AlcheMinT: Fine-grained Temporal Control for Multi-Reference Consistent Video Generation

Supplementary Material

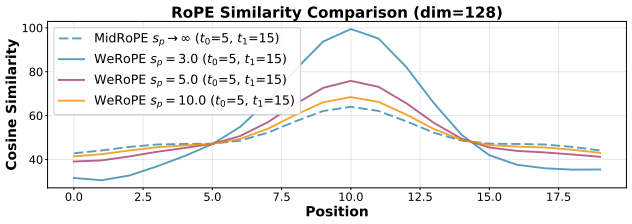

Figure 1. **Effect of scaling factor for WeRoPE.**

We provide additional details, and qualitative/quantitative results from the main paper. Sec. A includes further ablations with variants closer to our final approach. Sec. B discusses the details of our data collection pipeline in depth, Sec. C provides information about our inference CFG setup with multi text and reference conditions, and Sec. D provides background information for RoPE. We provide qualitative visualizations in Sec. E and computational analysis in Sec. F. Please refer to our project website (https://snap-research.github.io/Video-AlcheMinT) for additional video visualizations.

A. Ablations

We ablate the design of our pipeline, namely text conditioning, as well as WeRoPE, on our benchmark similar to Table 2 in the main paper but also including dense captions via CrossAttention + ReRoPE [8]. For a fair comparison, we compare with models trained for 25K iterations on a subset of the test set with 2 references, to highlight the effect of disentanglement. We include the baseline setting of no temporal RoPE for the reference tokens as well, effectively having a 0 timestamp. Results are summarized in Tab. 1. Similar to the observations in Sec. 4.3 of the main paper, we see that including reference text embeddings improves overall subject preservation with higher $\text{CLIP}_{\text{text}}$, CLIP_{ref} while maintaining similar timestamp following with lower t-L2, but also marginally lower t-IOU. This is due to text conditioning not affecting timestamp following of the generated video but to disambiguate appearances between multiple subjects as highlighted by the CLIP scores. Additionally, not providing any temporal RoPE for reference tokens increases the error in temporal following even further highlighting the importance of the RoPE mechanism for controlling the timing of the references.

Effect of WeRoPE parameters. We provide additional details regarding the scaling factors used for WeRoPE. We use scaling factors, positive s_p and negative $s_n = -1$ and define $w_p = \frac{s_p}{s_p+2*s_n}$, $w_n = \frac{s_n}{s_p+2*s_n}$. We visualize the RoPE similarity curve in Fig. 1 by varying s_p. We set $s_p = 5.0$ for our experiments as $s_p = 3.0$ introduces training instability with large activation magnitudes as seen by the cosine similarity, while $s_p = 10.0$ marginally reduces the attention score decay with interval length. $s_p \to \infty$ converges to MidRoPE which cannot control decay rate with any event length. $s_p = 5.0$ corresponds to values of $w_p \approx 1.67, w_n \approx -0.33$ which we use as the default values for all our experiments.

Table 1. **Ablation study.** We ablate the effect of incorporating text conditioning into the network as well as the 2 RoPE variants of MidRoPE and WeRoPE, by utilizing dense captions [8] into the model.

Variant	t-L2↓	t-IOU↑	$\text{CLIP}_{\text{text}}$ ↑	CLIP_{ref} ↑
w/o Ref. Text Embedding	0.313	**0.419**	0.231	0.724
w/ Ref. Text Embedding	**0.283**	0.402	**0.251**	**0.761**
No RoPE	0.360	0.323	0.234	0.728
MidRoPE	0.336	0.346	0.221	0.702
WeRoPE	**0.294**	**0.368**	**0.250**	**0.757**

B. Data collection pipeline details

Our dataset is built on top of an existing video-text paired datasets. We additionally collect dense timed captions for each temporal event similar to [8].

B.1. Entity Word Extraction

We extract, for every dense caption, the set of phrases that denote *physically groundable* entities, similar to [3]. We use Qwen 2.5 [1] with a fixed prompt template that includes the input caption and task description, and we augment the prompt with the following constraints:

- Each entity phrase is an exact substring of the caption.
- When multiple entities share similar labels, disambiguate using adjectives or referring expressions present in the caption.
- Exclude terms from a predefined blacklist (e.g., `letters`, `text`, and generic body parts such as `arm`, `leg`).

