remain intact (e.g., on KITTI). In contrast, foreground seg-

mentation tasks consistently benefit from joint training, whereas expression-referring segmentation experiences a slight degradation. Most notably, we empirically observe that joint training severely degrades 3D human keypoint estimation and compromises other dense tasks.

We presume this stems from two architectural factors. First, this token-based explicit coordinate regression diverges from the continuous pixel-space pre-training of the generative video model. Second, the additional learnable tokens, trained from scratch, disrupt the native attention mechanics of pre-trained DiT layers, damaging their optimized features and requiring more data to converge. This phenomenon suggest a valuable insight for post-training: one should strive for minimal architectural modifications to the pre-trained backbone, or alternatively, fundamentally redesign the pre-training paradigm to natively accommodate highly diverse downstream tasks.

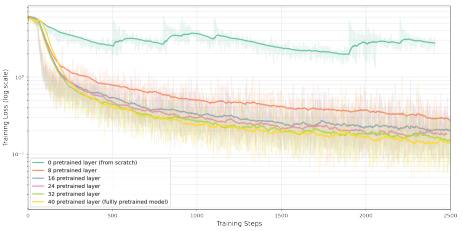


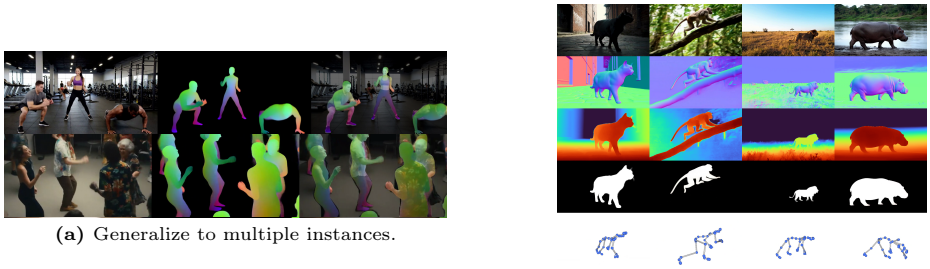
Fig. 7: Effect of transferring increasing number of layers from the pre-trained text-to-video model.

Importance of pretraining. We compare video generative pretraining backbone with other representation learning approaches. We use the largest available variants of V-JEPA and VideoMae V2 and fine-tune all on the same training set. As in Table 2, our diffusion-based pretraining (WAN 2.1) significantly outperforms other pretraining approaches. This serves as a suggestive empirical evidence that the generative diffusion objective itself—rather than just the dataset or model scale—is essential for extracting superior spatiotemporal priors. Furthermore, VideoMae and V-JEPA lack native support for a unified multi-task model, often requiring task-specific training, while our model can natively use text prompts to steer between tasks. Furthermore, as in Figure 7, training the randomly initialized DiT from scratch on our dataset yields a nearly flat learning curve. Performance improves as more pre-trained layers are used. This confirms that pre-training is essential for downstream convergence and performance.

Preliminary scaling property and data efficiency. As in Figure 2 and Table 2, the generative backbone demonstrates promising preliminary scaling properties across varying model sizes and datasets, with performance generally improving as model size and data volume increase. The model also shows exceptional data efficiency where it achieves comparable performance with leading models like D4RT and VGGT-Ω with 7× to 500× less training data.

4.5 Emergent Behaviors

In addition to competitive performance that matches or surpasses state-of-the-art methods, our model also exhibits intriguing emergent behaviors beyond its intended training objectives. We believe these behaviors arise from the rich repre-



(a) Generalize to multiple instances.

(b) Generalize to unseen categories.

Fig. 8: Generalization capabilities. Left: Trained on synthetic data with one single object within each video, our method generalizes in zero-shot to real videos with multiple objects. Right: Our approach has been trained only on videos of a single class (humans), but generalizes to a variety of other classes of articulated objects.

sentations learned in large-scale text-to-video generation models, shedding light on the broader potential of this paradigm for future exploration.

Generalization from synthetic training data: Our approach is trained purely on synthetic videos, yet it generalizes well to real-world videos, as shown in our quantitative and qualitative results. In particular, we observe that the fine-grained details present in the output of our model are exceeding the quality of our training data generated with conventional computer graphics methods in Blender. See, for example, the whiskers on the cat in Figure 8b (first column) or the soft segmentation of the hair in Figure 3 (second row).

