RefAlign: Representation Alignment for Reference-to-Video Generation

Lei Wang[1,2*,‡], Yuxin Song[2,‡], Ge Wu[1], Haocheng Feng[2], Hang Zhou[2], Jingdong Wang[2], Yaxing Wang[4†], Jian Yang[1,3†]

[1]PCA Lab, VCIP, College of Computer Science, Nankai University, China
[2] Baidu Inc., China
[3] PCA Lab, School of Intelligence Science and Technology, Nanjing University, China
[4] College of Artificial Intelligence, Jilin University, China
{scitop1998}@gmail.com, {songyuxinbb}@outlook.com,
{yaxing,csjyang}@nankai.edu.cn
Code: https://github.com/gudaochangsheng/RefAlign

Fig. 1: Reference-to-video generation using our proposed method, RefAlign.

Abstract. Reference-to-video (R2V) generation is a controllable video synthesis paradigm that constrains the generation process using both text prompts and reference images, enabling applications such as personalized advertising and virtual try-on. In practice, existing R2V methods typically introduce additional high-level semantic or cross-modal features alongside the VAE latent representation of the reference image and

† Corresponding authors. *Interns in Baidu Inc. ‡ Equal contribution

jointly feed them into the diffusion Transformer (DiT). These auxiliary representations provide semantic guidance and act as implicit alignment signals, which can partially alleviate pixel-level information leakage in the VAE latent space. However, they may still struggle to address copy–paste artifacts and multi-subject confusion caused by modality mismatch across heterogeneous encoder features. In this paper, we propose RefAlign, a representation alignment framework that explicitly aligns DiT reference-branch features to the semantic space of a visual foundation model (VFM). The core of RefAlign is a reference alignment loss that pulls the reference features and VFM features of the same subject closer to improve identity consistency, while pushing apart the corresponding features of different subjects to enhance semantic discriminability. This simple yet effective strategy is applied only during training, incurring no inference-time overhead, and achieves a better balance between text controllability and reference fidelity. Extensive experiments on the OpenS2V-Eval benchmark demonstrate that RefAlign outperforms current state-of-the-art methods in TotalScore, validating the effectiveness of explicit reference alignment for R2V tasks.

Keywords: Reference representation alignment · Reference consistency · Reference-to-video generation

1 Introduction

Diffusion models have driven rapid advances in video generation. Representative commercial systems (e.g., Sora) [2,3,35] and open-source models [37,40,44] now achieve high-fidelity text-to-video (T2V) and image-to-video (I2V) synthesis. However, T2V relies solely on the prompt and offers limited fine-grained control (e.g., identity and appearance), while I2V is more controllable but restricts diversity due to the reference image.

To bridge this gap, reference-to-video (R2V) [6,15,24] has attracted increasing attention. Conditioned on the prompt and the reference image, it aims to generate instruction-following videos while preserving subject identity and appearance, with applications in personalized advertising [5, 22] and virtual try-on [20,27]. A key challenge, however, is effective multi-modal conditioning during video generation.

To alleviate the aforementioned challenges, prior work [6, 11, 24] typically adopts a "two-stream reference" paradigm: it leverages a 3D VAE to extract reference features that provide low-level details; meanwhile, it also introduces an additional encoder to encode the reference images and supply higher-level semantic cues, which can partially mitigate the detail leakage caused by the former. However, because both reference images and textual prompts are often processed by separate, independent encoders, these methods may struggle to model complex cross-modal interactions—particularly for prompts that involve spatial relations and temporal reasoning.

Furthermore, some methods [8, 21, 29] leverage cross-modal representations from multimodal large language models (MLLM) to enhance deeper semantic

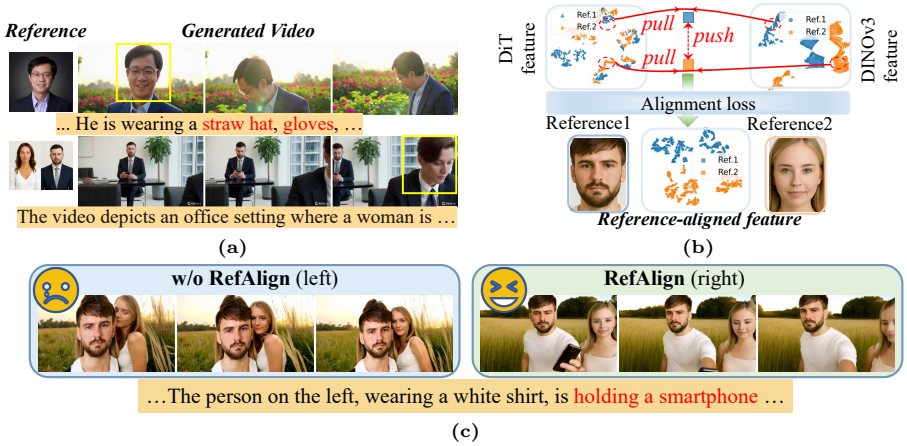


Fig. 2: Motivation of the proposed RefAlign method. (a) The R2V task suffers from copy–paste artifacts (top) and multi-subject confusion (bottom), both generated by Kling [17]. (b) t-SNE [26] visualization of reference feature distributions. DiT features (conditioned on VAE-encoded inputs) are highly entangled and overlap substantially across references, whereas DINOv3 features are more separable. RefAlign aligns DiT features to the DINOv3 feature space via an alignment loss, improving reference separability by *pulling* same-reference features closer and *pushing* different-reference features farther apart. (c) Visual comparison with and without RefAlign.

interactions among multimodal inputs. However, this often incurs substantial inference overhead (e.g., introducing models such as Qwen2/2.5-VL-7B [1, 38]). More importantly, these approaches still fundamentally rely on feeding additional semantic features together with cross-modal features into the diffusion transformer, attempting to mitigate modality mismatch via implicit alignment. *Here, modality mismatch refers to the discrepancy between DiT's internal reference representations derived from VAE latents and the external semantic reference representations injected for conditioning, making implicit alignment less effective.* Lacking explicit constraints on the alignment mechanism itself, the marginal benefit of the extra features may be limited. As a result, existing models can still suffer from copy–paste artifacts (*i.e.*, overly replicating the reference image in generated videos, Fig. 2(a)(top)) and multi-subject confusion (*i.e.*, mutual interference and blending of different subjects' appearances, Fig. 2(a)(bottom)).