This procedure yields an open-vocabulary inventory of entity phrases rather than a closed set of classes, and it naturally supports reference disambiguation (e.g., the two men are distinguished by their local modifiers and roles in the

Figure 2. **Segmented track visualization.** We show segmented tracks obtained via our data collection pipeline for a wide variety of objects along with the original videos. Objects appearing/disappearing are tracked consistently across frames with complex word descriptions.

caption). This also allows the model to bind word tags with the captions through our text conditioning strategy.

In practice, we occasionally observe intangible or non-physical descriptors (e.g., *left*, *right*, *area*) or blacklist violations, likely due to the model jointly performing extraction and rule following. To address this, we apply a lightweight post-filtering pass: the extracted list is fed back to Qwen with instructions to remove blacklist items and any terms that do not denote physically groundable entities.

We provide an example input caption and output pair below.

```
Caption: The camera dollies back and pans
    to the left showing a full shot of a
    fair-skinned man with a short brown
    beard, wearing a red t-shirt, blue jeans
    with white shoes, handing over the keys
    to the fair-skinned man standing on the
    left. The man has short black hair and
    beard wearing a white t-shirt with grey
    jeans and grey shoes.
Output: ['fair-skinned man with a short
    brown beard', 'fair-skinned man standing
    on the left', 'red t-shirt', 'blue
    jeans', 'white shoes', 'short black hair
    ', 'beard', 'white t-shirt', 'grey jeans
    ', 'grey shoes']
```

B.2. Grounded Entity Mask Tracking

To obtain temporally consistent instance masks for all entities across a video, we combine Grounding DINO [6] for text-conditioned detections with SAM2 [5] for mask track-

ing.

Entity detections. For each entity keyword, we take its associated dense caption and time interval in the video. We sample the $10^{\text{th}}, 50^{\text{th}}, 90^{\text{th}}$ percentile frames within that interval and run Grounding DINO using the keyword as the text prompt. Within each frame, we apply non-maximum suppression (NMS) to remove duplicate boxes, then select the remaining detection with the highest CLIP similarity to the keyword. Each selected detection stores the tuple: {word tag, caption and interval, frame index, bounding box}.

Tracking and mask propagation. For every detection, we invoke SAM2 with the detection box as a box prompt at the detection frame index and track the instance forward and backward to produce a per-frame mask track over the full interval.

Deduplication. Because multiple detections (from different dense captions) may correspond to the same physical entity, we compute the average mask IoU over all overlapping frames between track pairs and remove duplicates whose average IoU exceeds a threshold.

Post-processing. We remove tracks that are unlikely to be valid instances: (i) person-category tracks without any face detection, (ii) tracks whose area falls below a threshold, and (iii) a number of outliers/erroneous detections via a final manual pass.

Result. The resulting corpus contains long videos with time-aligned dense captions and multiple consistently tracked references (up to 15 per video). Consistent masks provide both reliable presence timestamps for each entity and enable mask-based data augmentation by sampling entity masks that lie outside the sampled training frames, thereby providing complex pose/lighting/appearance changes. We visualize some of the annotation results obtained in Fig. 2.

C. Multi-CFG

Due to presence of multiple input conditions in terms of reference text, images, time intervals as well as global, dense captions, we use multiple passes for Classifier-Free-Guidance [4](CFG) for the different conditions. However, all combinations of the input conditions would be prohibitively expensive growing exponentially. We therefore group the reference text and images into a joint reference condition for dropping. We also group the global and dense captions as a joint text condition similar to [8]. We do not drop/zero out reference time intervals as altering WeRoPE leads to undesirable patchy artifacts. As highlighted in [2, 3], we obtain the multi-CFG equation as

$$\begin{aligned}\tilde{e}_\theta(z_t, c_{\text{ref}}, c_{\text{text}}) = \quad & e_\theta(z_t, c_{\text{ref}}, c_{\text{text}}) \\ & + w_{\text{text}} \cdot \big(e_\theta(z_t, c_{\text{ref}}, c_{\text{text}}) - e_\theta(z_t, c_{\text{ref}}, \varnothing)\big) \\ & + w_{\text{ref}} \cdot \big(e_\theta(z_t, c_{\text{ref}}, c_{\text{text}}) - e_\theta(z_t, \varnothing, c_{\text{text}})\big) \\ & + w_{\text{both}} \cdot \big(e_\theta(z_t, c_{\text{ref}}, c_{\text{text}}) - e_\theta(z_t, \varnothing, \varnothing)\big).\end{aligned}$$

