ReferEverything: Towards Segmenting Everything We Can Speak of in Videos
Supplementary Material

Anurag Bagchi[1] Zhipeng Bao[1] Yu-Xiong Wang[2] Pavel Tokmakov[3†] Martial Hebert[1†]
[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign [3]Toyota Research Institute

https://refereverything.github.io/

In this supplementary material, we first include additional details for our Ref-VPS dataset in Section A. Next, we offer a deeper quantitative analysis, covering VAE-based mask reconstruction, failure modes, and computational costs, *etc*., in Section B. Section C presents additional qualitative evaluations, including visualizations on typical failure cases, challenging fight scenes, and ambiguous or overlapping scenarios. Finally, in Section D, we report all the implementation details.

A. Ref-VPS Dataset Details

A.1. Dataset collection pipeline and statistics

During our dataset collection, we first established a non-exhaustive taxonomy of five broad and possibly overlapping concepts. This taxonomy was designed to encompass key modes of dynamic change while offering a structured framework for the task. The concepts and their definitions are as follows:

- **Temporal Object Changes:** Phenomena where an object's state or shape evolves over time (*e.g*., object deformation, melting)
- **Motion Patterns:** Motion in amorphous or non-rigid regions (*e.g*., water ripples, flickering flames)
- **Dynamic Environmental Changes:** Environmental transformations affecting spatial regions over time (*e.g*., clouds moving across the sky, waves rising)
- **Interaction Sequences:** Events characterized by interactions between objects (*e.g*., bullet hitting glass, object collisions)
- **Pattern Evolution:** Progressive changes in patterns or textures (*e.g*., changing patterns of smoke dispersion, fluctuating light levels)

Our final dataset comprises 145 video clips representing 39 distinct dynamic process concepts. We report a comprehensive list of key statistics in Table A. Most of our samples are between 2.5 and 5 seconds in length, but can go up to more than 20 seconds. The distribution of our sample lengths is reported in Figure A.

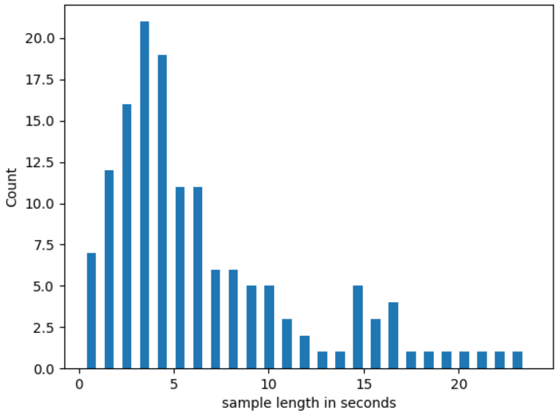

Figure A. Distribution of sample lengths in Ref-VPS. Most of our samples are between 2.5 to 5 seconds in length, but can go up to more than 20 seconds.

Clips	145
FPS	24
Frames	23,442
Concepts	39
Avg length (s)	6.74
Annotation FPS	6
Min-resolution	712 × 576
Max-resolution	1024 × 576

Table A. Statistics of our Ref-VPS benchmark. Our dataset contains 145 video clips covering 39 concepts for dynamic processes.

A.2. Annotation visualizations

Figure B showcases examples of our Ref-VPS segmentation mask annotations. Our annotations capture the full extent of target objects, as seen with the icicle (second row) and the glass (fourth row). For more ambiguous cases, such as glowing water (first row) or a dandelion being blown (third row), only the confident regions are labeled, while uncertain areas are marked as Ignore (yellow). These *Ignore Regions*