In this paper, we propose a simple yet effective strategy called RefAlign to mitigate the misalignment among multimodal features. Fig. 2(b) visualizes the latent distributions of reference images produced by different encoders. For the VAE, the features input to DiT are scattered and poorly structured, with strong cross-reference entanglement. As a result, discriminative boundaries are hard to establish. In contrast, features extracted by DINOv3 [34] demonstrate stronger separability: distributions from different reference images are more distinct, while features belonging to the same reference remain compact and consistent. Motivated by this observation, we design a vision foundation model (VFM)–guided

optimization strategy for the VAE-based reference representation in DiT. We find that VFM guidance can compensate for the limited semantic expressiveness of DiT latent features, thereby alleviating copy—paste artifacts and multi-subject confusion. Fig. 2(c) provides qualitative evidence supporting this claim.

RefAlign's main contribution is the proposed reference alignment (RA) loss, which aligns the DiT features of reference images into the feature space of a pre-trained VFM during training. Although directly incorporating VFM features as a complement to VAE features yields performance gains—likely because additional semantic information facilitates implicit multimodal alignment—this strategy introduces an extra encoder and increases inference cost. In contrast, we find that a carefully designed RA loss plays a key role in achieving effective alignment without additional inference overhead. RA loss is designed to regularize the feature distribution while preserving the expressive capacity of DiT reference features. Moreover, to prevent overly strong similarity constraints from causing representation collapse and exacerbating multi-subject confusion, our RA loss further incorporates a negative alignment mechanism. Finally, we systematically investigate the impact of different VFMs on alignment effectiveness.

To evaluate generation performance on the R2V task, we apply RA loss to one of the most representative video generation backbones, *Wan2.1* [37]. We conduct a comprehensive evaluation of RefAlign on the OpenS2V-Eval [47] benchmark, where it achieves state-of-the-art performance in TotalScore. Qualitative results further demonstrate clear improvements in subject fidelity and consistency. The main contributions of this paper are summarized as follows.

- We propose RefAlign, the first R2V framework that regularizes reference image features by leveraging a vision foundation model.
- We introduce the RA loss, which effectively alleviates both copy–paste artifacts and multi-subject confusion in R2V without introducing additional inference overhead.
- On the OpenS2V-Eval benchmark, RefAlign outperforms prior R2V methods, and extensive ablation studies further validate the effectiveness and necessity of the proposed design. *We will make the model and code publicly available to support reproducibility and further research.*

2 Related Work

2.1 Reference-to-Video Generation

Reference-to-video (R2V) aims to synthesize high-quality videos conditioned on text and reference images. R2V has evolved from human-centric identity-preserving generation [33, 43, 48, 51] to more general objects and scenes [11, 15, 21, 24, 50], enabling more flexible control. By reference-conditioning strategy, R2V methods fall into three categories: multi-encoder reference conditioning, reference disentanglement, and MLLM-based cross-modal guidance.

The first category introduces a semantic branch independent of the VAE encoder to impose global semantic constraints on the reference [6, 8, 11, 21, 24], mitigating pixel-level leakage and copy–paste artifacts. For example, Phantom [24]

injects fine-grained appearance cues via VAE features while anchoring subject semantics with CLIP [31] features.

The second category either decouples reference conditioning from video representations [50, 52] or explicitly aligns each reference subject with its corresponding text [9, 15], reducing multi-subject interference and identity confusion. For instance, Kaleido [50] adopts rotational positional encoding to distinguish image conditions from video tokens, whereas MAGREF [9] and ConceptMaster [15] bind subjects to their textual descriptions via masking and a decoupled attention module, respectively.

The third category leverages MLLMs for deeper vision–text interaction to provide cross-modal relational and semantic guidance [4, 8, 13, 14, 21, 29]. For example, BindWeave [21] uses Qwen2.5-VL [1] to facilitate multi-reference interaction, while PolyVivid [14] employs an internal LLaVA [23] to embed visual identity into the text space for precise semantic alignment. In addition, VACE [16] is a unified model that supports both R2V and video editing via a context adapter; however, due to the difficulty of multi-task optimization, it may still struggle to fully preserve identity consistency in reference-guided generation. Unlike the above methods, we use external visual encoder features as semantic anchors and apply explicit feature alignment to pull same-subject features closer and push different-subject features apart, improving identity consistency and semantic discriminability while reducing multi-subject confusion.

2.2 Alignment-based Method

Recently, REPA [46] accelerates training convergence by aligning DiT mid-block features to those of a Vision Foundation Model (VFM). Building on this, REPA-E [19] enables end-to-end VAE tuning, while DDT [39] improves the alignment paradigm by decoupling the encoder and decoder. In a parallel line of work, REG [41] and ReDi [18] jointly model image latents and semantic signals within the diffusion process. Beyond aligning the denoiser, VA-VAE [45] extends alignment to the VAE tokenizer; ARRA [42] transfers the alignment principle for autoregressive text-to-image generation; and VideoREPA [49] generalizes alignment to video generation to enhance physical consistency. In contrast, RefAlign aligns reference-conditioning features to VFM features, rather than aligning the generation target to VFM, which can strengthening conditional semantics and improving fine-tuning stability. This *reference-centric* alignment approach may also inspire video editing and may promote unified modeling of understanding and generation tasks.

3 Method

We first briefly introduce text-to-video generation and representation alignment (REPA) [46] in Sec. 3.1, which form the foundation of our work. Inspired by REPA, we present the overall RefAlign pipeline in Sec. 3.2 and its core design, the reference alignment (RA) loss, in Sec. 3.3. We further discuss the choice

of target encoders for alignment in Sec. 3.4. Finally, we summarize the key differences between RefAlign and REPA in Sec. 3.5.