with $e_\theta(z_t, c_{\text{ref}}, c_{\text{text}})$ being the score estimation function with the reference and text conditioning, c_{ref} and c_{text} respectively at denoising timestep t. We set $w_{\text{text}} = 8, w_{\text{ref}} = 2$, and $w_{\text{both}} = 3$. We perform denoising via rectified flow sampling for 40 steps at 288×512 resolution with time-shifting value of 5.66.

D. RoPE premilinaries

RoPE injects position by rotating each 2D feature subspace of a token with a phase that depends on its index [7]. Let the model dimension be d (even) and define a bank of angular frequencies

$$\boldsymbol{\omega} = \{\omega_i\}_{i=0}^{\frac{d}{2}-1}, \qquad \omega_i = 10000^{-2i/d}.$$

For a 1D index $n \in \mathbb{Z}$, the phase vector is $\boldsymbol{\phi}(n) = \{\, \phi_i(n) = n\,\omega_i \,\}_i$, grouping real coordinates into complex pairs $z_i = z_{2i} + \mathrm{i} z_{2i+1}$, $i = 0, \ldots, \frac{d}{2} - 1$. RoPE applies a rotation (complex multiplication) independently to each pair:

$$\langle \hat{\boldsymbol{q}}, \hat{\boldsymbol{k}} \rangle = \mathrm{Re}\left[\sum_{i=0}^{\frac{d}{2}-1} q_i (k_i)^* \exp\Big\{\mathrm{i}\big(\phi_i(m) - \phi_i(n)\big)\Big\}\right] \quad (\text{with } q_i = q_{2i} + \mathrm{i} q_{2i+1},\ k_i = k_{2i} + \mathrm{i} k_{2i+1}). \tag{1}$$

3D extension. For video tokens with coordinates $m = (x, y, t)$, channel pairs are split across the three axes (e.g., $d_x + d_y + d_t = d$) and use axis-specific frequency banks $\boldsymbol{\omega}^{(x)}, \boldsymbol{\omega}^{(y)}, \boldsymbol{\omega}^{(t)}$. A convenient notation is to write the total phase for pair i as the sum of axis phases,

$$\phi_i(m) \;=\; x\,\omega_i^{(x)} + y\,\omega_i^{(y)} + t\,\omega_i^{(t)},$$

and apply the same rotation rule per pair.

Relative-position property. Let $\boldsymbol{q}, \boldsymbol{k} \in \mathbb{R}^d$ be query/key at positions m and n, and let q_i, k_i be their complex pairs. After RoPE, the dot product depends only on the *relative* position:

$$\langle \hat{\boldsymbol{q}}, \hat{\boldsymbol{k}} \rangle \;=\; \mathrm{Re}\left[\sum_{i=0}^{\frac{d}{2}-1} q_i\,(k_i)^*\, e^{\mathrm{i}\,(\phi_i(m) - \phi_i(n))}\right].$$

