GROUPREF: TOWARDS IDENTITY-FAITHFUL VIDEO GENERATION WITH MULTI-PERSON REFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-person reference-to-video generation requires preserving every specified identity and assigning each action to the intended individual. Existing benchmarks provide limited coverage of larger groups and often allow appearance descriptions in prompts, making it unclear whether models rely on visual references or textual appearance cues to preserve the specified identities. To address these limitations, we introduce GroupRefBench for groups of three to ten people, with prompts that assign actions to individuals by image index and omit appearance descriptions. Using the reference-to-person correspondences established during identity evaluation, the benchmark further assesses whether each individual performs the assigned actions. To support training with indexed instructions, we construct GroupRef100K, which pairs videos with indexed action annotations and diversified reference images. We further propose Cross-Modal Identity Anchoring to strengthen index-to-reference binding by applying shared learnable anchor embeddings to image-index markers and their corresponding reference-image tokens. Controlled attention analysis provides evidence that the model uses these anchor correspondences for reference selection. The resulting GroupRef models demonstrate promising performance on both GroupRefBench and OpenS2V-Eval.

1 INTRODUCTION

Reference-to-video generation synthesizes videos from reference images and natural-language instructions (Liu et al., 2025; Wang et al., 2026). In multi-person scenes, successful generation requires all requested participants to appear with the correct identities and perform their assigned actions. Collaborative activities further require coordinated actions between participants with complementary roles. Recent systems increasingly support explicit image-index markers such as `@Image k` (Kling AI, 2026; BytePlus, 2026; Alibaba Cloud, 2026c; Vidu, n.d.). These markers let users refer to individuals by image index when specifying their actions, without repeating appearance descriptions. However, making this correspondence explicit in the input does not ensure correct identity preservation or action assignment during generation.

Existing benchmarks provide limited coverage of larger groups, as summarized in Table 1. These benchmarks also allow appearance descriptions in their natural-language prompts (Chen et al., 2025a; Yuan et al., 2025; Wei et al., 2025; Pan et al., 2026a; Wu et al., 2026; Zhang et al., 2026b), making it unclear whether models rely on visual references or these descriptions to preserve the specified identities. Moreover, identity fidelity alone does not establish participant coverage or count accuracy. Action evaluation must also determine whether the intended individual performs each action, rather than merely whether the action occurs.

To address these limitations, we introduce **GroupRefBench**, comprising 320 samples balanced across groups of three to ten people, with one reference image per person. Prompts refer to individuals by image index and specify their actions without describing their appearance. We establish evaluation metrics covering identity and composition, visual quality, and instruction following. A shared reference-to-person correspondence links identity, garment, and person-specific action assessment, so that identity and action are evaluated using the same person assignment.

To support training with indexed instructions, we construct **GroupRef100K**, containing 100,000 samples covering individual actions and multi-person collaboration. Our pipeline links identities

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Input and evaluation protocols of reference-conditioned video benchmarks.

Benchmark	Input Protocol			Evaluation Protocol		
	# References	Addressing	Appearance	IDCov	Count	Action/Role
MSRVTT-Personalization	1 – 4	Entity tags	Allowed	No	No	Text similarity
OpenS2V-Eval	1 – 3	Text	Allowed	No	No	Text similarity
UniVBench R2V	1 – 3	Text/indices	Allowed	No	No	Action checklist
IntelligentVBench	1 – 4	Ordinals	Allowed	No	No	Prompt following
FashionVideoBench	Unreported	Indices	Allowed	No	No	Prompt following
MultiRef-Compass	3 – 7	Indices	Allowed	No	No	Binding score
GroupRefBench	**3 – 10**	Indices	Excluded	Yes	Yes	Matched identities

across frames and produces indexed action annotations, while cross-shot reference selection and reference editing diversify the reference images to reduce copy–paste artifacts.

We further propose **Cross-Modal Identity Anchoring (CMIA)** to strengthen index-to-reference binding. CMIA adds a shared learnable anchor embedding to each image-index marker and its corresponding reference-image tokens, reusing the same anchor across repeated mentions. The shared anchor provides a common correspondence cue across the text and image modalities within each input. Controlled attention analysis and anchor perturbations provide evidence that the model uses the shared anchor correspondences for reference selection.

We fine-tune RefAlign (Wang et al., 2026) and MiniMax-H3 (MiniMax, 2026) on GroupRef100K to obtain the **GroupRef** models. Extensive experiments on GroupRefBench and OpenS2V-Eval demonstrate the advantages of GroupRef. We also conduct comprehensive ablation studies to validate the design choices of CMIA.

Our contributions are summarized as follows:

- We introduce **GroupRefBench** for reference-to-video generation with three to ten people. It uses indexed prompts without appearance descriptions and evaluates identity preservation, group composition, and whether matched individuals perform their assigned actions.
- We construct **GroupRef100K** with indexed action annotations and diversified reference images, and propose **Cross-Modal Identity Anchoring** to strengthen index-to-reference binding through shared learnable anchor embeddings.
- Our **GroupRef** models achieve promising results on GroupRefBench and OpenS2V-Eval, especially in preserving facial identity across multi-person scenes.

2 Related Work

2.1 Reference-to-Video Datasets and Benchmarks

Reference-to-video datasets associate reference images with subjects in target videos and their textual descriptions. Existing pipelines establish these correspondences through identity association, synthesis, or retrieval (Yuan et al., 2025; Chen et al., 2025b; Zhang et al., 2026a), with Vera using cross-clip identity retrieval to construct person-specific reference sets (Xu et al., 2026). Beyond reference selection, RefCaptioner grounds caption phrases in reference images (Liu et al., 2026a), ReBind specifies reference sources in editing instructions (Liu et al., 2026b), and GroupVideo provides descriptions of individual actions and interactions (Song et al., 2026). However, maintaining consistent action attribution throughout multi-person event sequences and across reference variations remains challenging. GroupRef100K addresses this challenge by combining constrained person association with indexed action annotation, while preserving the resulting reference–action correspondences during cross-shot pairing and reference editing.

For evaluation, existing benchmarks assess subject fidelity (Chen et al., 2025a; Yuan et al., 2025), instruction following (Wei et al., 2026; Pan et al., 2026a; Wu et al., 2026), and identity, attribute, and action binding (Zhang et al., 2026b). GroupRefBench complements these efforts by evaluating

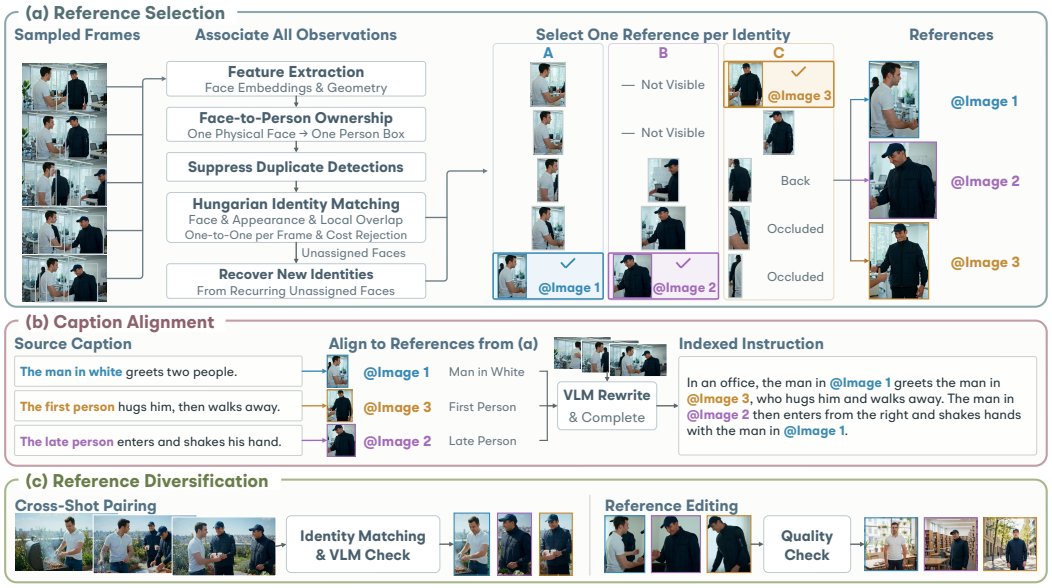

Figure 1: **Construction of GroupRef100K.** Person association establishes identities across frames. Reference selection and augmentation provide visual conditions. Indexed descriptions connect each retained person's actions to the corresponding image.

groups of three to ten people using indexed prompts without appearance descriptions and a shared reference-to-person assignment for identity coverage and person-specific action assessment.

2.2 Multi-Subject Reference-to-Video Generation

At the model level, existing methods explore visual conditioning (Liu et al., 2025; Fei et al., 2025; Jiang et al., 2025), reference-specific semantic fusion (Huang et al., 2025; Hu et al., 2025; Li et al., 2025), and identity-aware alignment or feedback (Wang et al., 2026; Pan et al., 2026b) for multi-subject generation. To distinguish references and connect them to textual mentions, prior work uses positional or category encodings (Zhang et al., 2025; Zhou et al., 2026), shared embeddings among images of the same entity (Cai et al., 2026), or structured attention relating identities and attributes (Xing et al., 2026). For explicit text–reference binding, Movie Weaver combines anchored prompts with learned concept embeddings applied only to reference-image tokens (Liang et al., 2025), while AlcheMinT shares index embeddings between reference tokens and separately encoded word tags (Girish et al., 2026). In contrast, we propose Cross-Modal Identity Anchoring, which adds the same learnable anchor to each reference's visual tokens and every occurrence of its image-index marker in the indexed prompt.

3 GroupRef100K

We construct GroupRef100K to support multi-person reference-to-video generation with explicit correspondence between each person's reference image and indexed action descriptions. As shown in Figure 1, the pipeline first establishes person identities across frames, then uses these identities to connect reference selection with caption annotation. This shared identity assignment is maintained during reference diversification, keeping person indices and associated actions consistent across reference variants.

To limit identity conflicts in multi-person scenes, we associate people across frames with one-to-one matching constraints within each frame. A vision-language model then matches each selected reference image to a person description extracted verbatim from the source caption. Each retained reference and its matched description receive a shared index for subsequent caption annotation. Using this mapping, the model rewrites the caption with person placeholders and is instructed to

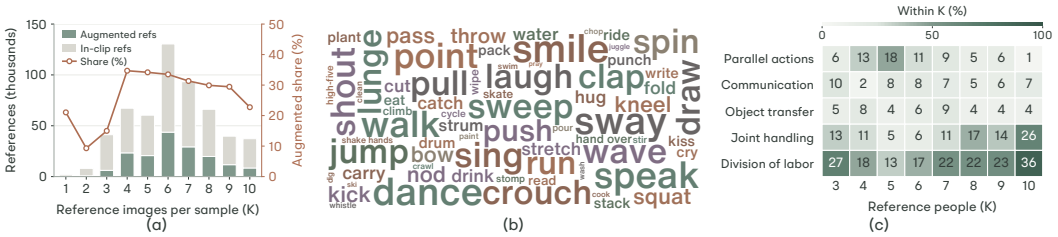

Figure 2: **Data and interaction coverage.** (a) Reference sources and augmentation shares in GroupRef100K by reference count. (b) Action terms in the training descriptions. (c) Interaction categories in GroupRefBench by group size.

preserve actions, interaction partners, spatial relations, and event order. A deterministic renderer then introduces each person as `the <label> in @Image k`, using the assigned reference index. Separating caption rewriting from index insertion allows flexible descriptions of interactions while keeping the rendered indices consistent with the assigned mapping.

To diversify reference conditions and reduce copy-paste artifacts (Zhang et al., 2025), we retrieve alternative views of the same person from other shots in the source video and edit reference backgrounds while retaining the person's appearance. These variants retain the assigned person indices and action annotations, extending reference diversity under the same indexed supervision. The resulting corpus includes in-clip crops and augmented references, with image provenance recorded. Each sample contains $K \in \{1, \ldots, 10\}$ reference images., with most samples containing multiple references. Here, K denotes the conditioning-set size rather than the total number of people visible in the target video. Figure 2 summarizes the reference sources and action vocabulary, and Appendix C provides further construction details and dataset statistics.