3.1 Preliminary

Text-to-Video Generation. Text-to-video (T2V) synthesis [37] typically performs generative denoising modeling in a low-dimensional latent space, thereby significantly reducing computational cost. Its training objective follows the Rectified Flow (RF) [10] paradigm. RF defines a straight-line trajectory in this space by linearly interpolating between the noise and data latents. Therefore, the training loss function can be formulated as:

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{z,\epsilon,c,t}\left[\left\|\varepsilon_{\Theta}\left(z_t, c, t\right) - \left(\epsilon - z_0\right)\right\|_2^2\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is a standard Gaussian noise sample, z_0 denotes the latent variable of the clean video data, z_t is the noisy latent variable at timestep t, c is the condition (e.g., prompt) and $\varepsilon_{\Theta}\left(z_t, c, t\right)$ is the output predicted by the network parameterized by Θ.

Representation Alignment. Representation Alignment (REPA) [46] is a regularization method designed to accelerate the convergence of diffusion transformers (DiT). It projects DiT hidden states through an MLP and aligns them with clean-image representations from a frozen pretrained vision encoder (e.g., DINOv2 [28]). The REPA loss can be formulated as:

$$\mathcal{L}_{\mathrm{REPA}} = -\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\cos\left(f_n^*, \hat{h}_n^t\right)\right], \tag{2}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity, n is the patch index, t denotes the timestep, $f^* \in \mathbb{R}^{N\times D}$ is the vision-encoder features with N patch tokens of dimension D, and $\hat{h}^t = \Psi(h^t) \in \mathbb{R}^{N\times D}$ is obtained by projecting the DiT hidden states h^t through an MLP projector $\Psi(\cdot)$.

3.2 Pipeline

The overall framework of RefAlign is illustrated in Fig. 3(a). Our model comprises a frozen T5 encoder [32] $\varepsilon_{\mathrm{T5}}$, a frozen Wan-VAE [37] $\varepsilon_{\mathrm{VAE}}$, and a trainable diffusion transformer (DiT) [37] parameterized by Θ. Given a text prompt c_{text}, reference images $I = \{I_m\}_{m=1}^{M}$, a target video x, and the noisy latent z_t at timestep t, we encode the prompt, references, and target video as:

$$\hat{c}_{\mathrm{text}} = \varepsilon_{\mathrm{T5}}\left(c_{\mathrm{text}}\right), \quad \hat{I} = \{\varepsilon_{\mathrm{VAE}}(I_m)\}_{m=1}^{M}, \quad z_0 = \varepsilon_{\mathrm{VAE}}\left(x\right). \tag{3}$$

We employ a DiT consisting of L transformer blocks. Following REPA [46], we apply the alignment only to the intermediate *reference image-token* features extracted from the first K blocks (Sec. 3.3). Importantly, all L blocks are jointly

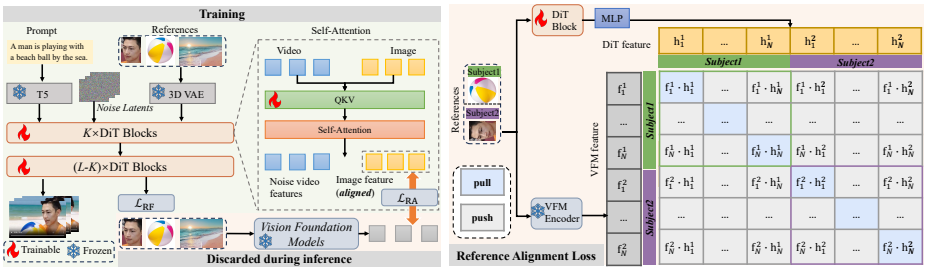


Fig. 3: (a) Overview of RefAlign. During training, we apply the proposed reference alignment loss $\mathcal{L}_{\mathrm{RA}}$ to intermediate features in selected DiT blocks and align them to target features extracted by a frozen vision foundation model (VFM). During inference, we discard the alignment process and the VFM. (b) Illustration of the reference alignment (RA) loss. RA loss aligns DiT reference features to their corresponding VFM teacher features by pulling matched (same-subject) pairs together and pushing mismatched (cross-subject) pairs apart, improving reference-consistent generation.

optimized under the RF objective. The DiT takes $(z_t, \hat{c}_{\text{text}}, \hat{I}, t)$ as input and outputs $\varepsilon_\Theta(z_t, \hat{c}_{\text{text}}, \hat{I}, t)$. The RF objective in Eq. (1) can be written as:

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{z_0,\epsilon,t}\left[\left\|\varepsilon_\Theta\left(z_t, \hat{c}_{\text{text}}, \hat{I}, t\right) - (\epsilon - z_0)\right\|_2^2\right]. \tag{4}$$

Inference. At inference, we remove the training-time VFM and MLP (Fig. 3(a)) to avoid extra overhead. Following Phantom [24] and Video Alchemist [6], we use an RF sampler with classifier-free guidance (CFG) [12]. At timestep t, the guided prediction is:

$$\begin{aligned}\hat{\varepsilon}_\Theta\left(z_t, \hat{c}_{\text{text}}, \hat{I}, t\right) = \varepsilon_\Theta\left(z_t, \varnothing, \varnothing, t\right) + \mu_1\Big(\varepsilon_\Theta\left(z_t, \varnothing, \hat{I}, t\right) - \varepsilon_\Theta\left(z_t, \varnothing, \varnothing, t\right)\Big) \\ + \mu_2\Big(\varepsilon_\Theta\left(z_t, \hat{c}_{\text{text}}, \hat{I}, t\right) - \varepsilon_\Theta\left(z_t, \varnothing, \hat{I}, t\right)\Big).\end{aligned} \tag{5}$$

where $\varnothing$ denotes dropping the corresponding condition (i.e., a null text embedding or a null reference embedding). μ_1 and μ_2 balance the reference-image and text conditions.