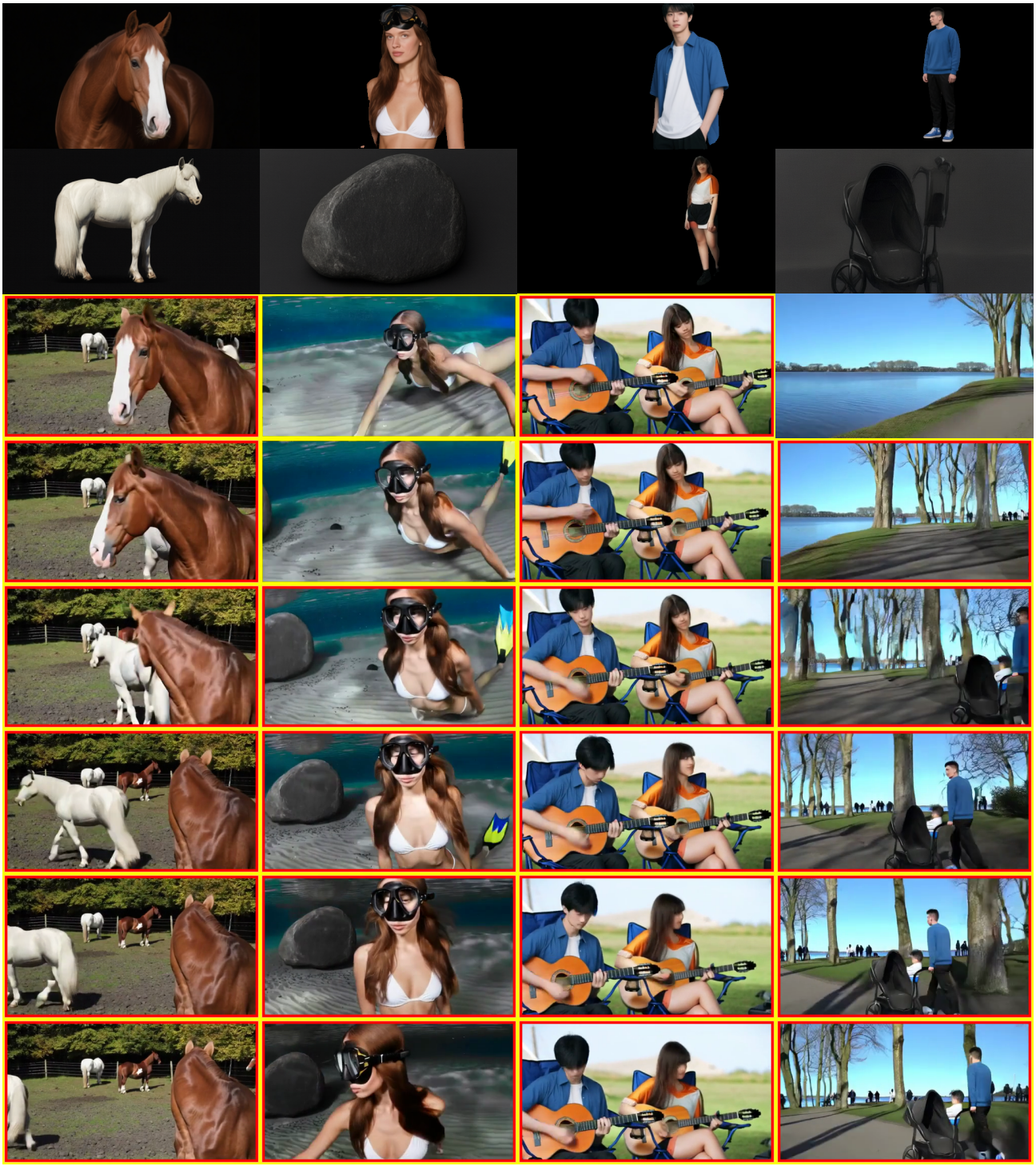

Figure 3. **Additional qualitative visualizations.** Yellow boxes indicate the expected occurence for the first reference and red boxes for the second. We obtain high fidelity generations which closely follow the input time interval.

In the common 1D case with $\phi_i(n) = n\,\omega_i$, this reduces to

$$\mathrm{Re}\left[\sum_{i=0}^{\frac{d}{2}-1} \mathbf{q}_{[2i:2i+1]}\,\mathbf{k}^*_{[2i:2i+1]}\,e^{\mathrm{i}(m-n)\,\omega_i}\right],$$

where $\mathbf{q}_{[2i:2i+1]} = q_{2i} + \mathrm{i}q_{2i+1}$ and similarly for $\mathbf{k}$, and $\mathrm{Re}[\cdot]$ denotes the real part. This complex/rotation view makes clear that RoPE enforces a phase difference proportional to the relative displacement, yielding a natural induc-