4 GroupRefBench

4.1 Benchmark Construction

To assess identity preservation and person-specific action assignment across group sizes, we construct GroupRefBench with 320 samples, including 40 for each group size $K \in \{3, \ldots, 10\}$. Each sample pairs one reference image per person with an indexed prompt. Reference sets contain distinct individuals drawn from a pool of over 1,000 synthetic identities generated from individual descriptions. Prompts refer to participants exclusively through `@Image k` and omit names and descriptions of age, hair, clothing, or occupation, so that reference images specify identities while indexed prompts assign actions. Representative examples are shown in Figure 6 in the appendix.

To cover individual activities and multi-person interactions, we first manually draft action and role assignments spanning parallel actions, communication, object transfer, joint handling, and division of labor. Qwen3.8-Max (Alibaba Cloud, 2026b) then expands these drafts with scene context, objects, spatial relations, and temporal progression while preserving the assigned roles. We subsequently review the resulting samples manually for identity distinctness, compliance with the prompt constraints, and the clarity, plausibility, and coherence of the instructions. Figure 2(c) summarizes the distribution of interaction categories across group sizes, with cell values reporting prompt counts and colors indicating proportions within each group size.

4.2 Evaluation Metrics

We establish an evaluation framework comprising nine metrics across three dimensions: identity and composition, visual quality, and instruction following. Identity, garment, and person-specific action evaluation share a single reference-to-person correspondence, ensuring that these scores refer to the same generated individuals.

Reference-to-person matching. We detect and track people in each generated video, detect faces within person boxes, and encode both generated and reference faces with the same recognition

model. Face embeddings are pooled within each track, and temporally non-overlapping tracks with similar embeddings are merged into *identity groups*. Hungarian assignment (Kuhn, 1955) matches groups one-to-one to references by maximizing total cosine similarity. A match is *accepted* if its similarity is at least 0.10 and the reference is the group's unique best match. Garment and action evaluation reuse this correspondence.

Identity and Composition (IC). (1) Face Similarity (Face) averages the assigned cosine similarity over all requested references before acceptance filtering; unmatched references and negative similarities contribute zero. (2) Identity Coverage (IDCov) measures the fraction of references with accepted matches. (3) Person Count Accuracy (Count) is computed as $(1+\overline{E})^{-1}$, where $\overline{E}$ is the mean relative count error $|N_v - K_v|/K_v$ across videos. Here, N_v counts all generated identity groups, including unmatched groups, and K_v is the requested number of people. (4) Temporal Identity Consistency (TempID) measures the fraction of reference–frame pairs in which the matched identity is confirmed by frame-level face evidence under the same acceptance criteria. Frames without visible faces remain in the denominator. (5) Garment Similarity (Garment) compares each reference with its accepted match using features from head-removed person segments. Scores are averaged over references with evaluable garments, with missing generated evidence contributing zero.

Visual Quality (VQ). (6) Image Quality (IQ) measures frame-level perceptual quality (Ke et al., 2021), and (7) Aesthetic Quality (Aes) assesses visual aesthetics (Schuhmann, 2022). Both are computed on sampled frames independently of identity matching.

Instruction Following (IF). (8) Person-Specific Action Following (Action) evaluates whether each person performs the assigned actions. (9) Other Instruction Following (Other) evaluates non-action requirements, including scene, props, and layout. A vision-language model (VLM) judge (Alibaba Cloud, 2026b) evaluates a fixed per-prompt checklist shared across models, using sixteen chronologically ordered frames with accepted identities labeled. Each item receives a rating from 1 to 5, or `null` when evidence is insufficient. Person-specific items are rated only when the named identity appears in the cited frames. Action assigns equal total weight to each person with action requirements and divides that weight equally among their items; Other weights all items equally. Each metric reports the mean across videos of the weighted rating over numerically judged items.

Overall Score (Total). Total combines the nine component scores normalized to $[0, 1]$, with weights of 0.20 for Face, 0.09 for IDCov, 0.10 each for Count and TempID, 0.05 for Garment, 0.08 each for IQ and Aes, 0.20 for Action, and 0.10 for Other. These weights emphasize identity preservation and person-specific action following. For Action and Other, `null` items retain their checklist weights and contribute zero to the normalized aggregate. Implementation, normalization, and aggregation details are provided in Appendix E.

5 METHOD

Given a prompt c and K reference images $\{I_k\}_{k=1}^{K}$, GroupRef aims to preserve each person's identity and generate the actions assigned to their corresponding `@Image k` marker. Cross-Modal Identity Anchoring (CMIA) encodes this index-to-reference correspondence through shared learnable anchors, providing a common signal across text and visual representations.

Conditioning interface. As shown in Figure 3, text and visual encoders with projections produce prompt features $\{t_i\} \subset \mathbb{R}^d$ and reference features $\{v_{k,p}\} \subset \mathbb{R}^d$, respectively. These features condition a DiT generator together with the noisy video latent z_t and timestep t. We instantiate GroupRef on RefAlign and MiniMax-H3, applying CMIA to their projected conditioning features.

Cross-modal identity anchoring. We maintain a learnable anchor table $E \in \mathbb{R}^{K_{\max} \times d}$, where $K_{\max}$ is the maximum supported reference count. The anchor e_k is added to both the tokens of `@Image k` and the visual tokens of its corresponding reference:

$$\widetilde{t}_i = t_i + \sum_{k=1}^{K} m_{ik} e_k, \qquad \widetilde{v}_{k,p} = v_{k,p} + e_k, \tag{1}$$

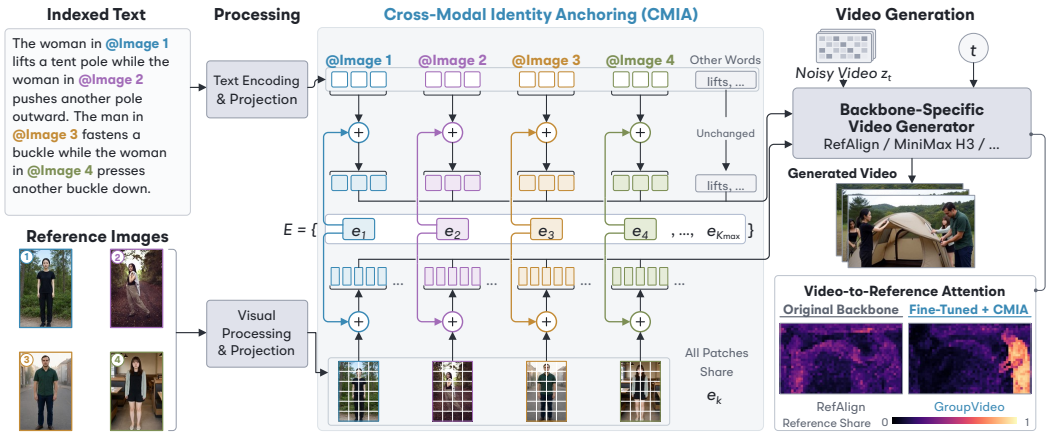

Figure 3: **Overview of GroupRef.** CMIA adds a shared anchor embedding e_k to the text tokens of `@Image k` and the visual tokens of the corresponding reference image, explicitly linking each textual index to its visual reference.

where $m_{ik} = 1$ if token i belongs to an occurrence of `@Image k`, identified by matching its tokenized sequence, and 0 otherwise. All occurrences of the same marker and all patches of its reference share e_k.

Each anchor denotes a reference index whose associated identity varies across samples. Sharing the anchor links the marker in its action context to the corresponding visual reference. The original features retain the instruction context and reference appearance, while the added anchor encodes their correspondence. CMIA introduces $K_{\max}d$ parameters and preserves token dimensions, sequence lengths, and positional encodings.

Training. Each backbone uses an independent anchor table initialized with $e_k \sim \mathcal{N}(0, \sigma^2 I_d)$. During training, references are randomly reordered and prompt indices updated accordingly, exposing each anchor to different identities. We jointly optimize the anchors and adaptation modules, including LoRA adapters, on GroupRef100K using the backbone's fine-tuning objectives, with pretrained backbone weights frozen. At inference, CMIA applies the learned anchors to the corresponding marker and reference features.

6 Experiments

6.1 Evaluation on GroupRefBench

We evaluate a broad range of existing reference-to-video models on GroupRefBench, covering both closed-source and open-source models. The closed-source models include MiniMax-H3 (MiniMax, 2026), HappyHorse-1.1 (Alibaba Cloud, 2026a), Kling3-Omni (Kling AI, 2026), and Wan-3.0 (Alibaba Cloud, 2026c), evaluated through their APIs. The open-source models include LoomVideo (Wu et al., 2026), OmniWeaving (Pan et al., 2026a), BindWeave (Li et al., 2025), Kaleido (Zhang et al., 2025), MAGREF (Deng et al., 2025), Phantom (Liu et al., 2025), VINO (Chen et al., 2026), UniVideo (Wei et al., 2025), VACE (Jiang et al., 2025), RefAlign (Wang et al., 2026), and the open-source version of MiniMax-H3 (MiniMax, 2026). All locally evaluated models, including GroupRef, generate videos for all 320 samples, whereas API constraints prevent complete coverage. Most missing API samples involve larger groups (higher K), so the available subsets favor smaller, generally easier groups and may yield higher aggregate scores than full-benchmark evaluation. Scores computed on these different subsets are therefore not directly comparable. We report the number of available videos as N and aggregate each metric over its available measurements, as detailed in Appendix E.

As shown in Table 2, closed-source models generally achieve higher reported scores than open-source models on their respective evaluation subsets, while multi-person reference-to-video genera-

Table 2: Quantitative results on GroupRefBench.

Method	N	IC					VQ		IF		Total
		Face	IDCov	Count	TempID	Garment	IQ	Aes	Action	Other	
MiniMax-H3	276	0.379	0.903	0.803	0.675	0.523	73.662	5.447	13.442	9.348	66.134
HappyHorse-1.1	280	0.414	0.919	0.716	0.691	0.527	75.246	5.413	12.350	9.381	65.333
Kling3-Omni	175	0.440	0.892	0.746	0.721	0.515	71.037	5.277	9.689	9.128	62.794
Wan-3.0	314	0.410	0.798	0.627	0.370	0.444	71.732	5.337	10.492	8.769	56.842
LoomVideo-5B	320	0.031	0.082	0.750	0.047	0.042	67.278	5.004	0.568	7.263	26.763
OmniWeaving-8.3B	320	0.039	0.111	0.669	0.059	0.056	70.983	4.943	0.755	7.149	26.894
BindWeave-14B	320	0.017	0.029	0.682	0.009	0.011	55.252	4.447	0.094	5.467	21.104
Kaleido-14B	320	0.060	0.210	0.697	0.129	0.116	64.034	5.012	1.057	7.480	29.606
MAGREF-14B	320	0.041	0.123	0.746	0.070	0.065	69.339	5.138	0.398	6.005	26.467
Phantom-14B	320	0.108	0.333	0.720	0.281	0.193	73.529	5.302	1.645	7.516	35.402
VINO-13B	320	0.135	0.419	0.695	0.344	0.231	69.994	5.204	2.028	7.359	37.174
UniVideo-13B	320	0.133	0.419	0.711	0.341	0.239	67.022	5.091	1.544	6.798	35.933
VACE-14B	320	0.034	0.094	0.743	0.052	0.054	74.682	5.232	0.401	7.072	27.392
RefAlign-1.3B	320	0.054	0.180	0.729	0.116	0.108	74.893	5.167	0.534	6.672	29.009
RefAlign-14B	320	0.099	0.314	0.743	0.224	0.187	76.139	5.341	1.423	6.715	33.913
MiniMax-H3-33B	320	0.349	0.735	0.711	0.551	0.399	73.397	5.326	8.742	8.635	55.724
GroupRef-1.3B	320	0.124	0.425	0.768	0.322	0.271	70.533	5.239	1.403	6.243	36.048
GroupRef-14B	320	0.171	0.538	0.761	0.461	0.317	69.438	5.063	1.998	7.193	40.860
GroupRef-33B	320	0.359	0.768	0.692	0.551	0.413	73.148	5.290	8.763	8.697	56.125

tion remains challenging for both groups. On this benchmark, all three fine-tuned GroupRef variants outperform their corresponding backbones in face similarity, identity coverage, garment similarity, and overall score. These results indicate improved preservation of the reference identities.