Generalization to OOD classes: our approach is trained purely on synthetic videos of human entities, yet it generalizes well to real-world videos—both of humans and of other categories such as animals and anthropomorphic characters—as shown in Figure 8b.

Generalization to multiple instances: trained on synthetic data with one single object within each video, our method generalizes in zero-shot to real videos with multiple objects as shown in Figure 8a.

5 Conclusion

We have presented GenCeption, a unified framework that identifies large-scale video generation as a foundational pre-training paradigm for visual perception. By repurposing a pre-trained video diffusion backbone into a high-efficiency, feed-forward model, we demonstrate that the rich spatiotemporal priors required for synthesis can be directly translated into precise perceptual capabilities without the need for costly iterative sampling. Our experiments across diverse vision tasks confirm that GenCeption not only matches the performance of specialized state-of-the-art models but also exhibits significant emergent behaviors, provides strong evidence for the existence of a universal "world model" within generative video backbones. We believe this paradigm shift, from task-specific engineering to generalist generative pre-training, paves a promising path toward a truly unified and scalable intelligence for understanding the physical world.

References

1. Acuaviva, P., Davtyan, A., Hassan, M., Stapf, S., Rahimi, A., Alahi, A., Favaro, P.: From generation to generalization: Emergent few-shot learning in video diffusion models. arXiv preprint arXiv:2506.07280 (2025)
2. Bachmann, R., Kar, O.F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., Zamir, A.: 4m-21: An any-to-any vision model for tens of tasks and modalities. Advances in Neural Information Processing Systems **37**, 61872–61911 (2024)
3. Bagchi, A., Bao, Z., Wang, Y.X., Tokmakov, P., Hebert, M.: Refereverything: Towards segmenting everything we can speak of in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23221–23231 (2025)
4. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
5. Bazavan, E.G., Zanfir, A., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Hspace: Synthetic parametric humans animated in complex environments (2022), `https://arxiv.org/abs/2112.12867`
6. Bin, Y., Hu, W., Wang, H., Chen, X., Wang, B.: Normalcrafter: Learning temporally consistent normals from video diffusion priors (2025), `https://arxiv.org/abs/2504.11427`
7. Bin, Y., Hu, W., Wang, H., Chen, X., Wang, B.: Normalcrafter: Learning temporally consistent normals from video diffusion priors. arXiv preprint arXiv:2504.11427 (2025)
8. Blender Online Community: Blender - a 3d modeling and rendering package (2025), `https://www.blender.org`
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
10. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. pp. 611–625. Springer (2012)
11. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
12. Carnegie Mellon University: CMU Graphics Lab Motion Capture Database. `http://mocap.cs.cmu.edu/` (2001), database created with funding from NSF EIA-0196217
13. Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation, 2023. URL https://arxiv. org/abs/2306.15595
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
15. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2694–2703 (2023)
16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)

17. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: European Conference on Computer Vision (ECCV) (2024)
18. Gabeur, V., Long, S., Peng, S., Voigtlaender, P., Sun, S., Bao, Y., Truong, K., Wang, Z., Zhou, W., Barron, J.T., et al.: Image generators are generalist vision learners. arXiv preprint arXiv:2604.20329 (2026)
19. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)
20. Garcia, G.M., Zeid, K.A., Schmidt, C., De Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 753–762 (2025)
21. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
22. He, J., Li, H., Sheng, M., Chen, Y.C.: Lotus-2: Advancing geometric dense prediction with powerful image generative model. arXiv preprint arXiv:2512.01030 (2025)
23. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020), `https://arxiv.org/abs/2006.11239`
25. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: Depthcrafter: Generating consistent long depth sequences for open-world videos (2024), `https://arxiv.org/abs/2409.02095`
26. Huang, J., Zhang, Y., He, X., Gao, Y., Cen, Z., Xia, B., Zhou, Y., Tao, X., Wan, P., Jia, J.: Unityvideo: Unified multi-modal multi-task learning for enhancing world-aware video generation. arXiv preprint arXiv:2512.07831 (2025)
27. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2821–2830 (2018)
28. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4d: Leveraging video generators for geometric 4d scene reconstruction. arXiv preprint arXiv:2504.07961 (2025)
29. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In: International Conference on Computer Vision (ICCV) (2023)
30. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9492–9502 (June 2024)
31. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1140–1147 (2022)
32. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
33. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.)