3.3 Reference Alignment Loss

By comparing the feature distributions of reference images encoded by DiT and DINOv3, we observe that DiT features are strongly coupled across references, whereas DINOv3 features are more separable. Motivated by this observation, we introduce a reference alignment (RA) loss that aligns DiT reference representations to the vision foundation model (VFM) feature space. As shown in Fig. 3(a), we align the *reference image-token* features produced inside the self-attention of

the first K DiT blocks to features extracted by a frozen VFM $\varepsilon_{\text{VFM}}(\cdot)$. Specifically, for each training sample we obtain the VFM features from I,

$$f = \{\varepsilon_{\text{VFM}}(I_m)\}_{m=1}^{M}, \in \mathbb{R}^{M\times N\times D}. \tag{6}$$

Let $h^{(l)} = \varepsilon_{\Theta^{(l)}}(I)$ be the *reference image-token* features from the self-attention module of the l-th DiT block ($l \leq K$), and we project them with an MLP $\Psi_{\text{proj}}(\cdot)$ to match the VFM feature dimension:

$$\hat{h}^{(l)} = \Psi_{\text{proj}}\left(h^{(l)}\right), \in \mathbb{R}^{M\times N\times D}. \tag{7}$$

As illustrated in Fig. 3(b), we define the RA loss with a positive (diagonal) term and a negative (off-diagonal) term. We pull DiT reference tokens closer to the corresponding VFM tokens for the same subject:

$$\mathcal{L}_{\text{pos}}^{(l)} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\left(1-\cos\left(\hat{h}_n^{(l),m}, f_n^m\right)\right). \tag{8}$$

To enforce subject-level separation, we push features from subject m away from features of other subjects $m' \neq m$ using a margin δ:

$$\mathcal{L}_{\text{neg}}^{(l)} = \frac{1}{M(M-1)}\sum_{m=1}^{M}\sum_{\substack{m'=1\\ m'\neq m}}^{M}\frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\left[\delta-\left(1-\cos\left(\hat{h}_n^{(l),m}, f_{n'}^{m'}\right)\right)\right]_+, \tag{9}$$

where $[x]_+ = \max(x, 0)$. Then, we average over the first K blocks and combine the two terms:

$$\mathcal{L}_{\text{RA}} = \frac{1}{K}\sum_{l=1}^{K}\left(\mathcal{L}_{\text{pos}}^{(l)} + \lambda\,\mathcal{L}_{\text{neg}}^{(l)}\right), \tag{10}$$

where λ controls the weight of the negative term. **Special case.** When $M = 1$ (i.e., only one reference), the inter-subject negative term is undefined and we set $\mathcal{L}_{\text{neg}}^{(l)} = 0$, so $\mathcal{L}_{\text{RA}}$ reduces to the positive alignment loss. Finally, combining Eq. (4) and Eq. (10), we optimize our model with the following overall objective:

$$\mathcal{L} = \mathcal{L}_{\text{RF}} + \eta\,\mathcal{L}_{\text{RA}}, \tag{11}$$

where η is a scalar hyperparameter balancing RA and RF losses.

3.4 Alignment Using Different Encoders

In RefAlign, $\mathcal{E}_{\text{VFM}}$ aligns the intermediate features of DiT's reference branch to establish a stable subject anchor. Therefore, the representational characteristics of $\mathcal{E}_{\text{VFM}}$ strongly influence the behavior of the alignment constraint. Below, we discuss the differences induced by three types of encoders.

DINOv3. DINOv3 provides self-supervised patch representations with strong instance-level discriminability [34,46]. When used as the alignment target, it encourages the reference branch to preserve identity-relevant cues while being less sensitive to appearance variations such as background, illumination, and pose. Its limitation is the lack of language supervision, resulting in relatively weaker high-level semantic constraints. *Therefore, using* $\mathcal{E}_{VFM}^{DINOv3}$ *steers RefAlign to emphasize intra-subject representation consistency and inter-subject separability, helping mitigate the effects of appearance variations and cross-subject interference on the reference-branch representations.*

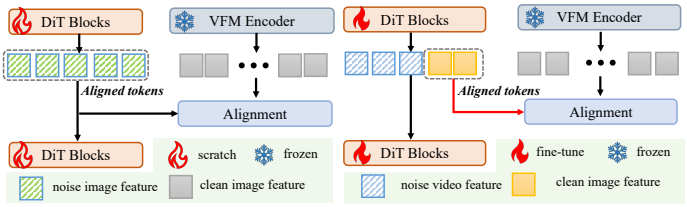
SigLIP2. SigLIP2 is trained using contrastive learning, which encourages patch representations to be consistent with global semantic separability [36]. As a result, local cues are often consolidated into more stable category- or attribute-level features. However, when the subject and background are highly co-occurring, this global discriminative bias can also encode background co-occurrence patterns as informative signals. *Therefore, when using* $\mathcal{E}_{VFM}^{SigLIP2}$ *as the alignment target, RefAlign tends to emphasize consistency along semantic-attribute dimensions: attribute cues that persist in the reference images and align with the prompt semantics are more likely to be amplified in the reference-branch representations.*

Qwen2.5-VL. Qwen2.5-VL vision tokens are strongly coupled with textual semantics through multimodal pre-training, making them well-suited for providing cross-modal semantic supervision [1]. When used as the alignment target, the reference branch is more likely to be pulled toward a prompt-consistent semantic direction (in contrast to SigLIP2, which tends to emphasize category/attribute-level semantics). A limitation is that multimodal fusion and sequence structuring can weaken the stability of local spatial correspondences, and the model is heavier (7B), incurring higher training cost and VRAM pressure. *Therefore, when using* $\mathcal{E}_{VFM}^{Qwen2.5\text{-}VL}$ *as the alignment target, RefAlign places more emphasis on contextualized compositional semantics, encouraging the reference-branch representation to organize reference cues into concept-level features aligned with the prompt semantics.*

Overall, RefAlign tends to benefit more from supervision that is **spatially consistent, identity-sensitive, and robust to appearance variations**. DINOv3 may align better with this form of supervision; by contrast, SigLIP2 tends to emphasize semantic discrimination and Qwen2.5-VL leans toward multimodal semantic fusion, making their alignment signals more likely to deviate from pure "subject anchoring".

3.5 Relationship to REPA

Our RefAlign and REPA both leverage representations from VFM to assist DiT training, but they differ fundamentally in motivation, objective, and mechanism (as shown in Fig. 4). **1) Different motivations**: REPA uses VFM to regularize DiT's *target* representations, easing semantic learning from noise. However, RefAlign regularizes the *reference condition* with VFM to mitigate VAE-induced pixel-level leakage and copy–paste artifacts, improving reference controllability. **2) Different objectives**: REPA targets training *from scratch*, using semantic



(a) REPA Training **(b)** RefAlign Training

Fig. 4: Training comparison between REPA and RefAlign. (a) REPA: Trained from scratch, aligning noisy generation targets with clean VFM features to accelerate DiT convergence. (b) RefAlign: Fine-tuned from *Wan2.1* [37] initialization, aligning clean reference-branch image features with clean VFM features to optimize reference representations and improve reference controllability.