6.2 Evaluation on OpenS2V-Eval

To assess performance beyond GroupRefBench, we evaluate GroupRef on OpenS2V-Eval (Yuan et al., 2025), which contains 180 samples with one to three reference images per sample. We compare against Saber (Zhou et al., 2025), Kling-1.6 (Kuaishou, n.d.), Vidu-2.0 (Bao et al., 2024), Pika-2.1 (Pika Labs, n.d.), VACE (Jiang et al., 2025), BindWeave (Li et al., 2025), VINO (Chen et al., 2026), Phantom (Liu et al., 2025), MAGREF (Deng et al., 2025), SkyReels-A2 (Fei et al., 2025), and RefAlign (Wang et al., 2026). As shown in Table 3, GroupRef-14B achieves the highest overall score among the compared open models. Both GroupRef variants improve face similarity over their same-scale RefAlign baselines (Wang et al., 2026). GroupRef-1.3B even achieves higher face similarity than several of the compared 14B models.

We also conduct qualitative comparisons on OpenS2V-Eval to examine identity preservation and action following in single- and multi-person scenes. As shown in Figure 4, GroupRef faithfully preserves the reference subjects' identities while generating actions consistent with the text prompts. In the multi-person example, it also maintains both subjects in a coherent shared scene, with clothing consistent with their reference images.

6.3 Ablation Study

To evaluate the effectiveness of CMIA, we conduct ablations with GroupRef-1.3B on the full GroupRefBench. As shown in Table 4, we compare the full model with five variants: (1) w/o CMIA, removing all anchors; (2) w/o image anchors, retaining anchors only on image-index markers; (3) w/o text anchors, retaining anchors only on reference-image tokens; (4) w/ separate anchors, using independent learnable parameters for the two modalities; and (5) w/ fixed hash anchors, replacing learnable anchors with fixed codes derived from SHA-256 hashes of the input reference images. GroupRef-1.3B achieves the highest scores across all five identity and composition metrics and the highest overall score. Both removing CMIA and using alternative configurations with

Table 3: Quantitative results on OpenS2V-Eval.

Method	Overall	Aes	Smooth	Amp	FaceSim	GME	Nexus	Natural
Saber	57.91%	42.42%	96.12%	21.12%	49.89%	67.50%	47.22%	72.55%
Kling-1.6	56.23%	44.59%	86.93%	**41.60%**	40.10%	66.20%	45.89%	**74.59%**
Vidu-2.0	51.95%	41.48%	90.45%	13.52%	35.11%	67.57%	43.37%	65.88%
Pika-2.1	51.88%	46.88%	87.06%	24.71%	30.38%	69.19%	45.40%	63.32%
VACE-1.3B	49.89%	**48.24%**	97.20%	18.83%	20.57%	**71.26%**	37.91%	65.46%
BindWeave	57.61%	45.55%	95.90%	13.91%	53.71%	67.79%	46.84%	66.85%
VINO	57.85%	45.92%	94.73%	12.30%	52.00%	69.69%	42.67%	71.99%
VACE-14B	57.55%	47.21%	94.97%	15.02%	55.09%	67.27%	44.08%	67.04%
Phantom-14B	56.77%	46.39%	96.31%	33.42%	51.46%	70.65%	37.43%	69.35%
MAGREF-480P	52.51%	45.02%	93.17%	21.81%	30.83%	70.47%	43.04%	66.90%
SkyReels-A2-P14B	52.25%	39.41%	87.93%	25.60%	45.95%	64.54%	43.75%	60.32%
RefAlign-1.3B	56.30%	42.96%	94.74%	20.74%	53.06%	66.85%	43.97%	66.25%
RefAlign-14B	60.42%	46.84%	**97.61%**	22.48%	55.23%	68.32%	**48.52%**	73.63%
GroupRef-1.3B	59.59%	43.58%	96.72%	14.98%	73.57%	63.52%	44.55%	63.61%
GroupRef-14B	**60.94%**	45.94%	97.35%	13.42%	**74.05%**	67.34%	44.60%	65.28%



Figure 4: Qualitative comparisons on OpenS2V-Eval. GroupRef preserves the reference subjects' identities more faithfully.

single-modality, separate, or fixed hash anchors yield lower overall scores, supporting the shared learnable anchor design of CMIA.

6.4 Reference-Attention Analysis

To examine reference discrimination and the role of shared anchors, we analyze reference attention in the 1.3B models. Prior work identifies subject-aligned attention in DiTs and shows that strengthening attention to the corresponding references during low-noise denoising can improve

Table 4: Ablation study of GroupRef-1.3B on GroupRefBench.

Method	IC					VQ		IF		Total
	Face	IDCov	Count	TempID	Garment	IQ	Aes	Action	Other	
GroupRef-1.3B	**0.124**	**0.425**	**0.768**	**0.322**	**0.271**	70.533	**5.239**	1.403	6.243	**36.048**
w/o CMIA	0.113	0.387	0.767	0.278	0.242	69.901	5.215	1.351	6.165	34.679
w/o image anchors	0.115	0.395	0.762	0.271	0.248	70.799	5.237	1.251	6.047	34.582
w/o text anchors	0.113	0.400	0.761	0.268	0.247	**70.918**	5.216	1.227	6.295	34.738
w/ separate anchors	0.109	0.392	0.766	0.263	0.241	70.716	5.216	**1.451**	**6.451**	34.929
w/ fixed hash anchors	0.123	0.416	0.760	0.303	0.258	69.958	5.229	1.292	5.933	35.127

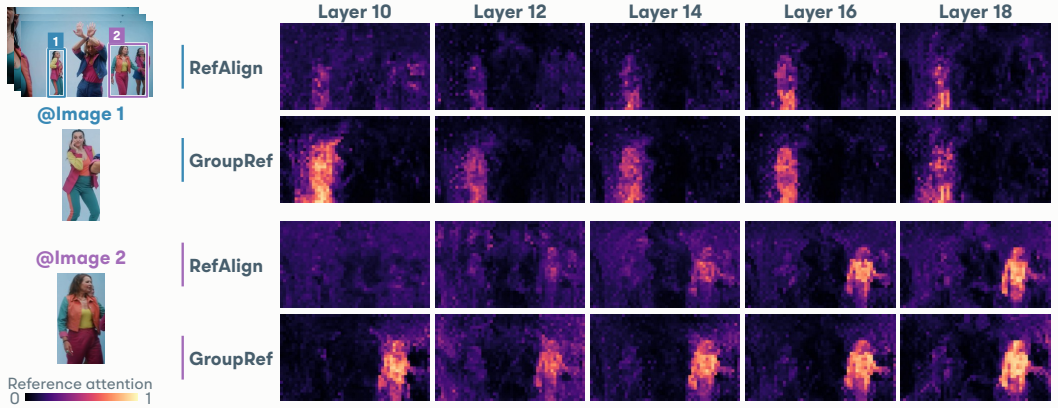

Figure 5: Reference-attention maps for RefAlign and GroupRef on the same example. GroupRef shows more localized, subject-aligned attention.

face similarity (Yu et al., 2026). We observe similar subject selectivity in RefAlign's intermediate blocks 10–18. Therefore, we analyze these blocks using identical noised ground-truth latents, references, and prompts, measuring the attention margin between the correct reference and competing references within annotated subject regions.

GroupRef shows more localized attention in Figure 5 and a larger margin than RefAlign. Most of this increase is already present after fine-tuning without CMIA, with a small further increase in GroupRef. With the GroupRef checkpoint fixed, mismatching image-side anchors reduces the margin, while moving anchors with permuted reference contents partially restores it relative to permuting contents alone. These observations are consistent with the identity gains and support the model's use of anchor correspondence for reference selection under ground-truth conditioning. Detailed results are provided in Appendix G.

7 Conclusion

We presented GroupRef for multi-person reference-to-video generation. GroupRefBench assesses identity preservation, group composition, and person-specific actions in groups of three to ten people, using indexed natural-language prompts without appearance descriptions. We also constructed GroupRef100K to support training with indexed action annotations and diverse reference images, while CMIA strengthens index-to-reference binding through shared learnable anchor embeddings. Experiments on GroupRefBench and OpenS2V-Eval demonstrate promising performance, particularly in preserving reference identities. Furthermore, controlled attention analysis and anchor perturbation experiments provide evidence that the model uses these anchor correspondences for reference selection. We hope that our data construction pipeline and GroupRefBench will support future research on identity preservation and person-specific action assignment in multi-person reference-to-video generation.

AI USE STATEMENT

AI assistance was used to revise manuscript prose, organize references, and check consistency between the manuscript and supplied experiment records. The dataset construction and evaluation also use vision-language models for descriptions, reference filtering, and checklist-based judgments, as specified in the method and appendices.

REFERENCES

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. URL https://arxiv.org/abs/2206.14651.

Alibaba Cloud. HappyHorse 1.1 Released, June 2026a. URL https://www.alibabacloud.com/en/news/product/happyhorse-11-released-mqn. Official product announcement. Accessed September 26, 2026.

Alibaba Cloud. Qwen3.8-max. Alibaba Cloud Model Studio documentation, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max. Official model documentation, updated September 11, 2026.

Alibaba Cloud. Wan3.0 Video Generation, 2026c. URL https://www.alibabacloud.com/help/en/model-studio/wan3-video-generation-guide. Official documentation, updated September 24, 2026. Accessed September 25, 2026.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models. *arXiv preprint arXiv:2405.04233*, 2024.

BytePlus. Dreamina Seedance 2.5 Tutorial, 2026. URL https://docs.byteplus.com/en/docs/ModelArk/2607688. Official documentation, updated September 24, 2026. Accessed September 25, 2026.

Yiyang Cai, Nan Chen, Rongchang Xie, Junwen Pan, Chunyang Jiang, Cheng Chen, Wen Zhou, Zhenbang Sun, Wei Xue, Wenhan Luo, and Yike Guo. HOMIE: Human-object Centric Video Personalization via Multimodal Intelligent Enhancement. *arXiv preprint arXiv:2607.18217*, 2026. URL https://arxiv.org/abs/2607.18217.

Chaofeng Chen and Jiadi Mo. IQA-PyTorch: PyTorch toolbox for image quality assessment. Software, 2022. URL https://github.com/chaofengc/IQA-PyTorch.

Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. VINO: A Unified Visual Generator with Interleaved OmniModal Context. *arXiv preprint arXiv:2601.02358*, 2026.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 6099–6110, June 2025a. URL https://arxiv.org/abs/2501.06187.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yi Zhang, Gen Li, Xinghui Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom-Data: Towards a General Subject-Consistent Video Generation Dataset. *arXiv preprint arXiv:2506.18851*, 2025b. URL https://arxiv.org/abs/2506.18851.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked Guidance for Any-Reference Video Generation with Subject Disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, and Yahui Zhou. SkyReels-A2: Compose Anything in Video Diffusion Transformers. *arXiv preprint arXiv:2504.02436*, 2025. URL https://arxiv.org/abs/2504.02436.

Sharath Girish, Viacheslav Ivanov, Tsai-Shien Chen, Hao Chen, Aliaksandr Siarohin, and Sergey Tulyakov. Alchemint: Fine-grained temporal control for multi-reference consistent video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23227–23237, June 2026.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RhB1AdoFfGE.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. doi: 10.1109/TIP.2020.2967829. URL https://arxiv.org/abs/1910.06180.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Jiangning Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. PolyVivid: Vivid Multi-Subject Video Generation with Cross-Modal Interaction and Enhancement. *arXiv preprint arXiv:2506.07848*, 2025. URL https://arxiv.org/abs/2506.07848.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2004.00288.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. ConceptMaster: Multi-Concept Video Customization on Diffusion Transformer Models Without Test-Time Tuning. *arXiv preprint arXiv:2501.04698*, 2025. URL https://arxiv.org/abs/2501.04698.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, October 2025. URL https://arxiv.org/abs/2503.07598.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11. Software, 2024. URL https://github.com/ultralytics/ultralytics. Official model documentation: https://docs.ultralytics.com/models/yolo11/.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality Adaptive Margin for Face Recognition. *arXiv preprint arXiv:2204.00964*, 2022.

Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. KeyPoint relative position encoding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–255, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kim_KeyPoint_Relative_Position_Encoding_for_Face_Recognition_CVPR_2024_paper.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kling AI. Kling VIDEO 3.0 Omni Model User Guide, February 2026. URL https://kling.ai/quickstart/klingai-video-3-omni-model-user-guide. Published February 6, 2026. Accessed September 25, 2026.

Kuaishou. Kling AI. https://klingai.com/, n.d. Official product website.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109. URL https://doi.org/10.1002/nav.3800020109.

Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. BindWeave: Subject-Consistent Video Generation via Cross-Modal Integration. *arXiv preprint arXiv:2510.00438*, 2025.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Movie Weaver: Tuning-Free Multi-Concept Video Personalization with Anchored Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13146–13156, 2025. URL https://arxiv.org/abs/2502.07802.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, October 2025. URL https://arxiv.org/abs/2502.11079.

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.

Tengfei Liu, Yang Shi, Yuran Wang, Xiaohan Zhang, Yuqing Wen, Yuqi Tang, Qixun Wang, Zhuoran Zhang, Xuanyu Zhu, Weihong Lin, Xinlei Yu, Yujie Wei, Xinwei Long, Fengxiang Wang, Xinlong Chen, Yue Ding, Jialu Chen, Haotian Wang, and Yuanxing Zhang. RefCaptioner: Multi-Reference Image-Grounded Video Captioning. *arXiv preprint arXiv:2607.28509*, 2026a. URL https://arxiv.org/abs/2607.28509.

Xinyu Liu, Shihao Li, Weihong Lin, Xinlong Chen, Yang Shi, Yujin Han, Yiyang Cai, Yanghao Wang, Ruibin Yuan, Yuanxing Zhang, Pengfei Wan, Wenhan Luo, and Yike Guo. ReBind: Multi-Reference Video Editing via Structured Instructions with Explicit Reference Relationships. *arXiv preprint arXiv:2607.14681*, 2026b. URL https://arxiv.org/abs/2607.14681.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

MiniMax. MiniMax H3: An Open Model Breaking the Boundaries Between Tasks and Modalities, July 2026. URL https://www.minimax.io/blog/minimax-h3. Published July 31, 2026. Accessed September 25, 2026.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, Yue Wu, Liefeng Bo, Siliang Tang, and Zhao Zhong. OmniWeaving: Towards Unified Video Generation with Free-form Composition and Reasoning. *arXiv preprint arXiv:2603.24458*, 2026a.

Panwang Pan, Jingjing Zhao, Yuchen Lin, Chenguo Lin, Chenxin Li, Hengyu Liu, Tingting Shen, and Yadong Mu. Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36627–36637, June 2026b. URL https://arxiv.org/abs/2511.00511.

Pika Labs. Pika. https://pika.art/, n.d. Official product website.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Christoph Schuhmann. CLIP+MLP aesthetic score predictor. Software, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Xinyang Song, Libin Wang, Jianxin Sun, Qi Li, Dandan Zheng, JingDong Chen, and Zhenan Sun. GroupVideo: Multi-Identity Customized Text-to-Video Generation. *arXiv preprint arXiv:2607.21027*, 2026. URL https://arxiv.org/abs/2607.21027.

Vidu. Reference to Video. Vidu API documentation, n.d. URL https://platform.vidu.com/docs/reference-to-video. Subjects-Reference to Video section. Accessed September 25, 2026.

Lei Wang, YuXin Song, Ge Wu, Haocheng Feng, Hang Zhou, Jingdong Wang, Yaxing Wang, and Jian Yang. RefAlign: Representation Alignment for Reference-to-Video Generation. *arXiv preprint arXiv:2603.25743*, 2026. URL https://arxiv.org/abs/2603.25743.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. UniVideo: Unified Understanding, Generation, and Editing for Videos. *arXiv preprint arXiv:2510.08377*, 2025. URL https://arxiv.org/abs/2510.08377.

Jianhui Wei, Xiaotian Zhang, Yichen Li, Yuan Wang, Yan Zhang, Ziyi Chen, Zhihang Tang, Wei Xu, and Zuozhu Liu. UniVBench: Towards Unified Evaluation for Video Foundation Models. *arXiv preprint arXiv:2602.21835*, 2026.

Jianzong Wu, Hao Lian, Jiongfan Yang, Dachao Hao, Ye Tian, Yunhai Tong, Jingyuan Zhu, Biaolong Chen, Qiaosong Qi, Aixi Zhang, Wanggui He, Mushui Liu, Jinlong Liu, Pipei Huang, and Hao Jiang. LoomVideo: Unifying multimodal inputs into video generation and editing. *arXiv preprint arXiv:2606.06042*, 2026.

Jiazheng Xing, Fei Du, Hangjie Yuan, Pengwei Liu, Hongbin Xu, Hai Ci, Ruigang Niu, Weihua Chen, Fan Wang, and Yong Liu. LumosX: Relate Any Identities with Their Attributes for Personalized Video Generation. *arXiv preprint arXiv:2603.20192*, 2026. URL https://arxiv.org/abs/2603.20192.

Yulong Xu, Xinyue Liu, Shujuan Li, Huafeng Shi, Yan Zhou, Jiwen Liu, Xintao Wang, Pengfei Wan, Yu-Shen Liu, and Huaibo Huang. Vera: Identity-Faithful Human Subject-to-Video Generation. *arXiv preprint arXiv:2607.20247*, 2026. URL https://arxiv.org/abs/2607.20247.

Niange Yu, Ye Tian, Biaolong Chen, Miao Lu, Aixi Zhang, Hao Jiang, Yunhai Tong, and Pipei Huang. Harnessing intrinsic subject-aware attention for controllable multi-subject video generation. *arXiv preprint arXiv:2609.11507*, 2026.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation. *arXiv preprint arXiv:2505.20292*, 2025. URL https://arxiv.org/abs/2505.20292.

Haojie Zhang, Di Wu, Bingyan Liu, Linjie Zhong, Yuancheng Wei, Xingsong Ye, Nanqing Liu, and Yaling Liang. MuSS: A Large-Scale Dataset and Cinematic Narrative Benchmark for Multi-Shot Subject-to-Video Generation. *arXiv preprint arXiv:2604.23789*, 2026a. URL https://arxiv.org/abs/2604.23789.

Xiaohan Zhang, Yuqing Wen, Junlin Chen, Yuqi Tang, Yiting He, Lizhuo Shao, Weiming Zhu, Tengfei Liu, Yang Shi, Jialu Chen, Yuanxing Zhang, and Huaxiong Li. MultiRef-Compass: Towards Comprehensive Evaluation of Multi-Reference-to-Audio-Video Generation. *arXiv preprint arXiv:2607.14189*, 2026b. URL https://arxiv.org/abs/2607.14189.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-Sourced Multi-Subject Reference Video Generation Model. *arXiv preprint arXiv:2510.18573*, 2025. URL https://arxiv.org/abs/2510.18573.

Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam Woh Ng, Tian Xie, Xiao Han, Yuren Cong, Hang Li, Chuyan Zhu, Aditya Patel, Tao Xiang, and Sen He. Scaling Zero-Shot Reference-to-Video Generation. *arXiv preprint arXiv:2512.06905*, 2025.

Zixiang Zhou, Zhentao Yu, Yifeng Ma, Hongmei Wang, Wenqing Yu, Cong Wang, Zilin Yang, Rui Chen, Jiarong Ou, Yezhou Liu, Yuan Zhou, and Qinglin Lu. Aura: Consistent Multi-Subject Video Generation via VLM-Grounded Semantic Alignment. *arXiv preprint arXiv:2607.04311*, 2026. URL https://arxiv.org/abs/2607.04311.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10492–10502, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhu_WebFace260M_A_Benchmark_Unveiling_the_Power_of_Million-Scale_Deep_Face_CVPR_2021_paper.html.

OVERVIEW OF SUPPLEMENTARY MATERIAL

The supplementary material is organized into seven sections:

- Section A presents supplementary video comparisons between GroupRef and selected baselines on GroupRefBench and OpenS2V-Eval, together with the organization of the video files.
- Section B reports quantitative comparisons on three paired sample subsets with a shared evaluation set for all models in each table.
- Section C details the construction of GroupRef100K, including identity association, reference selection, indexed action annotation, and reference diversification.
- Section D describes the construction of GroupRefBench prompts and illustrates tasks with different group sizes.
- Section E defines the shared identity correspondence, evaluation metrics, score aggregation, and VLM evaluation prompts.
- Section F describes the adaptation modules and shared training settings for the RefAlign-based GroupRef variants.
- Section G describes the reference-attention analysis protocol and reports comparisons of correct-reference attention margins.

A QUALITATIVE VIDEO COMPARISONS

We provide supplementary videos comparing our GroupRef models with selected baselines on GroupRefBench and OpenS2V-Eval. These videos complement the quantitative results and static visualizations by allowing inspection of identity preservation, group composition, and action execution over time.

The videos are organized by benchmark under the folder `2378_supp`. The `GroupRef` subfolder contains comparisons on GroupRefBench, and the `OpenS2V` subfolder contains comparisons on OpenS2V-Eval. Each subfolder contains three comparison videos, named `case1.mp4`, `case2.mp4`, and `case3.mp4`, as shown below.

```
2378_supp/
|-- GroupRef/
|   |-- case1.mp4
|   |-- case2.mp4
|   `-- case3.mp4
`-- OpenS2V/
    |-- case1.mp4
    |-- case2.mp4
    `-- case3.mp4
```

B Quantitative Results on Paired Subsets

We compare models on shared GroupRefBench samples to control for differences in video availability. For each model group, we first intersect the available videos and then retain samples with all nine valid measurements for every included model. All models and metrics within each table therefore use the same sample set. Table 5 summarizes the three subsets.

Table 5: Composition of the paired evaluation subsets. Common counts shared available videos; scored counts videos with all nine valid measurements for every included model.

Subset	Models	Common	Scored
$K \geq 6$, including Kling3-Omni	19	77	75
Unrestricted K, including Kling3-Omni	19	172	170
Excluding Kling3-Omni	18	271	268

The $K \geq 6$ intersection contains samples with $K = 6, 7$. Two videos with incomplete Wan-3.0 measurements are excluded from each subset; the subset excluding Kling3-Omni additionally excludes one video with incomplete UniVideo measurements. Each exclusion applies jointly to all models in that table.

Tables 6, 7, and 8 report the corresponding results. Metrics follow Section E; Action and Other are weighted contributions to Total with maxima of 20 and 10. Across all three subsets, GroupRef improves Face, IDCov, and Garment over each corresponding backbone. GroupRef-1.3B and GroupRef-14B also improve Total on all three subsets.

Table 6: Quantitative results on the paired GroupRefBench subset with $K \geq 6$. All 19 models use the same 75 videos, covering $K = 6, 7$.

