

Stable Part Diffusion: Multi-View RGB and Kinematic Parts Video Generation Supplementary Material

In the appendix, we provide the following supplementary materials: (1) Implementation Details, (2) our newly introduced dataset, KinematicParts20K, (3) Additional Qualitative Results, and (4) Evaluation Details.

A Implementation Details

Our model is implemented by directly extending the SV4D 2.0 framework (Yao et al., 2025). We retain the original U-Net architecture, latent VAE encoding, and diffusion setup, and introduce two key modifications: (1) an architecturally identical second branch that generates part segmentation outputs jointly with the existing RGB branch, and (2) Bidirectional Diffusion Fusion (BiDiFuse) modules inserted between each corresponding layer pair to enable cross-branch feature sharing. In the first stage, the RGB branch is trained following SV4D 2.0. The training setup — including optimizer, noise schedule, loss functions, and sampling strategy — follows SV4D 2.0 exactly. We adopt the EDM (Karras et al., 2022) training framework with an L2 loss and precompute VAE latents and CLIP features for all training images to accelerate convergence. The obtained network parameters are used to initialize both the RGB and part generation branches.

We train the full SPD model with BiDiFuse and our proposed contrastive part consistency loss on the KinematicParts20K dataset (as discussed below) for 40K iterations. Training is performed on 32 NVIDIA H100 GPUs with an effective batch size of 32, using 12 views and 4 frames per object sampled from the rendered dataset.

B KinematicParts20K Dataset

Our dataset is constructed by further filtering the SV4D 2.0 dataset, which is based on CC-licensed dynamic 3D assets from Objaverse and ObjaverseXL. We select only objects that contain rigging annotations, including bone hierarchies and skinning weights. To mitigate overly fine-grained or noisy bone structures, we apply a bone merging procedure based on two criteria: (1) the relative transformation between connected bones across all frames, and (2) the similarity of their projected part appearance in 2D using DINO features. Bone pairs with low motion discrepancy and high appearance similarity are merged. Objects with more than 100 bones after merging are discarded.

All objects are scaled to unit bounding boxes and rendered at 576×576 resolution using Blender’s Cycles renderer under a curated set of HDRI environment maps. We adopt orbit rendering with 24 azimuthal views and 24 video frames per object. In addition to RGB, we simultaneously render per-bone skinning weight maps. For each view and frame, we generate pixel-wise part segmentation labels by taking the argmax over the bone-specific skinning maps, resulting in multi-view, multi-frame kinematic part masks for supervision.

C More Qualitative Results

We show fixed-view cross-frame part tracking, fixed-frame cross-view part tracking, 3D decomposition, rigging, and animation results for synthetic data, real-world data, and zero-shot generated data. Please refer to the summary video in the supplementary material.

D Evaluation Details

D.1 Quantitative Metrics.

To evaluate the quality of kinematic part decomposition across multi-view and multi-frame settings, we report four standard metrics. Since the predicted part masks are label-free, we apply the Hungarian algorithm to align predicted and ground-truth parts based on respective IoU, for those metrics which require correspondences. The following metrics are computed:

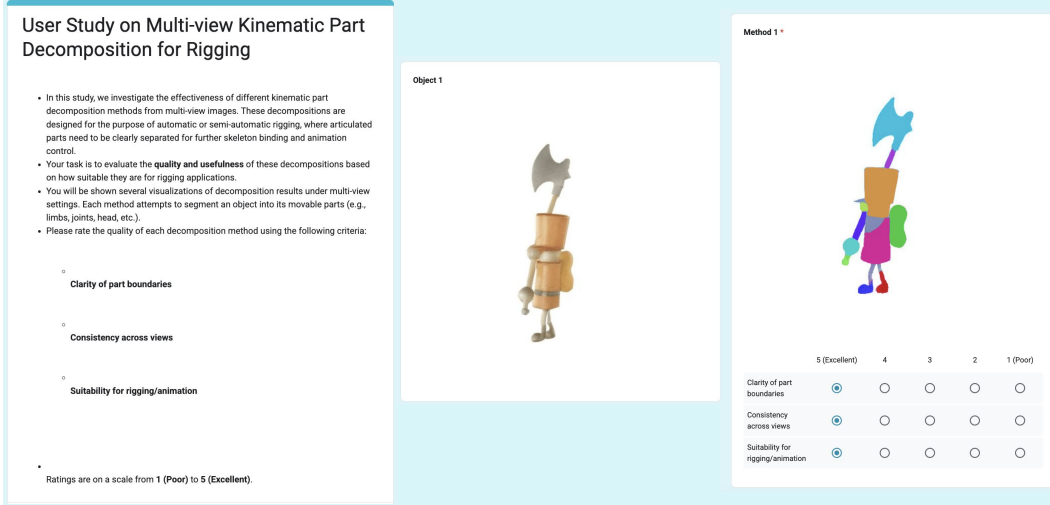


Figure 6: **User study interface for evaluating multi-view kinematic part segmentation.** Participants are presented with video results generated by different methods and asked to rank them based on part consistency, structural correctness, and motion coherence. The study compares SPD with baseline methods to assess perceptual quality and kinematic alignment.

- 852 • **mIoU** – Mean intersection-over-union across matched part masks.
- 853 • **ARI** – Adjusted Rand Index, which captures clustering similarity independent of label
- 854 permutation.
- 855 • **F1 Score** – The harmonic mean of precision and recall, reflecting pixel-level agreement.
- 856 • **mAcc** – Mean class-wise accuracy, indicating the average recall per ground-truth part.

857 D.2 User Study on Multi-view Kinematic Part Decomposition for Rigging

858 To evaluate the practical utility of different multi-view kinematic part decomposition methods for
 859 rigging tasks, we conducted a user study focusing on the perceived quality of part segmentation from
 860 a rigging perspective. These decompositions aim to separate articulated object parts (e.g., limbs,
 861 joints, head) to facilitate automatic or semi-automatic skeleton binding and animation control.

862 **Study Setup.** We randomly selected 20 sets of decomposition results, each containing visualizations
 863 from different methods applied to the same object. For each set, we generated animated GIFs showing
 864 the part decomposition from multiple viewpoints, allowing participants to better understand spatial
 865 consistency and articulation structure. All visualizations were presented anonymously to avoid bias.
 866 The study was conducted via a Google Form and received responses from 20 participants.

867 **Evaluation Criteria.** Participants were instructed to rate each method based on the following three
 868 criteria:

- 869 • **Clarity of part boundaries** – Are the decomposed part regions cleanly separated with
 870 well-defined borders?
- 871 • **Consistency across views** – Do the decomposed parts remain stable and coherent when
 872 viewed from different angles?
- 873 • **Suitability for rigging/animation** – Are the decomposed parts appropriate for assigning
 874 joints and performing realistic articulated motion?

875 Each criterion was rated on a scale from 1 (Poor) to 5 (Excellent).

876 **Goal.** The objective of this study is to assess the effectiveness of part decomposition methods in
877 real-world rigging scenarios, providing insight into their strengths and limitations for downstream
878 animation applications.