Computer Vision – ECCV 2024. pp. 206–228. Springer Nature Switzerland, Cham (2025)
34. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Asian conference on computer vision. pp. 123–141. Springer (2018)
35. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
36. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
37. Kuang, Z., Zhang, T., Zhang, K., Tan, H., Bi, S., Hu, Y., Xu, Z., Hasan, M., Wetzstein, G., Luan, F.: Buffer anytime: Zero-shot video depth and normal from image priors (2024), https://arxiv.org/abs/2411.17249
38. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: Generative models for human motion synthesis. arXiv preprint arXiv:2505.01425 (2025)
39. Li, T., He, K.: Back to basics: Let denoising generative models denoise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 36115–36125 (2026)
40. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, Z.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., Wang, Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models (2025), https://arxiv.org/abs/2501.18590
41. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
42. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8762–8771 (2021)
43. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 238–247 (2022)
44. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
45. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
46. Martin Garcia, G., Abou Zeid, K., Schmidt, C., de Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
47. Martinez, J., Kim, E., Romero, J., Bagautdinov, T., Saito, S., Yu, S.I., Anderson, S., Zollhöfer, M., Wang, T.L., Bai, S., et al.: Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. Advances in Neural Information Processing Systems **37**, 83008–83023 (2024)
48. Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4m: Massively multimodal masked modeling. Advances in Neural Information Processing Systems **36**, 58363–58408 (2023)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

50. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
51. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
53. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
54. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
55. Renderpeople: Renderpeople: 3d people for renderings., `https://renderpeople.com`
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 10684–10695 (2022)
57. Saleh, F., Aliakbarian, S., Hewitt, C., Petikam, L., Xiao-Xian, Criminisi, A., Cashman, T.J., Baltrušaitis, T.: DAViD: Data-efficient and accurate vision models from synthetic data (2025), `https://arxiv.org/abs/2507.15365`
58. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3260–3269 (2017)
59. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia Conference Proceedings (2024)
60. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
61. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
62. Uijlings, J., Zhou, X., Gu, X., Nagrani, A., Arnab, A., Fathi, A., Ross, D., Schmid, C.: Vocap: Video object captioning and segmentation from any prompt. arXiv preprint arXiv:2508.21809 (2025)
63. Wan, T.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
64. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
65. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupprecht, C.: Vggt-*omega*. arXiv preprint arXiv:2605.15195 (2026)

66. Wang, L., Zanfir, A., Bazavan, E.G., Andriluka, M., Sminchisescu, C.: Thfm: A unified video foundation model for 4d human perception and beyond. arXiv preprint arXiv:2603.25892 (2026)
67. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
68. Wang, Y., Wang, Z., Liu, L., Daniilidis, K.: Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In: European Conference on Computer Vision. pp. 467–487. Springer (2024)
69. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
70. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 (2024)
71. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 585–601 (2018)
72. Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., He, T.: Depth any video with scalable synthetic data (2025), https://arxiv.org/abs/2410.10815
73. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
74. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
75. Yang, P., Zhou, S., Zhao, J., Tao, Q., Loy, C.C.: Matanyone: Stable video matting with consistent memory propagation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7299–7308 (2025)
76. Yin, Y., Guo, C., Kaufmann, M., Zarate, J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Computer Vision and Pattern Recognition (CVPR) (2023)
77. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
78. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
79. Zhang, C., Moing, G.L., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., et al.: Efficiently reconstructing dynamic scenes one d4rt at a time. arXiv preprint arXiv:2512.08924 (2025)
80. Zhao, C., Sun, Y., Liu, M., Zheng, H., Zhu, M., Zhao, Z., Chen, H., He, T., Shen, C.: Diception: A generalist diffusion model for visual perceptual tasks. arXiv preprint arXiv:2502.17157 (2025)
81. Zhu, X., Zhu, J., Li, H., Wu, X., Li, H., Wang, X., Dai, J.: Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16804–16815 (2022)
82. Zoran, D., Parthasarathy, N., Yang, Y., Hudson, D.A., Carreira, J., Zisserman, A.: Recurrent video masked autoencoders. arXiv preprint arXiv:2512.13684 (2025)