Method	N	**IC**					**VQ**		**IF**		**Total**
		Face	**IDCov**	**Count**	**TempID**	**Garment**	**IQ**	**Aes**	**Action**	**Other**	
MiniMax-H3	75	0.366	0.924	0.785	0.663	0.545	73.492	5.466	13.615	9.580	66.286
HappyHorse-1.1	75	0.410	0.925	0.753	0.709	0.536	75.362	5.439	12.272	9.500	65.971
Kling3-Omni	75	0.367	0.815	0.762	0.609	0.465	69.158	5.161	7.791	8.818	56.987
Wan-3.0	75	0.402	0.780	0.671	0.335	0.436	71.448	5.330	9.774	8.659	55.728
LoomVideo-5B	75	0.031	0.071	0.768	0.040	0.039	67.548	5.001	0.630	7.622	27.178
OmniWeaving-8.3B	75	0.041	0.125	0.649	0.067	0.064	71.356	4.987	0.650	7.623	27.411
BindWeave-14B	75	0.018	0.026	0.702	0.005	0.009	58.201	4.465	0.139	5.825	21.892
Kaleido-14B	75	0.047	0.179	0.684	0.091	0.100	63.301	4.987	0.827	7.912	28.595
MAGREF-14B	75	0.036	0.088	0.774	0.046	0.046	69.460	5.110	0.140	6.458	26.185
Phantom-14B	75	0.098	0.325	0.746	0.266	0.189	73.940	5.306	1.838	7.871	35.814
VINO-13B	75	0.124	0.380	0.739	0.306	0.213	69.989	5.167	1.661	7.733	36.546
UniVideo-13B	75	0.113	0.363	0.752	0.290	0.209	66.888	5.077	1.330	6.870	34.606
VACE-14B	75	0.033	0.080	0.777	0.042	0.045	74.799	5.264	0.289	7.485	27.758
RefAlign-1.3B	75	0.053	0.172	0.760	0.110	0.102	75.633	5.162	0.499	7.271	29.767
RefAlign-14B	75	0.095	0.287	0.790	0.203	0.168	76.424	5.335	1.240	7.141	34.021
MiniMax-H3-33B	75	0.344	0.748	0.786	0.545	0.413	73.576	5.337	8.699	9.098	56.926
GroupRef-1.3B	75	0.106	0.381	0.782	0.261	0.240	70.162	5.210	1.261	6.493	34.708
GroupRef-14B	75	0.164	0.514	0.804	0.436	0.307	69.106	5.041	1.894	7.319	40.616
GroupRef-33B	75	0.352	0.783	0.804	0.544	0.427	73.425	5.312	8.746	8.917	57.492

Table 7: Quantitative results on the paired GroupRefBench subset including Kling3-Omni, selected without a restriction on K. All 19 models use the same 170 videos.

Method	N	IC					VQ		IF		Total
		Face	IDCov	Count	TempID	Garment	IQ	Aes	Action	Other	
MiniMax-H3	170	0.414	0.939	0.788	0.743	0.551	73.875	5.489	14.160	9.502	68.754
HappyHorse-1.1	170	0.436	0.949	0.729	0.737	0.548	75.702	5.435	13.366	9.446	67.883
Kling3-Omni	170	0.439	0.891	0.744	0.720	0.515	70.995	5.273	9.675	9.124	62.706
Wan-3.0	170	0.456	0.837	0.607	0.398	0.474	71.513	5.380	11.282	8.929	59.315
LoomVideo-5B	170	0.035	0.098	0.787	0.058	0.052	67.622	5.043	0.689	7.515	27.947
OmniWeaving-8.3B	170	0.046	0.124	0.602	0.072	0.065	71.223	4.982	0.895	7.458	27.138
BindWeave-14B	170	0.020	0.040	0.691	0.013	0.016	55.471	4.455	0.144	5.712	21.734
Kaleido-14B	170	0.078	0.277	0.735	0.183	0.158	67.268	5.146	1.311	7.868	32.704
MAGREF-14B	170	0.049	0.154	0.743	0.098	0.084	69.822	5.135	0.489	6.223	27.587
Phantom-14B	170	0.141	0.426	0.760	0.365	0.249	74.790	5.376	2.257	7.915	39.608
VINO-13B	170	0.164	0.504	0.688	0.420	0.280	70.252	5.185	2.553	7.824	40.436
UniVideo-13B	170	0.165	0.517	0.724	0.431	0.302	66.984	5.110	1.843	7.043	39.350
VACE-14B	170	0.042	0.125	0.776	0.074	0.076	74.984	5.279	0.505	7.280	28.846
RefAlign-1.3B	170	0.066	0.224	0.772	0.153	0.136	74.599	5.217	0.703	6.897	31.013
RefAlign-14B	170	0.126	0.398	0.779	0.296	0.238	76.353	5.388	2.042	7.177	37.677
MiniMax-H3-33B	170	0.406	0.825	0.680	0.631	0.451	73.665	5.370	10.314	9.013	60.433
GroupRef-1.3B	170	0.146	0.500	0.812	0.378	0.321	70.718	5.250	1.679	6.406	38.870
GroupRef-14B	170	0.212	0.662	0.765	0.578	0.396	70.369	5.141	2.658	7.732	45.733
GroupRef-33B	170	0.412	0.860	0.654	0.630	0.466	73.425	5.326	10.126	9.002	60.427

Table 8: Quantitative results on the paired GroupRefBench subset excluding Kling3-Omni. The remaining 18 models use the same 268 videos.

Method	N	IC					VQ		IF		Total
		Face	IDCov	Count	TempID	Garment	IQ	Aes	Action	Other	
MiniMax-H3	268	0.378	0.902	0.802	0.675	0.524	73.670	5.443	13.433	9.328	66.074
HappyHorse-1.1	268	0.414	0.918	0.717	0.693	0.528	75.267	5.413	12.368	9.357	65.357
Wan-3.0	268	0.428	0.812	0.618	0.379	0.455	71.723	5.347	10.639	8.877	57.646
LoomVideo-5B	268	0.033	0.089	0.762	0.052	0.047	67.481	5.011	0.647	7.495	27.387
OmniWeaving-8.3B	268	0.041	0.113	0.650	0.062	0.058	71.182	4.959	0.770	7.436	27.128
BindWeave-14B	268	0.018	0.033	0.687	0.010	0.013	56.109	4.456	0.112	5.612	21.474
Kaleido-14B	268	0.067	0.235	0.712	0.148	0.131	65.376	5.067	1.173	7.730	30.890
MAGREF-14B	268	0.042	0.128	0.749	0.075	0.068	69.512	5.137	0.396	6.184	26.828
Phantom-14B	268	0.119	0.364	0.730	0.309	0.212	74.071	5.330	1.800	7.730	36.822
VINO-13B	268	0.144	0.449	0.701	0.373	0.247	70.288	5.203	2.175	7.594	38.449
UniVideo-13B	268	0.144	0.450	0.714	0.371	0.258	67.030	5.091	1.601	7.071	37.174
VACE-14B	268	0.037	0.101	0.756	0.057	0.059	74.904	5.249	0.464	7.242	27.969
RefAlign-1.3B	268	0.057	0.192	0.745	0.126	0.115	74.842	5.187	0.571	6.962	29.817
RefAlign-14B	268	0.107	0.343	0.760	0.246	0.203	76.299	5.363	1.588	6.903	35.195
MiniMax-H3-33B	268	0.369	0.762	0.700	0.578	0.417	73.518	5.340	9.204	8.883	57.349
GroupRef-1.3B	268	0.132	0.451	0.778	0.339	0.288	70.611	5.243	1.304	6.039	36.480
GroupRef-14B	268	0.182	0.575	0.764	0.496	0.340	69.696	5.078	2.161	7.486	42.403
GroupRef-33B	268	0.378	0.796	0.689	0.580	0.431	73.213	5.302	9.155	8.904	57.729

C Data Construction Details

The construction procedure associates person observations across frames, selects reference images, and uses reference indices in action descriptions.

Identity association and reference selection. In crowded scenes, a person crop can contain a neighboring person's face, and overlapping detections can describe the same person more than once. We detect people with Grounding DINO (Liu et al., 2023), detect and embed faces with InsightFace antelopev2, and associate faces with person boxes jointly within each frame. Table 9 summarizes the detection and association settings. Frames are sampled at a target rate of four per second. The initial IoU tracks retain observations from nearby frames and exclude small, short-lived detections; subsequent identity association operates on their individual detections. The track-gap limit is measured in sampling intervals, and the box-size threshold applies to the median short side along a track.

Face detections are merged when their boxes have IoU at least 0.50, or when IoU is at least 0.30 and face cosine similarity is at least 0.70. A face–person pair is eligible when the person box contains at least 70% of the face box and the face center lies within the upper 55% of the person box. Eligible pairs are scored using containment, relative position, and face size, then assigned one-to-one using Hungarian matching (Kuhn, 1955). Each face is assigned to at most one person box. Duplicate person boxes are suppressed when their IoU exceeds 0.80, the smaller box is more than 85% contained in the larger one, or their IoU exceeds 0.30 and the embeddings of the largest face in each box have cosine similarity at least 0.40. Duplicate suppression precedes identity initialization.

The reference filters use a landmark-based frontality score f and a face-quality score q. For left and right eye positions $\mathbf{p}_\ell, \mathbf{p}_r$, nose coordinate x_n, face detection score d, and face-box short side ℓ_f in pixels, we compute

$$f = 1 - \min\left(1, \frac{|x_n - (x_\ell + x_r)/2|}{d_{\text{eye}}}\right), \qquad q = d \min\left(1, \frac{\ell_f}{160}\right)(0.5 + 0.5f), \tag{2}$$

where $d_{\text{eye}} = \|\mathbf{p}_r - \mathbf{p}_\ell\|_2 + \epsilon$ and $\epsilon = 10^{-6}$ for crop construction. Frontality measures the horizontal displacement of the nose relative to the eyes, and face quality also accounts for confidence and resolution. The frame with the most retained person detections initializes the identities, with ties resolved by their summed face quality. Its detections seed separate identities. Each seed initializes an appearance template and, when available, a face template.

Remaining frames are processed in increasing temporal distance from the initialization frame. For detection i and identity j, face similarity s_{ij}^{face} is the maximum cosine similarity to its retained face templates. CLIP ViT-L/14 crop features (Radford et al., 2021) are normalized, centered by the sample's mean normalized crop feature, and normalized again. Appearance similarity s_{ij}^{app} takes the maximum inner product with the identity's retained appearance features, retaining up to eight most recently assigned features. Spatial similarity s_{ij}^{IoU} uses the nearest assigned box in time and is zero beyond the spatial-support interval. These cues define the association cost

$$c_{ij} = \begin{cases} 1 - \dfrac{s_{ij}^{\text{face}} + 0.35 s_{ij}^{\text{app}} + 0.25 s_{ij}^{\text{IoU}}}{1.60}, & \text{with face evidence,} \\ 1 - 0.85 \dfrac{0.35 s_{ij}^{\text{app}} + 0.25 s_{ij}^{\text{IoU}}}{0.60}, & \text{otherwise.} \end{cases} \tag{3}$$

The first branch requires a face feature for the detection and at least one face template for the identity. Hungarian matching computes a one-to-one assignment at each frame, after which pairs with cost above 0.85 are rejected. Accepted faces update the templates when $q \geq 0.25$, retaining up to eight observations ranked by face quality. Unmatched faces with sufficient quality and low similarity to every initial identity can seed an additional identity. Further unmatched faces satisfying the same quality and initial-identity separation conditions join an additional identity at the recovery cosine threshold, subject to one observation per frame. The minimum-observation requirement in Table 9 applies to these additional identities.

Reference selection searches the observations assigned to each identity, allowing different people to use different frames. Candidates meeting the face-score, face-size, and frontality thresholds in Table 9 form the preferred pool. If this pool is empty, selection uses the available observations.

Table 9: Thresholds for constructing in-clip references. Track gap uses sampling intervals; spatial support uses original video frame indices.

Detection and tracks	
Sampling rate	4 fps
Box / text score	0.30 / 0.25
Track IoU	≥ 0.30
Track gap	≤ 3 intervals
Track observations	≥ 3
Median short side	≥ 80 px
Face assignment	
Face score	≥ 0.50
Face short side	≥ 60 px
Face containment	≥ 0.70
Face-center height	≤ 0.55
Identity association	
Assignment cost	≤ 0.85
Spatial support	≤ 12 frames
Template update q	≥ 0.25
Face templates	≤ 8

Additional identity	
Face quality q	≥ 0.40
Seed face cosine	< 0.35
Recovery cosine	≥ 0.35
Observations	≥ 3
Reference selection	
Face score	≥ 0.65
Face short side	≥ 90 px
Frontality f	≥ 0.35
Crop short side	≥ 224 px
Other containment	≤ 0.30
Duplicate cosine	≥ 0.60
Duplicate boxes	
Box IoU	> 0.80
Box containment	> 0.85
Face-aided IoU	> 0.30
Face cosine	≥ 0.40

Candidates are ranked by $0.5q + 0.3c$, where crop completeness is $c = 1 - b/4$ and b counts image boundaries touched within two pixels. The top twelve are reranked after adding $0.2\,h$, where h is the crop's grayscale Laplacian variance divided by 500 and capped at one. Crowding is measured by $o(C) = \max_{B \in \mathcal{B}_{\text{other}}} |C \cap B|/|B|$ for crop C, excluding this identity's own boxes and suppressed duplicates, with $o(C) = 0$ when no other box is present. We prefer candidates with $o(C) \leq 0.30$ and use the least crowded candidate meeting the crop-size requirement if none meets the containment threshold. Finally, for identities with eligible frontal faces, duplicate checking takes the maximum face cosine over their eligible observations and removes the selected reference with lower face quality when it reaches 0.60.

