
UniTransfer: Video Concept Transfer via Progressive Spatial and Timestep Decomposition

Anonymous Author(s)

Affiliation

Address

email

Web Page: <https://anonymous-name-9.github.io/neurips-anonymous/>

1 More Details about self-supervised Pretraining

In practice, obtaining large amounts of high-quality annotated data remains challenging. Although efficient segmentation tools like SAM2[1] are available, they still require extensive manual interactions, such as point prompts, bounding boxes, or object-specific filtering. To address this limitation and reduce the reliance on manual annotations, we introduce a self-supervised pretraining approach, the detailed algorithmic pipeline of which is outlined below Algorithm 1 and Figure 1.

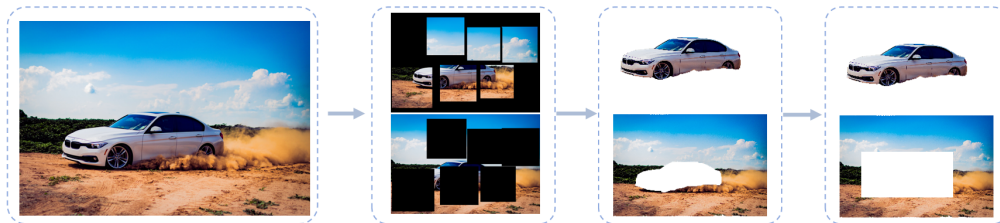


Figure 1: Self-supervised Pretrain

2 More Details about CoP Guidance

The detailed algorithmic pipeline of our Chain-of-Prompt (CoP) guidance is demonstrated in Algorithm 2.

3 More Results

Our framework enables flexible foreground and background transfer, including part-level object replacement, such as garment transfer. The results of different transfer tasks are shown in Figure 2, Figure 3, Figure 4. Our curated animal-centric dataset OpenAnimal is demonstrated in Figure 5.

4 Limitation

Although our model can achieve subject transfer and background replacement in videos, there are still cases where the subject and background appear with artifacts. This issue might be due to current

Algorithm 1: The pipeline of our self-supervised pretraining.

Input: *image*: input image

coverage: target coverage ratio (default: 0.5)

min_block_size: minimum block size (default: 320)

max_block_size: maximum block size (default: 640)

Output: *foreground*: processed image with coverage

background: inverse masked image

Function *random_white_blocks*(*image*, *coverage*, *min_block_size*, *max_block_size*)

if *image* is *None* **then**

 raise *ValueError*("Input image is empty");

if *image* is *grayscale* **then**

result \leftarrow convert *image* to BGR color space;

else

result \leftarrow copy of *image*;

 (*h*, *w*) \leftarrow height and width of *result*;

total_area $\leftarrow h \times w$;

covered_area $\leftarrow 0$;

target_area \leftarrow *total_area* \times *coverage*;

mask \leftarrow zero matrix of size (*h*, *w*);

result_2 \leftarrow gray matrix (127.5) with same size as *result*;

max_iter $\leftarrow 500$;

while *covered_area* < *target_area* **and** *max_iter* > 0 **do**

max_iter \leftarrow *max_iter* - 1;

block_size \leftarrow random integer between *min_block_size* and *max_block_size*;

x \leftarrow random integer between 0 and *w* - *block_size*;

y \leftarrow random integer between 0 and *h* - *block_size*;

if *mask*[*y* : *y* + *block_size*, *x* : *x* + *block_size*] contains any 1 **then**

 continue;

 Special overlapping effect *result_2*[*y* : *y* + *block_size* - 5, *x* : *x* + *block_size* - 5] \leftarrow

result[*y* : *y* + *block_size* - 5, *x* : *x* + *block_size* - 5];

result[*y* : *y* + *block_size* + 5, *x* : *x* + *block_size* + 5] \leftarrow [127.5, 127.5, 127.5];

mask[*y* : *y* + *block_size*, *x* : *x* + *block_size*] \leftarrow 1;

covered_area \leftarrow *covered_area* + *block_size* \times *block_size*;

if *covered_area* > 1.1 \times *target_area* **then**

 break;

return *foreground*, *background*;

18 segmentation models cannot fully separate foreground and background elements, leading to imperfect
19 composite results in the final output. In the future, we plan to address this by leveraging large-scale
20 models for enhanced video scene understanding, further improving the quality of generated videos.