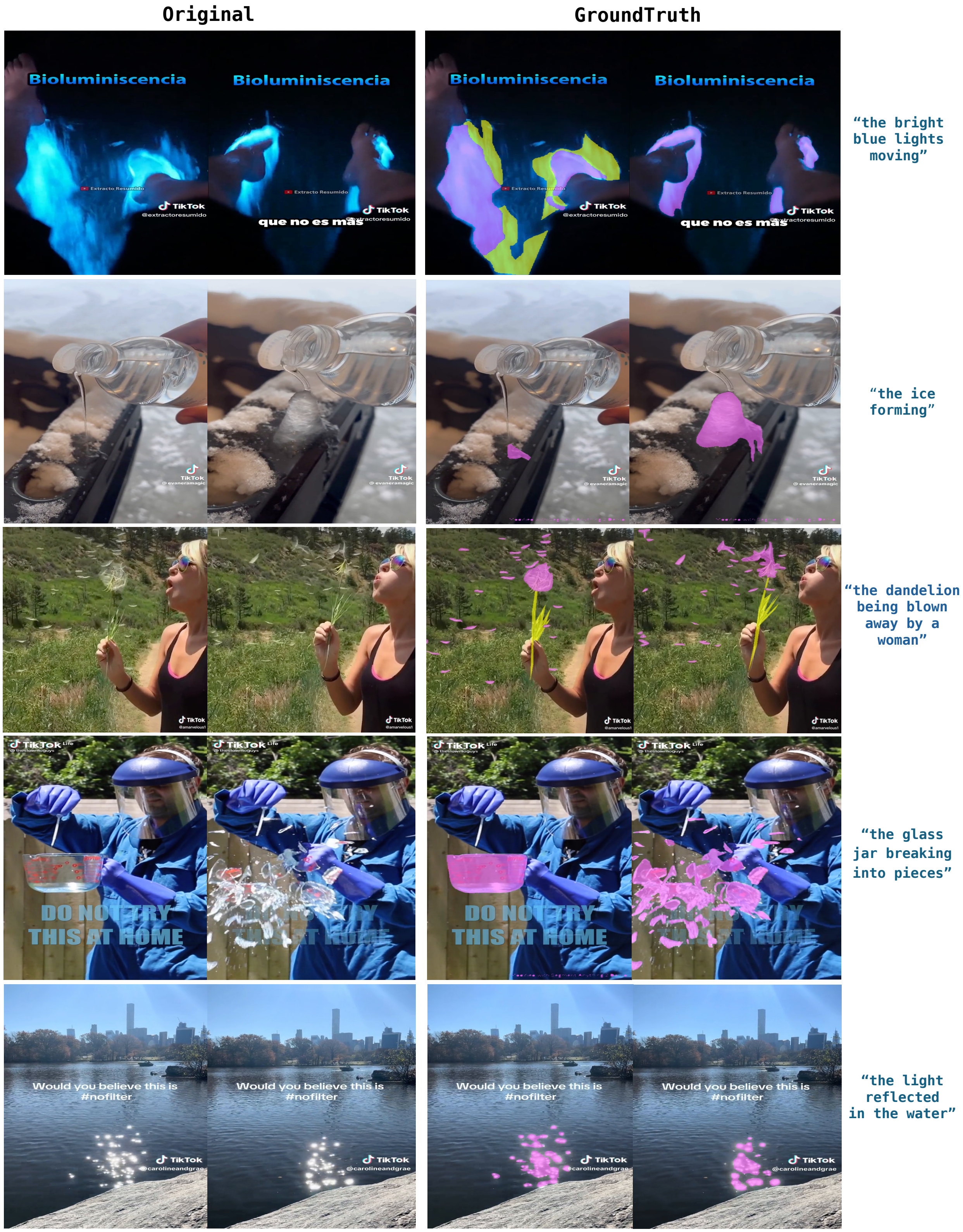
Original
GroundTruth
Bioluminiscencia
que no es más
"the bright blue lights moving"
"the ice forming"
"the dandelion being blown away by a woman"
DO NOT TRY THIS AT HOME
"the glass jar breaking into pieces"
Would you believe this is #nofilter
"the light reflected in the water"
TikTok

Figure B. Samples from our Ref-VPS dataset. Ground-truth masks are shown in pink, and the Ignore regions are shown in yellow. Pixels inside the Ignore regions are not included in the metric calculation.

are excluded from metric computation, ensuring that evaluations focus on reliable mask regions and are not penalized for inherently ambiguous boundaries.

A.3. Mask annotation accuracy evaluation

To assess the quality of our annotations, we compute inter-annotator agreement on the Ref-VPS dataset. Specifically, an independent annotator relabeled a subset of the dataset, covering all 39 dynamic concepts, using the same annotation protocol. Following the evaluation approach of Benenson et al. [2], we report an inter-annotator mean IoU (mIoU) of **87.1%**, significantly higher than the ∼80% agreement number reported for COCO [4]. This high agreement demonstrates the effectiveness of our annotation protocol, particularly the use of Ignore labels to handle ambiguity in subjective scenarios.

B. Additional Quantitative Evaluations

In this section, we provide additional quantitative evaluations for our proposed REM. Same as our ablation study in the main paper, we conduct these experiments using the MS-1.4B version, unless staged otherwise.

B.1. Mask reconstruction accuracy analysis

In designing our REM model, we repurpose a pre-trained VAE as the mask decoder, based on the intuition that large-scale pre-training enables the VAE to effectively reconstruct masks as images. To validate this assumption, we quantitatively evaluate the VAE's reconstruction performance on binary mask images, following the methodology of Marigold [11]. Specifically, we assess reconstruction accuracy on 3,471 binary masks from the Ref-YTB training set (one per video). The VAE achieves a mean absolute error (MAE) of **0.0144** for mask reconstruction. In comparison, reconstructing the corresponding RGB frames yields a higher MAE of *0.1236*, reflecting the greater difficulty of the RGB task. Furthermore, the VAE attains a mask reconstruction mIoU of **99.33%** between the input and output masks. These results support our approach of converting masks into 3-channel inputs for compatibility with pre-trained auto-encoders, effectively mitigating concerns about domain mismatch.