Action descriptions tied to reference indices. Action annotation uses the reference indices to specify who participates in each described event. A vision-language model first receives the selected references and the source caption, and is asked to return a caption span and a short person label for each reference. Span validation checks substring membership, allowing surrounding quotation marks and whitespace differences. The matched descriptions provide context for distinguishing references during generation. Caption-based rewriting retains references with valid matched spans and renumbers them together with their mapping. A complementary route retains the selected references without requiring a valid span for each. It supplies sampled target-video frames when available and asks the model to describe the actions of people without matched spans from the visual inputs.

Both routes generate descriptions using person placeholders tied to the reference mapping. The prompts request actions, interaction partners, spatial relationships, and event order, interleaving concurrent events and avoiding appearance descriptions unrelated to the actions. Validation requires the expected set of placeholder indices before rendering and checks that each retained image index occurs afterward. A deterministic renderer converts placeholders into indexed person mentions, including `the <label> in @Image k`; pronouns may refer back to earlier mentions. The source descriptions and the VLM's interpretation of the visual inputs provide the semantic associations; automatic validation checks index format and coverage.

Reference diversification. In-clip crops tie the reference view to an observation in the target video. Cross-shot pairing searches other shots of the same source video, matching identities that are mutual nearest neighbors between the target and candidate shots. Matching uses the maximum InsightFace cosine over all face observations satisfying the reference face-eligibility criteria in Table 9. Candidates must meet all initial conditions in Table 10. InsightFace similarity uses antelopev2

Table 10: Cross-shot and reference-editing thresholds. Cross-shot candidates satisfy every initial condition and either stronger face threshold after VLM assessment.

Cross-shot screening	
InsightFace	≥ 0.60
CurricularFace	≥ 0.50
Appearance cosine	≥ 0.78
Shot-index distance	≤ 120
Stronger face match	
InsightFace	≥ 0.65
or	
CurricularFace	≥ 0.55

Editing source	
Face short side	≥ 60 px
Frontality f	≥ 0.35
Other containment	≤ 0.35
Edited output	
Detected faces	1
Frontality f	≥ 0.30
Identity cosine	≥ 0.20

embeddings, CurricularFace (Huang et al., 2020) compares aligned faces in the paired crops, and appearance similarity uses CLIP crop features. For each target identity, candidates are ranked by CurricularFace similarity, with ties resolved by InsightFace similarity, appearance similarity, and smaller shot-index distance, in that order. The top candidate receives a VLM assessment of identity agreement and reference usability. Acceptance requires positive judgments on both criteria and either stronger face threshold in the table; rejection leaves the in-clip reference in place.

Reference editing requests a natural standing pose in a single-person full-body or knee-up portrait, a different background, and retention of the person's face, hair, and clothing. Source eligibility is assessed for each reference, separately from output acceptance (Table 10). For output checking, InsightFace antelopev2 detects faces with scores of at least 0.50. Identity similarity is the dot product of normalized embeddings from the largest detected face in the source and output images. Editing validation uses Eq. 2 with $d_{\text{eye}} = \max(\|\mathbf{p}_r - \mathbf{p}_\ell\|_2, \epsilon)$. The output must contain a single detected face and pass both the frontality and identity-similarity thresholds. An accepted image replaces its source reference; otherwise, the source image is retained. Replacement preserves the reference index, target video, and instruction within this editing branch. Face count, frontality, and identity similarity determine output acceptance; clothing similarity is recorded as an auxiliary measurement.

D GroupRefBench Construction and Examples

We use Qwen-Image-3.0-Pro to render reference images for GroupRefBench. Each sample provides an ordered list of person references, with image indices linking participants to their assigned roles in the instruction.

D.1 Prompt Construction

Construction begins with human-written activity descriptions. These are developed into scene plans specifying the group objective, spatial arrangement, and each participant's visible action or state. Roles are mapped to reference indices before English prompt generation. Qwen3.8-max expands the plans and provides critiques of the resulting text. Human reviewers compare candidate prompts and manually revise selected descriptions. Automatic checks test index coverage, consistency with the assigned cast, duplicate text, and the text encoder's input-length limit.

The plans target short, silent clips, expressing activities through observable movements, poses, and object interactions. Shared activities connect individual contributions to a group objective. Participants may manipulate different parts of one object, pass an object to a specified recipient, or perform complementary actions. Waiting participants can be assigned visible states, such as watching or holding an object ready. This design specifies each person's involvement within the activity.

D.2 Illustrative Benchmark Cases

Fig. 6 illustrates benchmark tasks with three, four, five, eight, and ten references. Each panel includes the indexed references, the instruction, and four frames from a generated video.

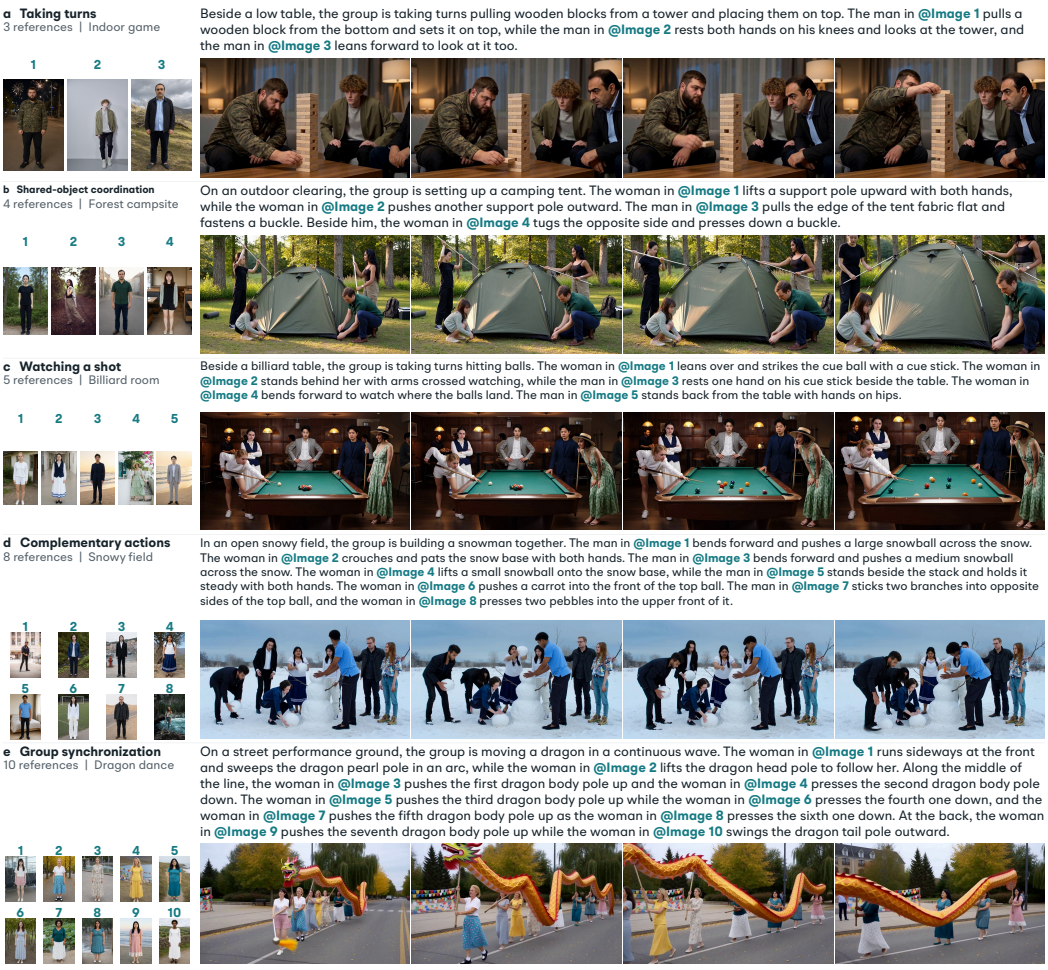

Figure 6: Each GroupRefBench example shows indexed references, a prompt, and four generated frames for an everyday interaction.

The prompt in panel (a) assigns one person to move a wooden block and the others to watch. Panel (b) specifies complementary operations on tent poles and fabric. Panel (c) assigns one participant to strike the cue ball while the others watch or hold specified poses around the billiard table. The snowman prompt in panel (d) distributes rolling, placing, stabilizing, and decorating across the group. Panel (e) assigns distinct roles for the dragon pearl, head, body, and tail during a coordinated dragon dance. Together, the prompts illustrate individual responsibilities and their relationships within a group scene.

E Evaluation Definitions

This section defines the shared identity correspondence and nine metrics in Section 4.2. Table 11 gives their reporting scales and Total weights.

E.1 Identity Extraction and Shared Correspondence

For video v, K_v is the requested identity count and T_v its complete length in frames, indexed from zero. Person tracking covers the full video. Garment, quality, and instruction assessment use the separate sampling rules specified below.

Table 11: Metric names, table reporting scales, and Total weights. Action and Other are reported as their weighted contributions to Total.

Measurement	Metric	Scale	Weight
Face similarity	Face	$[0, 1]$	20%
Identity coverage	IDCov	$[0, 1]$	9%
Count accuracy	Count	$(0, 1]$	10%
Temporal coverage	TempID	$[0, 1]$	10%
Garment similarity	Garment	$[0, 1]$	5%
Image quality	IQ	$[0, 100]$	8%
Aesthetic quality	Aes	$[0, 10]$	8%
Assigned actions	Action	$[0, 20]$	20%
Other requirements	Other	$[0, 10]$	10%

Person tracks and face observations. YOLO11x (Jocher & Qiu, 2024) detects people, and BoT-SORT (Aharon et al., 2022) associates person boxes across frames with appearance re-identification and global motion compensation disabled. We retain tracks observed in at least three frames with median box height at least one tenth of the image height. SCRFD's 10 GF detector (Guo et al., 2022) detects faces and five landmarks in every frame containing retained tracks.

To associate faces with people in a frame, let $\mathcal{F}_a$ be face box a, $\mathcal{P}_b$ person box b, and $\gamma_{ab} = |\mathcal{F}_a \cap \mathcal{P}_b|/|\mathcal{F}_a|$ their containment score. We solve the partial one-to-one assignment

$$\begin{aligned} A^\star = \underset{A \in \{0,1\}^{n_f \times n_p}}{\arg\max} & \sum_{a=1}^{n_f} \sum_{b=1}^{n_p} A_{ab} \gamma_{ab}, \\ \text{subject to} \quad \sum_b A_{ab} \le 1, & \qquad \sum_a A_{ab} \le 1, \\ A_{ab} = 0 \quad & \text{if } \gamma_{ab} < 0.30, \end{aligned} \tag{4}$$

Here n_f and n_p count face and retained person boxes. Hungarian matching (Kuhn, 1955) with zero-score dummy assignments allows unmatched boxes. Frames without an associated face retain their person track but provide no face embedding. Table 12 gives the remaining settings.

Face encoding and track representations. Faces are aligned by a five-point similarity transform to 112×112 crops and encoded with KP-RPE ViT-B (Kim et al., 2024), trained with the AdaFace objective (Kim et al., 2022) on WebFace12M (Zhu et al., 2021). Inputs are RGB intensities in $[-1, 1]$ and aligned landmark coordinates divided by 112; outputs are unit-normalized. Reference embedding r_j uses the highest-confidence SCRFD face and the same preprocessing. An unreadable reference or missing reference face invalidates extraction without reducing K_v.