Table 1: Quantitative comparison of RefAlign and other methods on zero-shot OpenS2V-Eval results. ↑ denotes that higher is better.

Method	TotalScore↑	Aesthetics↑	MotionSmoothness↑	MotionAmplitude↑	FaceSim↑	GmeScore↑	NexusScore↑	NaturalScore↑
Closed-source								
Vidu2.0 [2]	51.95%	41.48%	90.45%	13.52%	35.11%	67.57%	43.37%	65.88%
Pika2.1 [30]	51.88%	46.88%	87.06%	24.71%	30.38%	69.19%	45.40%	63.32%
Kling1.6 [17]	56.23%	44.59%	86.93%	41.60%	40.10%	66.20%	45.89%	74.59%
Saber-14B [52]	57.91%	42.42%	96.12%	21.12%	49.89%	67.50%	47.22%	72.55%
Open-source								
VACE-1.3B [16]	49.89%	48.24%	97.20%	18.83%	20.57%	71.26%	37.91%	65.46%
VACE-P1.3B [16]	48.98%	47.34%	96.80%	12.03%	16.59%	71.38%	40.19%	64.31%
Phantom-1.3B [24]	54.89%	46.67%	93.30%	14.29%	48.56%	69.43%	42.48%	62.50%
RefAlign-1.3B	**56.30%**	42.96%	94.74%	**20.74%**	**53.06%**	66.85%	**43.97%**	**66.25%**
VACE-14B [16]	57.55%	47.21%	94.97%	15.02%	55.09%	67.27%	44.08%	67.04%
SkyReels-A2-P14B [11]	52.25%	39.41%	87.93%	25.60%	45.95%	64.54%	43.75%	60.32%
Phantom-14B [24]	56.77%	46.39%	96.31%	33.42%	51.46%	70.65%	37.43%	69.35%
MAGREF-480P [9]	52.51%	45.02%	93.17%	21.81%	30.83%	70.47%	43.04%	66.90%
VINO [4]	57.85%	45.92%	94.73%	12.30%	52.00%	69.69%	42.67%	71.99%
Kaleido [50]	56.59%	46.15%	98.11%	12.67%	34.56%	67.51%	41.70%	82.18%
BindWeave [21]	57.61%	45.55%	95.90%	13.91%	53.71%	67.79%	46.84%	66.85%
RefAlign-14B	**60.42%**	46.84%	97.61%	22.48%	**55.23%**	68.32%	**48.52%**	73.63%

regularization to speed convergence. By contrast, RefAlign targets *fine-tuning* pretrained video generators (e.g., *Wan2.1* [37]), balancing text and reference conditions while largely preserving generative quality. **3) Different task suitability**: In multi-reference settings, REPA's one-to-one alignment can be overly hard, potentially collapsing different reference representations toward an average and thus reducing discriminability and increasing subject confusion (see Tab. 2 and Fig. 6). RefAlign instead enforces inter-reference separability via cross-subject discrimination (e.g., $\mathcal{L}_{\text{neg}}$), making it well suited to multi-reference-to-video generation. **4) Different alignment representations**: REPA pulls *noisy* representations toward VFM, making it more suitable for training from scratch. In contrast, RefAlign aligns the *clean* DiT reference-image representations to VFM, reducing interference with the generative backbone for stable fine-tuning.

4 Experiment

4.1 Experiment Setting

Datasets and Evaluation. RefAlign is fine-tuned on a high-quality subset (360K samples) curated from OpenS2V-5M [47] and Phantom-Data [7]. The subset consists of image–text–video triplets, with a regular-pair to cross-pair ratio of 6:4. To validate effectiveness, we generate 180 videos on the OpenS2V-Eval [47] benchmark. We report Aesthetics (visual quality), MotionSmoothness (motion continuity), MotionAmplitude (motion magnitude), FaceSim (face fidelity), NexusScore (subject consistency), NaturalScore (naturalness), and Gme Score (video–text alignment).

Implementation Details. RefAlign is fine-tuned on the *Wan2.1* [37] T2V backbone. The MLP for $\mathcal{L}_{\text{RA}}$ has two linear layers with SiLU, and its parameters are not shared across DiT blocks. The model is trained for 3000 iterations with AdamW [25] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0.01$), a learning rate of $5e$-5, and a global batch size of 128. To mitigate copy–paste artifacts, we apply data augmentation to the reference images in regular pairs (e.g., random rotation, scaling, horizontal flip, affine transformations (including shear), Gaussian blur, and color jitter). We set λ and η in Eqs. (10) and (11) to 1.0. During inference, we use a 50-step Euler sampler, with μ_1 and μ_2 in Eq. (5) set to 5.0 and 7.5, respectively.

Baselines. We choose *closed-source* models (e.g., Vidu [2], Pika [30], Kling [17], and Saber [52]) and *open-source* methods (e.g., VACE [16], SkyReels-A2 [11], Phantom [24], MAGREF [9], VINO [4], Kaleido [50], and BindWeave [21]) as baselines for comparison with RefAlign.