21 5 Social Impact

22 Our research decouples video generation into foreground, background, and their corresponding
23 motion. This technology will enhance the efficiency of video production, drive innovation in industries
24 such as film and gaming, and enable more immersive entertainment and educational experiences.
25 However, as this technology becomes more widespread, society may face ethical and legal challenges,
26 including concerns over the authenticity of video content, characters, and backgrounds. Therefore,
27 establishing appropriate regulatory frameworks to ensure responsible use of this technology is a
28 critical task that demands our attention.

```

Input: initial_noisy_video: Initial noisy video
base_prompt: Original text description
total_steps: Total denoising steps (default: 50)
T1, T2: Stage transition steps
Output: generated_video: Final generated video
Function hierarchical_denoising(initial_noisy_video, base_prompt, total_steps = 50)
    // Phase partitioning (all steps)  $T1 \leftarrow 35$ ; // First phase
     $T2 \leftarrow 15$ ; // Second phase
    // Generate hierarchical prompts using LLM
    prompts  $\leftarrow$  QwQ32B_GenerateHierarchicalPrompts(base_prompt);
    // Returns: {'stage1':coarse, 'stage2':detailed, 'stage3':fine}
    current_video  $\leftarrow$  initial_noisy_video;
    for step  $\leftarrow$  total_steps to 1 do
        if step  $\geq T1$  then
            | guidance_prompt  $\leftarrow$  prompts['stage1']; // Coarse prompt
            | guidance_weight  $\leftarrow$  1.5; // Strong guidance
        else if step  $\geq T2$  then
            | guidance_prompt  $\leftarrow$  prompts['stage2']; // Detailed prompt
            | guidance_weight  $\leftarrow$  2.0;
        else
            | guidance_prompt  $\leftarrow$  prompts['stage3']; // Fine prompt
            | guidance_weight  $\leftarrow$  1.0;
        // Execute denoising step current_video  $\leftarrow$ 
        DenoiseStep(current_video, guidance_prompt, guidance_weight, step);
    return current_video;

```

[1] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.



Figure 2: More Visual Results.(Left is reference video, middle is reference image, right is output)

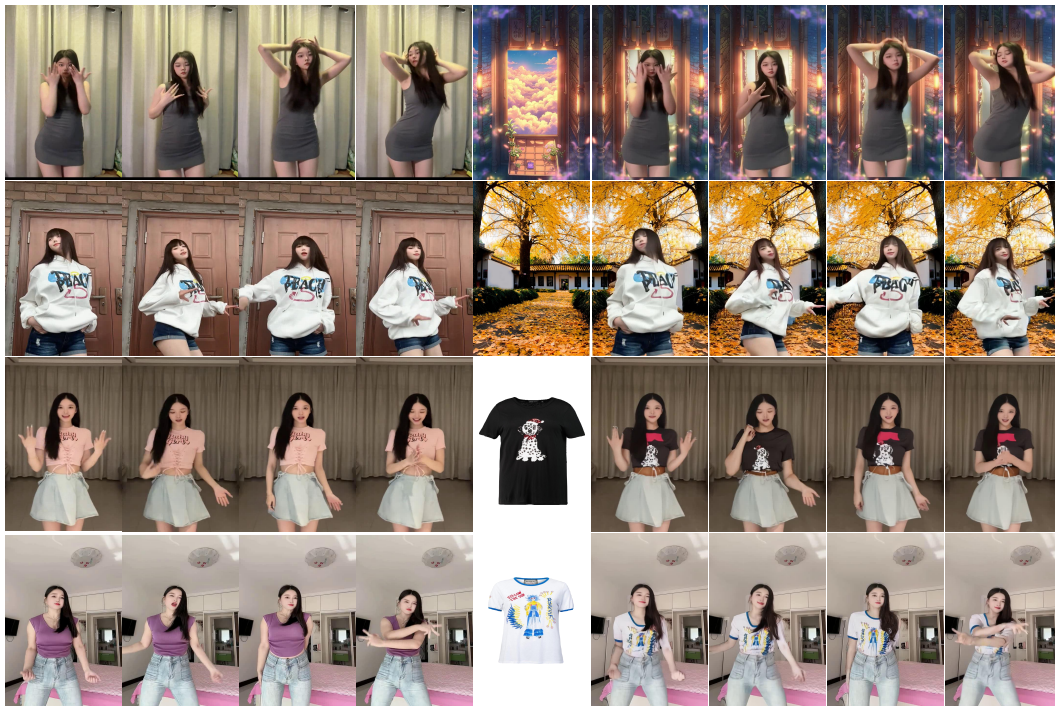


Figure 3: More Visual Results.(Left is reference video, middle is reference image, right is output)



Figure 4: Animal Motion Transfer.(Left is reference video, right is output)



Figure 5: OpenAnimals Datasets