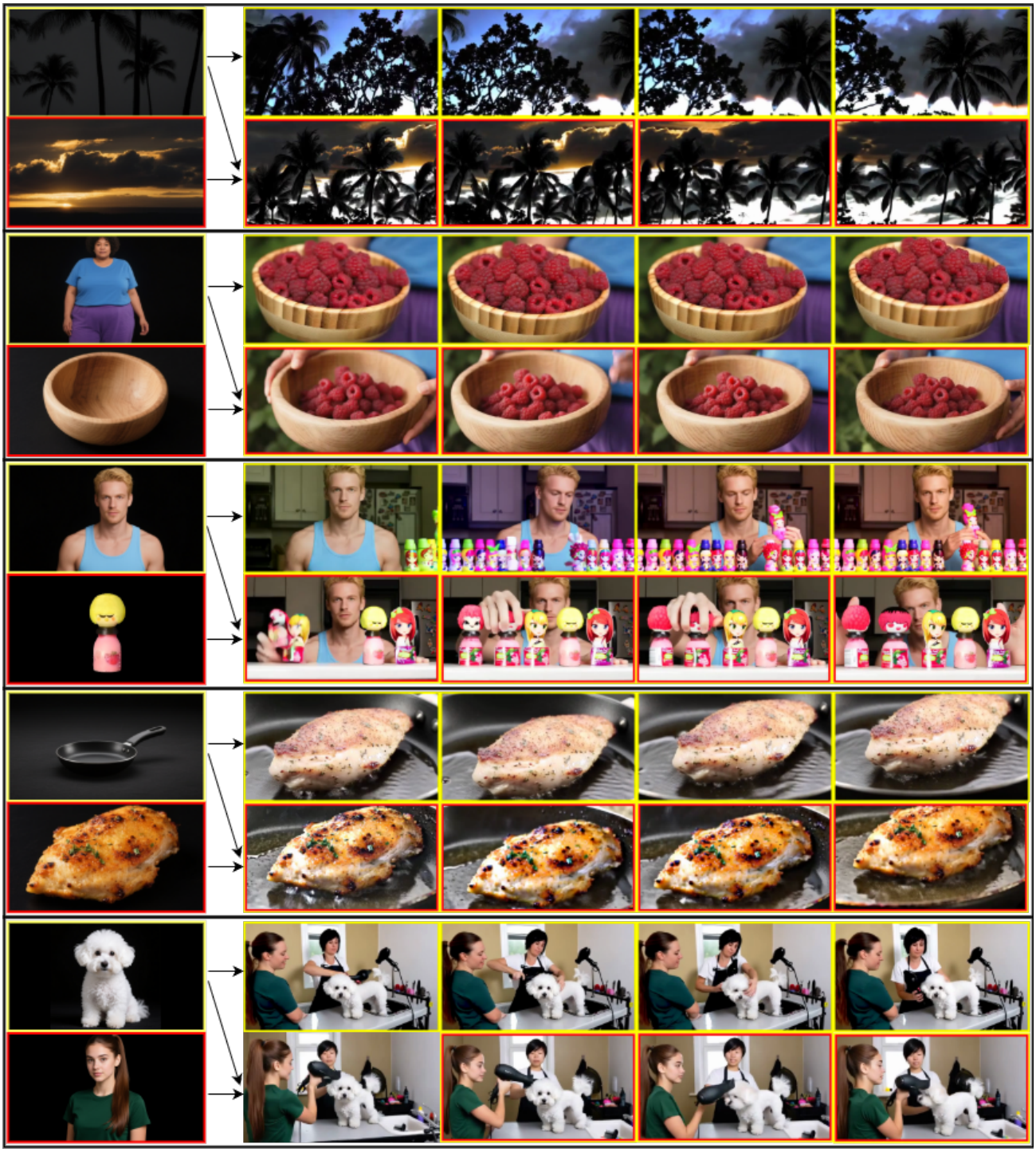

Figure 4. **Varying number of references .** We qualitatively visualize videos generated with a single reference as input or 2 references. We use the same set of text prompts as input for both generations. As expected, we see that including the second reference produces generations which are more consistent with the input preserving the identity appearance information.

tive bias for long-range, relative attention.

E. Additional visualizations

In addition to the qualitative visualizations in the main paper, we provide further results on our benchmark videos.