B.2. Failure mode analysis

We conducted a quantitative evaluation of our REM failure modes on Ref-VPS in Figure C. We measure performance degradation across motion and shape changes (following Dave et al. [7]), and prompt complexity (sentence length). Our analysis reveals that significant shape change is the primary failure mode, with motion and prompt complexity having secondary impacts. These results further illustrate the challenge of segmenting dynamic concepts in videos.

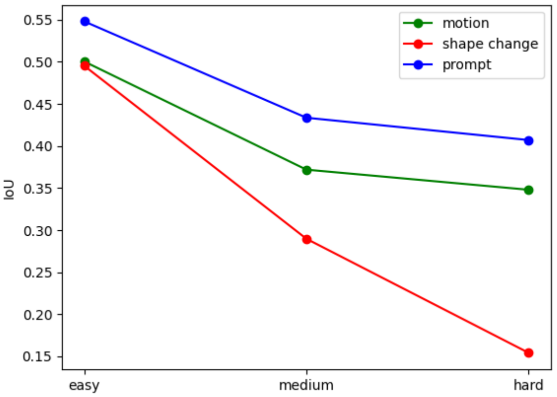

Figure C. Quantitative evaluation of our REM (MS-1.4B) failure modes on Ref-VPS: Significant shape change is the primary failure mode, with motion and prompt complexity having secondary impacts.

Method	Memory (GB)	Speed (FPS)
MUTR	34.1	13.6
UNINEXT	9.7	3.3
VD-IT	72.8	7.1
REM (MS-1.4B)	41.8	7.1

Table B. Inference costs of REM and top RVS methods on Ref-DAVIS. Both the memory requirements and the runtime of REM are on par with other models in the literature.

B.3. Computational cost

We report the inference speed and memory consumption of REM (MS-1.4B) alongside key baselines in Table B, using their official public implementations. All measurements are conducted on 32-frame clips from Ref-DAVIS using a single NVIDIA A100 GPU, with averages computed over 80 runs. As shown, the inference costs of REM align with those of other state-of-the-art approaches.

For training, REM requires 174 hours on four A100 GPUs. Since most baselines do not disclose training costs, we estimate them under equivalent hardware and conditions (excluding I/O time), and report the results in Table C, including per-GPU memory consumption. Our training efficiency is on par with prior works. Notably, UNINEXT, the current state-of-the-art in RVOS, requires approximately 6.3 times longer to train than REM due to its reliance on over ten supervised datasets to achieve strong object segmentation performance. In contrast, REM leverages Internet-scale pre-training to attain comparable performance on in-domain benchmarks and significantly outperforms UNINEXT in out-of-distribution scenarios, all while incurring a fraction of the training cost.

Method	Memory (GB)	Total Runtime (hr)
MUTR	30.4	134
UNINEXT	30.2	1906
VD-IT	68.5	260
REM (MS-1.4B)	61.8	174

Table C. Training costs of REM and top RVS methods. Our costs are on par with prior work and are notably significantly lower compared to UNINEXT, the state-of-the-art RVOS approach.

Training stages	Training strategy	Ref-YTB	Ref-VPS
		($\mathcal{J}\&\mathcal{F}$)	$\mathcal{J}$
Two-stages (Ours)	Images → Images & Videos	**68.4**	**49.0**
Single-stage	Images & Videos	66.3	46.33

Table D. Comparison between our default two-stage training and single-stage training strategy. The two-stage training strategy allows the model to first learn strong spatial representations from image-only data before incorporating the more complex temporal dynamics present in videos. Therefore, it achieves better results for both the standard RVS benchmark and our out-of-distribution Ref-VPS dataset.

B.4. Ablation of the training strategy

Our model adopts a two-stage training strategy (detailed in Section D), where we first pre-train on image-only data to learn spatial representations, followed by joint fine-tuning on mixed image and video data. In this section, we compare the two-stage approach to a single-stage alternative and analyze the impact of varying the amount of image data used in the second stage.

Benefits of two-stage training. As reported in Table D, the two-stage training strategy yields superior performance on both the standard RVS benchmark and the out-of-distribution Ref-VPS dataset. This improvement stems from allowing the model to first acquire strong spatial priors from image-only data before incorporating the more complex temporal dynamics of videos. Additionally, initializing with well-trained spatial weights enhances training stability and convergence.