Let $\mathcal{O}_b$ index all frames on retained track b with an associated face of short-side length $p_{bt} \ge 12$ pixels. Write f_{bt} for its unit embedding and d_{bt} for detector confidence. Generated tracks require $|\mathcal{O}_b| \ge 3$ and $\max_{t \in \mathcal{O}_b} p_{bt} \ge 20$ to receive a face representation; individual references have no temporal eligibility condition. All observations in such a track contribute as

$$\omega_{bt} = d_{bt} \min\{1, p_{bt}/80\}, \qquad h_b = \operatorname{norm}\left(\sum_{t \in \mathcal{O}_b} \omega_{bt} f_{bt} \right), \tag{5}$$

where $\operatorname{norm}(x) = x/\|x\|_2$. Throughout face and garment pooling, a weighted sum with norm at most 10^{-9} is unavailable and is not normalized.

From tracks to identity groups. We consider track pairs in descending order of their original face cosine similarities. A pair is merged if its cosine is at least 0.50 and the current groups have disjoint frame supports, updated after every merge. The original track-level face similarities remain fixed throughout merging. Merging uses face embeddings and temporal support; tracks without face embeddings remain separate groups.

Table 12: Person tracking, face extraction, and identity-association settings.

Component	Setting	Value
Person detector	Input size	1280
	Confidence / NMS IoU	0.25 / 0.50
Tracker	High / low threshold	0.25 / 0.10
	New-track threshold	0.25
	Match threshold	0.80
	Lost-track buffer	30 frames
Track retention	Min. observations	3
	Median height ratio	≥ 0.10
Face detector	Input size	640×640
	Confidence threshold	0.50
Face assignment	Min. containment	0.30
Face alignment	Min. short side	12 pixels
	Aligned image size	112×112
Face pooling	Min. observations	3
	Largest short side	≥ 20 pixels
Group merging	Face cosine	≥ 0.50
Reference matching	Group / frame cosine	0.10 / 0.10

After merging, group q pools its valid track embeddings $\{h_b : b \in \mathcal{B}_q\}$ using effective observation counts:

$$\nu_b = \frac{\left(\sum_{t \in \mathcal{O}_b} \omega_{bt}\right)^2}{\sum_{t \in \mathcal{O}_b} \omega_{bt}^2}, \qquad u_q = \operatorname{norm}\left(\sum_{b \in \mathcal{B}_q} \nu_b h_b\right). \tag{6}$$

The effective count ν_b accounts for unequal observation weights. Only groups with a valid pooled face enter reference matching.

Fixed matching and identity acceptance. Let N_v count all retained groups and $Q_v \leq N_v$ those with valid face embeddings. Suppressing v, define the cosine matrix $S_{qj} = u_q^\top r_j$. Hungarian assignment selects a one-to-one matching of cardinality $\min(Q, K)$ on these raw similarities:

$$\mathcal{M}^\star = \underset{\substack{\mathcal{M} \text{ one-to-one} \\ |\mathcal{M}| = \min(Q,K)}}{\arg\max} \sum_{(q,j) \in \mathcal{M}} S_{qj}. \tag{7}$$

Write q_j for the group assigned to reference j, or $\emptyset$ when unassigned, and set $s_j = S_{q_j j}$ or zero, respectively. An assignment is accepted only when it meets the fixed threshold and its reference is the group's unique best match:

$$a_j = \begin{cases} \mathbf{1}[s_j \geq \tau_g]\, \mathbf{1}[s_j > \max_{\ell \neq j} S_{q_j \ell}], & q_j \neq \emptyset, \\ 0, & q_j = \emptyset, \end{cases} \qquad \tau_g = 0.10. \tag{8}$$

Unassigned references and exact ties are rejected without reassignment. Garment and person-specific VLM judgments reuse this fixed correspondence. The reported identity coverage is the fraction of references accepted by these fixed embedding-similarity criteria.

E.2 Identity and Composition (IC)

Face similarity (Face) and identity coverage (IDCov). The first metric retains continuous similarity, while the second counts accepted reference matches:

$$F_v = \frac{1}{K_v} \sum_{j=1}^{K_v} \max\{0, s_{vj}\}, \qquad C_v = \frac{1}{K_v} \sum_{j=1}^{K_v} a_{vj}. \tag{9}$$

Both average over all K_v requested identities. Face uses assigned cosines before acceptance filtering, truncated at zero.

Count accuracy (Count). Count includes every retained group in N_v, including unmatched groups and those without usable faces. Its error and score over a valid-video set $\mathcal{V}$ are

$$E_v = \frac{|N_v - K_v|}{K_v}, \qquad \mathrm{Count}(\mathcal{V}) = \left(1 + \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} E_v\right)^{-1}. \tag{10}$$

Count converts the mean relative error in group count into a score that penalizes both missing and excess groups.

Temporal identity coverage (TempID). Let f_{vqt} be the recorded unit face embedding for group q at frame t. For an assigned group, the frame-level indicator records whether its face embedding meets the cosine cutoff and uniquely favors the assigned reference:

$$H_{vjt} = \begin{cases} \mathbf{1}[f_{vq_{vj}t}^\top r_{vj} \geq \tau_f]\,\mathbf{1}[f_{vq_{vj}t}^\top r_{vj} > \max_{\ell\neq j} f_{vq_{vj}t}^\top r_{v\ell}], & \substack{q_{vj}\neq\emptyset,\\ f_{vq_{vj}t}\text{ exists},} \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

with $\tau_f = 0.10$. TempID is

$$U_v = \frac{1}{K_v T_v}\sum_{j=1}^{K_v} a_{vj} \sum_{t=0}^{T_v-1} H_{vjt}. \tag{12}$$

TempID measures the fraction of all requested reference–frame pairs supported by accepted face evidence, with $0 \leq U_v \leq C_v$. Every frame remains in the denominator; frames in which a person is absent or their face is unobservable contribute zero for that reference.

Garment similarity (Garment). Clothing features are extracted from the person boxes of the same assigned groups at up to twelve uniformly spaced frames, $\mathcal{T}_v^{\mathrm{cloth}} = \mathrm{unique}\{\mathrm{round}[r(T_v - 1)/11] : r = 0, \ldots, 11\}$. An observation requires a recorded group face, a unique associated person box, and a valid garment region. The frame-level acceptance test in Eq. 11 is not applied. SAM ViT-B (Kirillov et al., 2023) uses the person box to predict a person mask; masks covering less than two percent of the crop are rejected. The crop and mask are letterboxed to 448×448. We exclude the transformed face region expanded by 0.70 face widths on either side, 1.10 face heights above, and 0.60 below.

The distilled DINOv3 ViT-L/16 pretrained on LVD-1689M (Siméoni et al., 2025) encodes the full person crop. The garment mask supplies pooling weights for the resulting patch features. Excluding class and register tokens, let z_p be a patch token and α_p its mask coverage fraction. The garment embedding is

$$\phi = \mathrm{norm}\left(\sum_p \alpha_p z_p\right). \tag{13}$$

Reference crops use the YOLO11x box with greatest face containment among boxes containing the selected face center. The same mask and embedding extraction applies to references and generated observations. Let $b_{vj} = 1$ when reference j has a valid person box, mask, and embedding, and zero otherwise; this reference-only mask is shared across generation models. Let $\mathcal{T}_{vj}^{\mathrm{valid}}$ contain the sampled frames with valid garment features from its accepted group, or be empty if that group or reference garment is unavailable. Then

$$g_{vj} = \begin{cases} \mathrm{median}_{t\in\mathcal{T}_{vj}^{\mathrm{valid}}}\left[(\phi_{vj}^{\mathrm{ref}})^\top \phi_{vjt}^{\mathrm{gen}}\right], & \mathcal{T}_{vj}^{\mathrm{valid}} \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases} \qquad G_v = \frac{\sum_{j=1}^{K_v} b_{vj} a_{vj} g_{vj}}{\sum_{j=1}^{K_v} b_{vj}}. \tag{14}$$

Missing generated garment evidence contributes zero; reference exclusions reduce the garment denominator identically for all models. A video with no evaluable reference garment is unscorable. Fourteen of the benchmark's 2,080 reference instances are excluded. Cosines are not truncated, so G_v can lie in $[-1, 1]$; the nominal reporting range is checked separately when forming Total.

E.3 Visual Quality (VQ)

IQ and Aes use unannotated RGB frames independently of identity matching. The sixteen temporal segment midpoints are

$$t_{vr}^{\mathrm{quality}} = \min\left\{T_v - 1, \mathrm{round}\left[\frac{(r + 1/2)T_v}{16}\right]\right\}, \qquad r = 0, \ldots, 15. \tag{15}$$

Table 13: Inputs to image-quality and instruction evaluation. Their frame-selection formulas differ.

Component	Setting	Value
Quality	Frames	16 midpoints
	Short edge cap	480 pixels
IQ	Pretrained data	KonIQ-10k
	Implementation	PyIQA 0.1.16
Aes	Image preprocessing	CLIP center crop
VLM	Judge	Qwen3.8-max
	Frames	16, with endpoints
	Long edge cap	1024 pixels
	JPEG quality	90
	Temperature	0
	Thinking	Disabled
	Output token limit	6,000
	Reference images	None

Each image is resized with Lanczos interpolation to a short edge of at most 480 pixels, preserving its aspect ratio without upsampling. If a target frame cannot be decoded, neighboring offsets $-1, +1, -2, +2$ are tried in that order. All sixteen requests must yield an image for a valid quality measurement.

IQ uses KonIQ-10k pretrained MUSIQ (Ke et al., 2021; Hosu et al., 2020) through PyIQA 0.1.16 (Chen & Mo, 2022), with RGB inputs in $[0, 1]$. Aes uses the LAION regressor (Schuhmann, 2022) on normalized CLIP ViT-L/14 features (Radford et al., 2021) with center cropping. Per-video scores M_v and A_v are the respective arithmetic means of sixteen frame predictions, rounded to four decimal places. Both metrics assess the visual appearance of sampled frames.

E.4 Instruction Following (IF)

Checklists and identity-bound evidence. Each benchmark prompt has a fixed, text-derived checklist separating person-indexed Action requirements from non-action requirements (Other). The extraction instructions preserve specified objects, partners, direction, timing, outcomes, and manner, assigning scene, prop, layout, and object-state constraints to Other. A checklist sentence is scored as written, including any coupled conditions. All evaluated generators share the same checklist for each prompt.

For instruction assessment, we sample sixteen frames including the video endpoints:

$$t_{vr}^{\text{judge}} = \text{round}\left[\frac{r(T_v - 1)}{15}\right], \qquad r = 0, \ldots, 15. \tag{16}$$

Accepted reference labels from Eq. 8 are propagated along their person tracks and rendered on the corresponding boxes, including tracked frames without a scorable face. This label propagation uses the group-level acceptance criterion; the frame-level threshold τ_f applies to TempID. Qwen3.8-max (Alibaba Cloud, 2026b) receives the temporally ordered, identity-labeled full frames and checklist. Its person-specific judgments use the supplied identity assignment, with reference images excluded from the judge input. The recorded evaluator identifier is the service alias `qwen3.8-max`. Table 13 summarizes the image and inference settings.

Ratings and evidence sufficiency. Each item receives an integer rating $r_i \in \{1, \ldots, 5\}$ or `null`. The rubrics in Section E.6 distinguish visible degrees of fulfillment from insufficient evidence. The latter receives `null`; clear contradictory evidence can support a low numerical rating.

Let $\mathcal{R}_i$ contain the reference identities named by item i, and let $\mathcal{E}_i \subseteq \{0, \ldots, 15\}$ be its cited judge-image indices. Write $\mathcal{L}_{vr}^{\text{judge}}$ for the reference labels attached to boxes in supplied image r, whose source-video frame is t_{vr}^{judge}. A valid numerical judgment must satisfy

$$\mathcal{E}_i \neq \emptyset, \qquad \mathcal{R}_i \subseteq \bigcup_{r \in \mathcal{E}_i} \mathcal{L}_{vr}^{\text{judge}}. \tag{17}$$

Named participants may be established in different cited frames. This condition checks identity availability; fulfillment remains a visual judgment. Generic scene items can have $\mathcal{R}_i = \emptyset$. Items identified as unjudgeable by the binding check are recorded as `null`; the judge assesses evidence sufficiency for the remaining items. Numerical outputs lacking required bindings fail validation. Each accepted response records the rating, observed and missing content, and cited frames.