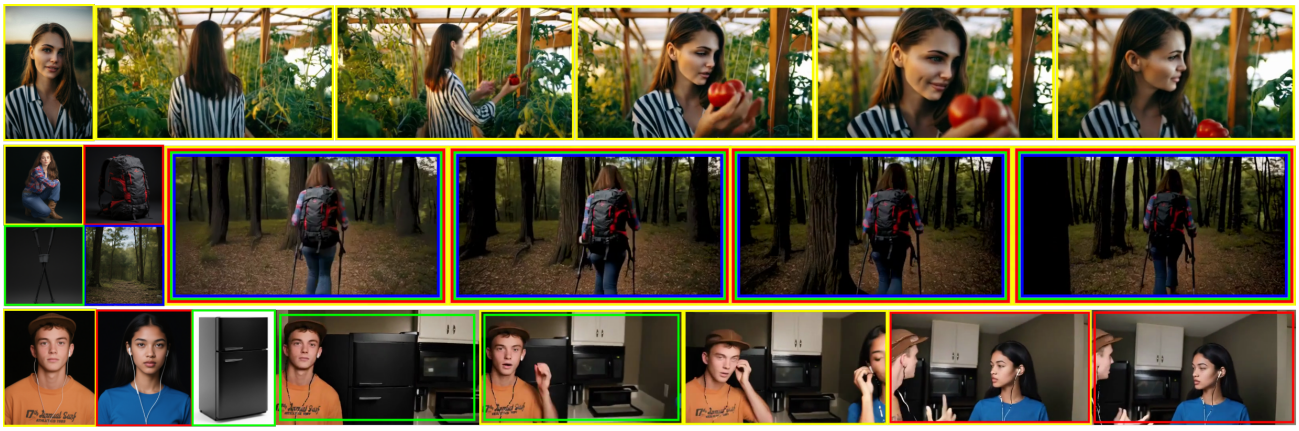

Figure 5. **Longer video and more reference generation.** We show results with longer videos upto 10 seconds in the first row as well as generations with 3 or 4 references as input (2nd,3rd row). We obtain consistent realistic generations while also following the reference timestamps as observed in the 3rd row.

The supplementary material includes short video files demonstrating generations with single- and two-reference inputs, as well as frame-wise visualizations for multi-reference cases (Figs. 3 and 4). In the figures, yellow boxes mark the expected presence window of the first reference (from the input interval), and red boxes mark the second reference. As illustrated in Fig. 3, the model produces high-quality, reference-consistent samples that largely respect the specified timing (e.g., the rock, red box, in column 2; the man with stroller in column 4). Fig. 4 also shows generations with only the first reference, and both references with the same prompt. We see that providing the second reference produces generations consistent with the image in terms of its corresponding attributes showing that the model is able to produce consistent generations with the subject/reference condition.

Note that there are slight mismatches between interval inputs and when a reference actually appears in the generation due to a) the error in temporal downsampling for the latents (by a factor of 4) and also b) RoPE producing a gradual decay in attention score and not a sharp falloff as visualized in Fig. 3 of the main paper. This is however, necessary, for producing smooth generations with references gradually appearing in the video without abrupt unnatural transitions.

Longer video or larger multi-reference generation. Our method can extend beyond 2 references and 6s long videos by incorporating training data with more references, and longer length videos. We train model variants with increased number of input references and longer videos. Fig. 5 shows a 10s long generated video (1st row) or with 3/4 references (2nd,3rd row), showing our scalability to arbitrarily large number of input references or longer videos simply by scaling up the training data due to the flexible nature of our architecture.

F. Computation costs

We briefly discuss the computation costs and overhead for our approach. We introduce a negligible amount of additional parameters compared to the baseline MinT model [8]. The dense cross attention branch from MinT introduces a 10% parameter overhead, while the text MLP results in a 3% increase. Each reference index embedding has the latent dimension (4096), initialized with standard normal. It is directly added to its reference tokens, with negligible parameter/inference cost. Therefore, while it scales linearly with respect to number of references, it is orders of magnitude smaller than the number of total parameters in the DiT which can be several billion in scale. In terms of FLOPs or inference speed, while the number of reference tokens increases linearly, each reference has $1/20^{th}$ number of video tokens for a 6s video (73 frames with 4x temporal downsample), thus having a 5% inference overhead, which reduces for longer videos. Further computational gains can be obtained by excluding tokens belonging to masked regions of the reference. Therefore, our model induces only a small compute overhead and even smaller parameter overhead compared to the base DiT architecture.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

[3] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *CVPR*, 2025. 1, 3

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[7] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 3

[8] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *CVPR*, 2025. 1, 3, 6