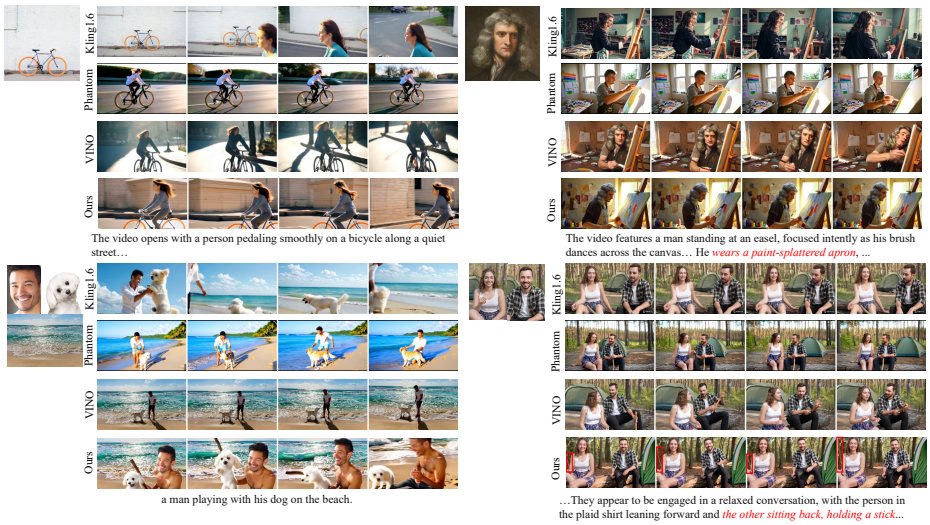


Fig. 5: Qualitative results. We compare RefAlign with three representative methods, namely Kling1.6 [17], Phantom [24], and VINO [4].

Table 2: Ablation study on the impact of the RA loss design in RefAlign on the OpenS2V-Eval. **Config D** adopts a dual-encoder input, feeding both VAE and DINOv3 features into DiT simultaneously.

Configuration	TotalScore↑	Aesthetics↑	MotionSmoothness↑	MotionAmplitude↑	FaceSim↑	GmeScore↑	NexusScore↑	NaturalScore↑
A: $\mathcal{L}_{\text{RA}}$	**55.73%**	41.53%	90.00%	29.53%	53.15%	67.54%	46.23%	62.96%
B: $\mathcal{L}_{\text{RA}}$(w/o $\mathcal{L}_{\text{neg}}$)	51.75%	41.12%	91.80%	17.67%	48.67%	61.31%	46.61%	53.75%
C: w/o $\mathcal{L}_{\text{RA}}$	49.93%	36.46%	88.97%	28.62%	68.45%	58.83%	38.63%	40.46%
D: VAE + DINOv3	52.15%	37.14%	87.98%	39.26%	35.11%	67.58%	45.78%	66.06%

4.2 Quantitative Results

Tab. 1 reports the R2V generation performance of RefAlign, open-source baselines, and closed-source baselines on the OpenS2V-Eval benchmark. For the 1.3B model, RefAlign achieves the state-of-the-art (SOTA) TotalScore, indicating a better overall trade-off across evaluation metrics than competing methods. Specifically, RefAlign obtains the best results on NexusScore and FaceSim, demonstrating its clear advantage in subject consistency and fidelity. Meanwhile, RefAlign also delivers competitive performance on NaturalScore. For the 14B model, RefAlign likewise achieves the SOTA TotalScore, further validating its effectiveness across different model scales. To the best of our knowledge, RefAlign is the first model to achieve a TotalScore of 60.42% among existing publicly reported results.

4.3 Qualitative Results

Fig. 5 presents qualitative comparisons between RefAlign and existing SOTA R2V methods. Overall, RefAlign achieves a better balance between preserving the appearance consistency of the reference subject and following the text prompt, producing videos with more stable temporal coherence and smoothness. In contrast, Kling1.6 shows noticeable copy–paste artifacts in some examples (top-left), and the subject fidelity may degrade in certain scenarios (bottom-left); Phantom exhibits similar issues. In addition, both methods appear to underutilize the background reference image in some cases, leading to reduced background consistency (bottom-left). The strongest open-source baseline, VINO, still shows erroneous copying of clothing textures in some results (top-right), and its instruction following becomes weaker under some prompts (bottom-right).

4.4 Ablation Study

To validate the effectiveness of the RA loss, we conduct a systematic ablation study from three aspects: loss design, alignment depth, and the alignment encoder. Notably, all experiments are conducted under a unified setting of ***1800 training iterations*** to ensure a fair comparison. The baseline configuration (w/o $\mathcal{L}_{\text{RA}}$) only encodes the reference image with the VAE and feeds it into the DiT, without introducing the RA loss.

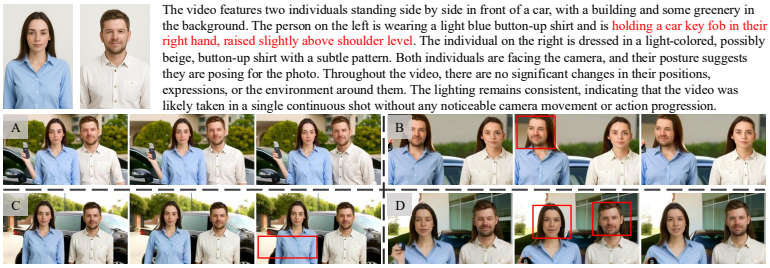


Fig. 6: Qualitative Ablation Study. RA loss (A) improves reference fidelity and instruction following. Removing the negative loss (B) leads to multi-reference confusion; removing RA loss (C) causes copy–paste artifacts; using DINOv3 features as input (D) reduces subject fidelity.

Effectiveness of RA loss. To validate the design of the proposed RA loss, we conduct an ablation study with different configurations. As shown in Tab. 2, configuration A achieves the best TotalScore (55.73%), indicating a favorable trade-off. Configuration B lowers TotalScore mainly due to drops in FaceSim and NaturalScore, showing $\mathcal{L}_{\text{neg}}$ is important for suppressing incorrect alignments and improving naturalness. Without RA loss (configuration C), TotalScore drops sharply (55.73%→49.93%) while FaceSim increases markedly, indicating RA loss mitigates copy–paste artifacts. Configuration D improves over C (52.15% vs. 49.93%), but still trails A, suggesting direct alignment to DINOv3 is more effective than using DINOv3 as an extra input, likely due to reduced modality mismatch and more stable fine-tuning.

The qualitative comparisons in Fig. 6 visually support the above analysis. Configuration B shows clear confusion in gender-related attributes; configuration C retains high facial fidelity but fails to follow "holding a car key"; configuration D better follows the action instruction, yet facial fidelity drops and slight identity confusion emerges. Overall, configuration A offers a more stable trade-off between instruction following and identity consistency.