Both metrics use the joint system prompt. Invalid items are re-evaluated with validation feedback while accepted ratings are retained; rubrics and item weights remain fixed.

Person weighting, coverage, and evidence scores. For video v, let P_v be the number of people with action requirements, L_{vp} the number of actions assigned to person p, and L_v^{other} the number of non-action requirements. Item weights are

$$w_{vp\ell}^{\text{act}} = \frac{1}{P_v L_{vp}}, \qquad w_{v\ell}^{\text{other}} = \frac{1}{L_v^{\text{other}}}. \tag{18}$$

Thus Action gives each person equal planned weight, and Other weights items equally. Each applicable checklist sums to one. An empty checklist is inapplicable, distinct from insufficient evidence for an existing requirement.

For either metric, suppress its superscript and let $\mathcal{J}_v$ be the numerically judged items. Define judgment coverage c_v, conditional rating s_v, and normalized evidence score d_v by

$$c_v = \sum_{i \in \mathcal{J}_v} w_{vi}, \qquad s_v = \frac{\sum_{i \in \mathcal{J}_v} w_{vi} r_{vi}}{c_v} \quad (c_v > 0), \tag{19}$$

$$d_v = \sum_{i \in \mathcal{J}_v} w_{vi} \frac{r_{vi} - 1}{4} = \begin{cases} c_v(s_v - 1)/4, & c_v > 0, \\ 0, & c_v = 0. \end{cases} \tag{20}$$

When $c_v = 0$, s_v is undefined. For $c_v > 0$, the conditional rating divides the weighted sum of judged ratings by their total original weight c_v. The evidence score retains the original item weights, with `null` items contributing zero, so $0 \le d_v \le c_v \le 1$.

For a set $\mathcal{V}^{\text{app}}$ of videos with completed, applicable assessments, let $\mathcal{V}^+ = \{v \in \mathcal{V}^{\text{app}} : c_v > 0\}$. We report

$$\overline{c} = \frac{1}{|\mathcal{V}^{\text{app}}|} \sum_{v \in \mathcal{V}^{\text{app}}} c_v, \qquad \overline{s}_+ = \frac{1}{|\mathcal{V}^+|} \sum_{v \in \mathcal{V}^+} s_v, \qquad D = \frac{1}{|\mathcal{V}^{\text{app}}|} \sum_{v \in \mathcal{V}^{\text{app}}} d_v. \tag{21}$$

For a nonempty completed set with no numerical judgments, $\overline{s}_+$ is undefined and $D = 0$. The Action and Other columns in Tables 2 and 4 report $20D^{\text{act}}$ and $10D^{\text{other}}$, respectively. These are their weighted contributions to Total, with maxima of 20 and 10. The conditional rating $\overline{s}_+$ and judgment coverage $\overline{c}$ describe the fulfillment and availability of judged evidence separately. The aggregate D is computed by averaging per-video evidence scores.

E.5 Overall Score and Evaluation Sets

Fixed-weight aggregation. Each component mean gives equal weight to videos in its specified measurement set, irrespective of group size or checklist length. With the weights in Table 11,

$$\begin{aligned} \text{Total} = 20\overline{F} + 9\overline{C} + \frac{10}{1 + \overline{E}} + 10\overline{U} + 5\overline{G} \\ + 0.08\overline{M} + 0.8\overline{A} + 20D^{\text{act}} + 10D^{\text{other}}. \end{aligned} \tag{22}$$

Instruction terms use the applicable-video sets in Eq. 21. IQ and Aes are normalized by 100 and 10, respectively. The weights in Table 11 are fixed across models. The instruction evidence scores satisfy $0 \le d_v \le 1$ and $0 \le D \le 1$, giving weighted table ranges of $[0, 20]$ and $[0, 10]$. The reporting routine checks per-video inputs and aggregate components against their corresponding reporting ranges, rejecting values outside those ranges by more than a 10^{-12} numerical tolerance without substantive clipping. Accepted totals therefore lie on the zero-to-100 scale. Aggregation first computes each component across its measurement set, including Count from the mean relative count error, and then applies the fixed weights.

Missing evidence and missing measurements. A completed `null` judgment records insufficient evidence and contributes zero to d_v. API or decoding failures, malformed responses, failed output validation, and unavailable videos are recorded as missing measurements, separately from completed numerical or `null` judgments. Reference-side garment exclusions and valid absence of generated evidence follow the metric-specific rules above. Total requires applicable valid measurements for every component and retains the fixed component weights.

Let $\mathcal{V}_{m,k}$ contain the videos with valid measurements of component k for model m, including completed all-`null` assessments. Available-measurement reporting evaluates each component on its own $\mathcal{V}_{m,k}$, allowing different denominators within a row. Paired reporting uses $\mathcal{V}^{\text{paired}} = \bigcap_{m,k} \mathcal{V}_{m,k}$ for every component and model. Comparisons must identify the evaluation set, measurement counts, and group-size composition so that the denominators are explicit.

The main results in Table 2 use available measurements; the ablations in Table 4 use the same 320 videos with all nine measurements.

E.6 Evaluation Prompts

The following verbatim excerpts present the shared evaluation scope, scoring rubrics, and response format in source order. Evaluation uses the complete joint prompt, including the examples and repeated instructions omitted from these excerpts. Within the prompt, `IF` denotes Other; the implementation names `MUSIQ` and `AesCrop` correspond to IQ and Aes.

The user message supplies sixteen annotated images with indices and timestamps, requested items, complete checklists, identity bindings, and validation feedback. Only requested Action and Other items are rated, together in one JSON array; context items provide background. Aggregation uses the returned ratings.

Shared evaluation scope (verbatim excerpt).

> Evaluate two separate metrics using the SAME 16 frames. Apply the Action rubric only to payload.Action.items and the IF rubric only to payload.IF.items. Keep the judgments independent: good scene fulfillment does not establish a person's action, and a failed action does not by itself violate a non-action requirement.

Action rubric (verbatim excerpt).

> You receive 16 chronological full video frames and Action checklist items. R1 represents @Image 1, R2 represents @Image 2, and so on; labels are identities, not screen positions. Rate only the supplied items.
>
> Judge the named person's action, including the stated object, direction, recipient and timing. Another person's action does not fulfill it. Consider partners only when the item requires interaction. Missing or ambiguous bindings mean identity is unresolved, not that the person is absent. Use the whole sequence; a movement needs visible change, while a holding state or pose can be satisfied at a suitable moment. Do not require persistence unless stated.
>
> First decide whether the supplied evidence lets you distinguish fulfillment from nonfulfillment. Use null for unresolved identity, occlusion, details too small, or an uncaptured transition when these prevent that decision. Not seeing a brief event in 16 frames does not by itself prove it never happened. A visible contradiction or clearly assessable unsuccessful attempt can still be scored. Do not turn uncertainty into a middle score.
>
> For judgeable items use an integer from 1 to 5:
>
> 1 = the target and relevant behavior are adequately visible, but the requested action fails or has no meaningful task-specific progress.
>
> 2 = a clear attempt at this exact action, with little achieved.
>
> 3 = substantial partial achievement, with a substantial requested part visibly unfinished.

4 = the main action is achieved, with only a small, concrete visible shortfall.

5 = the stated action is fully supported at the supplied visual resolution.

A score of 2, 3 or 4 needs a visible unfulfilled part of the actual sentence, not a confidence discount.

Action response format (verbatim excerpt).

Return JSON only: {"ratings":[{"id":"A1.1","score":5,"observed":"Concrete visible evidence about the specified person.","missing":"none, the concrete unmet requirement, or what cannot be resolved","evidence_frames":[0,15]}]}.

Return each supplied ID exactly once with these fields. Cite supplied frame IDs only. Numeric scores require evidence for the named bindings, including stated interaction partners. For null explain the unresolved evidence; never describe an unbound person as absent solely because the label is missing. Keep observed and missing consistent with the score. Do not compute an overall score.

Other rubric (verbatim excerpt).

Judge each sentence as written, using ordinary contextual meaning. Person-specific relations require the named bindings; scene and generic props do not. Missing or ambiguous bindings mean identity is unresolved, not that a person is absent. A condition may be fulfilled at a suitable moment unless the sentence explicitly requires persistence. Do not re-evaluate a person's actions or add unstated conditions.

Use an integer score from 1 to 5:

1 = the relevant scene, object or relation is adequately visible, but the requirement is contradicted or essentially unmet.

2 = only a small part of the stated requirement is satisfied; major discrepancies remain.

3 = a meaningful part of the stated requirement is satisfied, but a substantial part remains unmet.

4 = the main requirement is satisfied, with only a small, concrete visible discrepancy.

5 = the stated requirement is fully supported at the supplied visual resolution.

Use null when the evidence is insufficient to decide fulfillment, for example because of unresolved identities, occlusion or indistinguishable small details. Scores 2–4 express visible partial fulfillment, not confidence. For a simple condition with no meaningful partial fulfillment, do not invent a middle grade just to express doubt.

F Implementation and Training Details

GroupRef variants are initialized from the released RefAlign and MiniMax-H3 checkpoints. We apply LoRA (Hu et al., 2022) to the query, key, value, and output projections in every self-attention block, while also training the reference-alignment projections and CMIA anchor embeddings. For the RefAlign-based variants, the flow-matching and representation-alignment objectives jointly optimize the adaptation modules and CMIA anchors. We use AdamW (Loshchilov & Hutter, 2019) with linear learning-rate warmup followed by a constant learning rate. Table 14 summarizes the shared settings for the RefAlign-based variants.

G Reference-Attention Analysis

We compare the 1.3B models on 128 videos from the fine-tuning set, each with three to ten references. All models receive identical noised ground-truth video latents, reference images, and prompts. We extract the query and key vectors used by self-attention in intermediate DiT blocks

Table 14: Shared training settings for RefAlign-based GroupRef variants.

Setting	Value
Training examples	100,000
Training steps	3,200
Effective batch size	32
Learning rate	10^{-4}
Warmup steps	100
LoRA rank / alpha	64 / 64
Anchor initialization	$\mathcal{N}(0, 0.01^2)$

10 through 18 (zero-based), after the learned projections, RMS normalization, and rotary positional encoding.

At each layer and noise condition, let q_{hu} denote video query u in attention head h, and k_{hv} the key at token position v. The attention mass assigned to reference image j is

$$a_{uj} = \frac{1}{H} \sum_{h=1}^{H} \sum_{v \in \mathcal{R}_j} \frac{\exp(q_{hu}^\top k_{hv} / \sqrt{d_h})}{\sum_{w \in \mathcal{S}} \exp(q_{hu}^\top k_{hw} / \sqrt{d_h})}, \tag{23}$$

where H is the number of heads, d_h is the head dimension, $\mathcal{R}_j$ contains the tokens of reference j, and $\mathcal{S}$ contains all video and reference tokens. Thus, we first normalize over the full key sequence, sum the probabilities within each reference image, and average over heads.

We map the annotated person masks to the video-token grid and sample foreground queries with valid annotations and low overlap with other people. For a person associated with reference i, let Ω_i be its nonempty set of valid sampled foreground queries and $w_{iu} > 0$ the corresponding weights, proportional to person-mask occupancy. The query-level correct-reference margin and its person-level average are

$$g_i(u) = a_{ui} - \frac{1}{K-1} \sum_{j \neq i} a_{uj}, \qquad G_i = \frac{\sum_{u \in \Omega_i} w_{iu} g_i(u)}{\sum_{u \in \Omega_i} w_{iu}}, \tag{24}$$

where K is the number of supplied references. Query positions and weights are shared across models, layers, and noise conditions. We average G_i equally over people with valid foreground support, then over the selected layers, noise realizations, and noise levels within each video, and finally over videos. All supplied references enter the softmax; only references $j \neq i$ enter person i's competing-reference mean.

Over this layer range, the mean margins are 0.97, 3.39, and 3.42 percentage points for the original RefAlign, reference-alignment fine-tuning without CMIA, and GroupRef, respectively.