Effectiveness of alignment depth. To study the effect of alignment depth for the RA loss, we treat w/o $\mathcal{L}_{\text{RA}}$ as depth = 0 and compare different depths. As shown in Fig. 7 (a), TotalScore follows an "increase-then-decrease" trend and peaks at depth = 9, suggesting that moderate depth yields a favorable trade-off. FaceSim decreases monotonically with depth (Fig. 7 (b)), indicating that overly deep alignment weakens subject identity fidelity, while NaturalScore generally increases with minor fluctuations (Fig. 7 (c)), indicating that deeper alignment improves naturalness. Based on this observation, we set depth = 9 by default.

Effectiveness of alignment encoder. To evaluate the impact of the alignment encoder on the RA loss, we ablate encoder size and type. As shown in Tab. 3, using an alignment encoder consistently outperforms w/o $\mathcal{L}_{\text{RA}}$ under both settings, suggesting the benefit of external representations as RA targets. **In the size ablation**, DINOv3-B/L/H+ all bring substantial gains, with only 0.33% to 0.43% variation in TotalScore, indicating that RA loss is robust to encoder size. Considering the trade-off between performance and computational cost, we

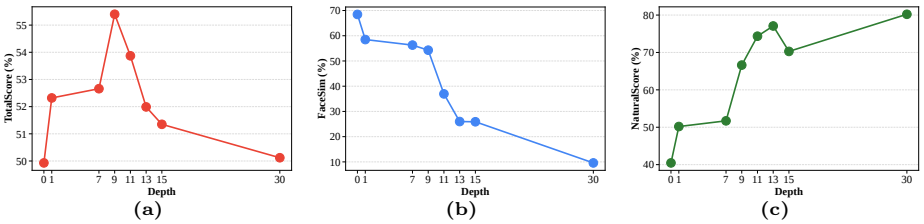


Fig. 7: Effect of alignment depth on three key metrics. Subfigures (a), (b), and (c) illustrate how alignment depth affects TotalScore, FaceSim, and NaturalScore, respectively. We treat w/o $\mathcal{L}_{RA}$ as depth 0.

Table 3: Ablation study on the impact of the alignment encoder in RefAlign on the OpenS2V-Eval.

Encoder	TotalScore↑	Aesthetics↑	MotionSmoothness↑	MotionAmplitude↑	FaceSim↑	GmeScore↑	NexusScore↑	NaturalScore↑
w/o $\mathcal{L}_{\text{RA}}$	49.93%	36.46%	88.97%	28.62%	68.45%	58.83%	38.63%	40.46%
Encoder size ablation								
DINOv3-B	55.40%	38.83%	93.04%	22.49%	54.28%	66.07%	41.92%	66.62%
DINOv3-L	**55.73%**	41.53%	90.00%	29.53%	53.15%	67.54%	46.23%	62.96%
DINOv3-H+	55.30%	43.72%	94.25%	20.09%	53.07%	66.90%	44.24%	61.48%
Encoder type ablation								
DINOv3-L	**55.73%**	41.53%	90.00%	29.53%	53.15%	67.54%	46.23%	62.96%
SigLIP2-So	54.25%	43.95%	94.79%	17.90%	46.00%	68.43%	40.81%	65.00%
Qwen2.5-VL-7B	53.12%	37.50%	92.09%	32.13%	44.96%	64.40%	45.07%	63.43%

use DINOv3-L as the default. **In the type ablation**, the improvement generalizes across different encoders (DINOv3-L, SigLIP2-So, and Qwen2.5-VL-7B), indicating it is not tied to a specific architecture. They also show distinct metric preferences: DINOv3-L performs best on consistency-related metrics (e.g., NexusScore and FaceSim), SigLIP2-So favors quality and naturalness metrics (e.g., Aesthetics, MotionSmoothness, and NaturalScore), while Qwen2.5-VL-7B performs better on MotionAmplitude. Overall, DINOv3-L achieves a favorable trade-off.

4.5 User Study

We conduct a user study with 30 volunteers evaluating visual quality, reference fidelity, and video–text alignment. For each comparison pair, two videos from different methods are shown anonymously in random order, and participants choose the better one or a tie. As shown in Fig. 8, RefAlign achieves the highest overall human preference.

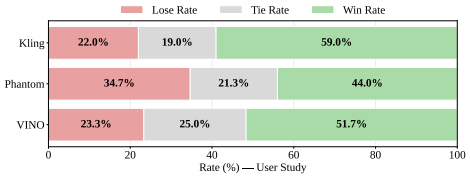


Fig. 8: User study results comparing RefAlign with Kling, Phantom, and VINO.

5 Conclusion

This paper presents RefAlign, a representation alignment framework to enhance controllability in R2V generation. RefAlign aligns the reference-branch representation to an underlying vision foundation model (VFM) at the feature level via a reference alignment (RA) loss, helping alleviate copy–paste artifacts and multi-subject confusion from modality mismatch. The alignment is applied only during training and removed at inference, incurring no inference-time overhead. Extensive experiments show that RefAlign outperforms SOTA methods (e.g., Kling1.6, Phantom, and VINO) in reference subject fidelity and consistency. We hope this work provides new insights into reference-driven video generation and video editing, and further may promote a unified modeling paradigm for visual understanding and generation.

Limitations and future work. Limited training data diversity currently prevents an optimal balance between instruction following and reference fidelity. Moreover, the underlying foundation model limits RefAlign to 81-frame videos, restricting long-video generation. In addition, a single VFM guidance signal may be incomplete, as different VFM features emphasize different aspects of visual representations. Future work includes using more diverse data to improve the trade-off, extending R2V to longer videos, and combining multiple VFM features as alignment targets for more robust and comprehensive alignment.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62361166670 and U24A20330, and by the Supercomputing Center of Nankai University (NKSC).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 3, 5, 9
2. Bao, F., Xiang, C., Yue, G., He, G., Zhu, H., Zheng, K., Zhao, M., Liu, S., Wang, Y., Zhu, J.: Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv:2405.04233 (2024) 2, 10, 11
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024) 2
4. Chen, J., He, T., Fu, Z., Wan, P., Gai, K., Ye, W.: Vino: A unified visual generator with interleaved omnimodal context. arXiv preprint arXiv:2601.02358 (2026) 5, 10, 11
5. Chen, S., Ge, C., Zhang, Y., Zhang, Y., Zhu, F., Yang, H., Hao, H., Wu, H., Lai, Z., Hu, Y., et al.: Goku: Flow based video generative foundation models. In: CVPR. pp. 23516–23527 (2025) 2

6. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation. In: CVPR. pp. 6099–6110 (2025) 2, 4, 7
7. Chen, Z., Li, B., Ma, T., Liu, L., Liu, M., Zhang, Y., Li, G., Li, X., Zhou, S., He, Q., Wu, X.: Phantom-data: Towards a general subject-consistent video generation dataset. ICLR (2026) 11
8. Deng, Y., Guo, X., Wang, Y., Fang, J.Z., Wang, A., Yuan, S., Yang, Y., Liu, B., Huang, H., Ma, C.: Cinema: Coherent multi-subject video generation via mllm-based guidance. arXiv preprint arXiv:2503.10391 (2025) 2, 4, 5
9. Deng, Y., Guo, X., Yin, Y., Zhiyuan Fang, J., Yang, Y., Wang, Y., Yuan, S., Wang, A., Liu, B., Huang, H., et al.: Magref: Masked guidance for any-reference video generation. ICLR (2026) 5, 10, 11
10. 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: ICML (2024) 6
11. Fei, Z., Li, D., Qiu, D., Wang, J., Dou, Y., Wang, R., Xu, J., Fan, M., Chen, G., Li, Y., et al.: Skyreels-a2: Compose anything in video diffusion transformers. arXiv preprint arXiv:2504.02436 (2025) 2, 4, 10, 11
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop on Deep Generative Models and Downstream Applications (2021) 7
13. Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., Lu, Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation. arXiv preprint arXiv:2505.04512 (2025) 5
14. Hu, T., Yu, Z., Zhou, Z., Zhang, J., Zhou, Y., Lu, Q., Yi, R.: Polyvivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. NeurIPS (2025) 5
15. Huang, Y., Yuan, Z., Liu, Q., Wang, Q., Wang, X., Zhang, R., Wan, P., Zhang, D., Gai, K.: Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv preprint arXiv:2501.04698 (2025) 2, 4, 5
16. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: ICCV. pp. 17191–17202 (2025) 5, 10, 11
17. kling: Image to video elements feature. https://klingai.com/image-to-video/multi-id/new/ (2024) 3, 10, 11
18. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Boosting generative image modeling via joint image-feature synthesis. NeurIPS (2025) 5
19. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In: ICCV. pp. 18262–18272 (2025) 5
20. Li, D., Zhong, W., Yu, W., Pan, Y., Zhang, D., Yao, T., Han, J., Mei, T.: Pursuing temporal-consistent video virtual try-on via dynamic pose interaction. In: CVPR. pp. 22648–22657 (2025) 2
21. Li, Z., Qian, D., Su, K., Diao, Q., Xia, X., Liu, C., Yang, W., Zhang, T., Yuan, Z.: Bindweave: Subject-consistent video generation via cross-modal integration. ICLR (2026) 2, 4, 5, 10, 11
22. Liang, F., Ma, H., He, Z., Hou, T., Hou, J., Li, K., Dai, X., Juefei-Xu, F., Azadi, S., Sinha, A., et al.: Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In: CVPR. pp. 13146–13156 (2025) 2
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023) 5

24. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. In: ICCV. pp. 14951–14961 (October 2025) 2, 4, 7, 10, 11
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. ICLR (2019) 11
26. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html 3
27. Nguyen, H., Nguyen, Q.Q.V., Nguyen, K., Nguyen, R.: Swifttry: Fast and consistent video virtual try-on with diffusion models. In: AAAI. vol. 39, pp. 6200–6208 (2025) 2
28. 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. TMLR (2024) 6
29. Pan, P., Zhao, J., Lin, Y., Lin, C., Li, C., Liu, H., Shen, T., Mu, Y.: Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. CVPR (2026) 2, 5
30. Pika: Pikascenes. https://pika.art/ingredients/ (2024) 10, 11
31. 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: ICML. pp. 8748–8763. PmLR (2021) 5
32. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR **21**(140), 1–67 (2020) 6
33. Sang, S., Zhi, T., Gu, T., Liu, J., Luo, L.: Lynx: Towards high-fidelity personalized video generation. In: CVPR (2026) 4
34. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 3, 9
35. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025) 2
36. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 9
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 2, 4, 6, 10, 11
38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 3
39. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. CVPR (2026) 5
40. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025) 2
41. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. NeurIPS (2025) 5
42. Xie, X., Liu, J., Lin, Z., Fan, H., Han, Z., Tang, Y., Qu, L.: Unleashing the potential of large language models for text-to-image generation through autoregressive representation alignment. AAAI (2026) 5

43. Xue, B., Duan, Z.P., Yan, Q., Wang, W., Liu, H., Guo, C.L., Li, C., Li, C., Lyu, J.: Stand-in: A lightweight and plug-and-play identity control for video generation. CVPR (2026) 4
44. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025) 2
45. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: CVPR. pp. 15703–15712 (2025) 5
46. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: ICLR (2025) 5, 6, 9
47. Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Lin, B., Luo, J., Yuan, L.: Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. NeurIPS (2025) 4, 11
48. Yuan, S., Huang, J., He, X., Ge, Y., Shi, Y., Chen, L., Luo, J., Yuan, L.: Identity-preserving text-to-video generation by frequency decomposition. In: CVPR. pp. 12978–12988 (2025) 4
49. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models. NeurIPS (2025) 5
50. Zhang, Z., Teng, J., Yang, Z., Cao, T., Wang, C., Gu, X., Tang, J., Guo, D., Wang, M.: Kaleido: Open-sourced multi-subject reference video generation model. arXiv preprint arXiv:2510.18573 (2025) 4, 5, 10, 11
51. Zhong, Y., Yang, Z., Teng, J., Gu, X., Li, C.: Concat-id: Towards universal identity-preserving video synthesis. In: ICCVW. pp. 1906–1915 (2025) 4
52. Zhou, Z., Liu, S., Liu, H., Qiu, H., An, Z., Ren, W., Liu, Z., Huang, X., Ng, K.W., Xie, T., et al.: Scaling zero-shot reference-to-video generation. In: CVPR. pp. 9253–9262 (2026) 5, 10, 11