Impact of image data used in the second stage. In our default two-stage setup, we use an equal amount of image and video data during the second stage. To investigate the effect of image data volume, we consider two variants: one with twice as many images and one with no image data. As shown in Table E, using no image data significantly degrades generalization on Ref-VPS, underscoring the importance of image supervision. Conversely, doubling the image data leads to performance degradation on both benchmarks, suggesting that excessive reliance on static visual information can hinder the learning of spatiotemporal dynamics. These results highlight the importance of a balanced integration of image and video data for effective training.

Images:Videos	Ref-YTB	Ref-VPS
	($\mathcal{J}\&\mathcal{F}$)	$\mathcal{J}$
2:1	67.0	48.0
1:1 (Ours)	68.4	**49.0**
No images	**68.7**	40.7

Table E. Impact of the image data volume used in the second stage. Incorporating image data in the second stage significantly enhances the model's generalization on the Ref-VPS dataset compared to the version trained without images, while using more image data yields suboptimal results. Overall, a balanced integration of image and video data is key to the success of our approach.

Benchmark	SD2.1	VC-1	VC-2	MS-1.4B
Ref-YTB ($\mathcal{J}\&\mathcal{F}$)	60.2	57.5	64.9	**63.5**
Ref-VPS ($\mathcal{J}$)	29.8	28.0	36.8	**40.0**

Table F. Analysis of the effects of generative pre-training on Ref-YTB and Ref-VPS. Both large-scale image pre-training as well as learning to model video-language interactions are important for robust RVS performance.

B.5. Effect of generative pre-training on RVS

We focus on how generative pre-training affects the RVS performance in this section. We focus on comparing our MS-1.4B model variants with other pre-trained diffusion models of a similar parameter size.

We begin by evaluating the effect of Image generation pre-training in Table F. As a baseline, we first fine-tune Stable Diffusion 2.1 [3] (an image generation model, denoted as SD2.1) on individual frames (column 1 in the table). This variant has no temporal modeling capacity, but neither does UNINEXT [23] - the state-of-the-art approach for RVOS. However, it strongly underperforms compared to our best video-based variants, not only on Ref-VPS but also on the object-centric Ref-YTB. This shows that while generative pre-training relies heavily on images, video data is crucial for learning effective representations for tracking.

Next, we evaluate two variants of the VideoCrafter model [5, 6] (denoted as VC-1 and VC-2 in Table F), both initialized from Stable Diffusion 2.1 [3] and trained on 600M images and 10-20M videos. VC-2 focuses on high-quality data curation, which has been shown to be important for representation learning in the past [9, 18], and leads to substantial performance gains across both benchmarks. Finally, the ModelScope [21] approach is also initialized from Stable Diffusion, but trained on the larger LAION 2B and a similar amount of high-quality video data (last column in Table F). It performs comparably to VC2 on Ref-YTB, while demonstrating the best zero-shot generalization to Ref-VPS among all the variants, making it our default representation. These results highlight that large-scale image pre-training, combined with generative video-language modeling, is important for generalization in RVS.

Noise Level	Ref-YTB ($\mathcal{J}$&$\mathcal{F}$)	Ref-YTB ($\mathcal{J}$)
200	59.2	36.2
50	**62.9**	35.0
0 (Ours)	**62.9**	**40.5**

Table G. Ablation study on the choice of the noisy timestep. The best performance is achieved with minimal noise ($t = 0$), validating our design.

Method	Ref-VPS		Ref-DAVIS	
	$\mathcal{J}$	Temp. Con.	$\mathcal{J}$	Temp. Con.
MUTR	24.1	2.9	64.8	3.4
UNINEXT	26.3	5.2	68.2	5.2
VD-IT	35.3	4.7	66.2	3.1
REM (MS-1.4B)	49.0	2.8	69.9	2.1

Table H. Temporal Consistency comparison to the state of the art on Ref-VPS and Ref-DAVIS. Our approach demonstrates the best temporal consistency on both object-centric and non-object-centric datasets.

B.6. Ablation study on the noisy timestep

In the main paper, we default to a noise timestep of $t = 0$, based on the observation that our task formulation focuses on direct mask latent prediction rather than denoising. This design choice eliminates the need for injecting noise into the latent space. To empirically validate this decision, we conduct a single-stage training experiment on the Ref-YTB dataset and report the results in Table G. The findings confirm our hypothesis: minimal noise (*i.e.*, $t = 0$) consistently leads to the best performance, reinforcing the suitability of this choice for our predictive framework.

B.7. Temporal consistency evaluation

Evaluating temporal consistency in video segmentation remains a challenging task, as it is difficult to disentangle variations caused by model inconsistency from those arising due to genuine object deformations. For example, the temporal consistency metric initially introduced in the DAVIS dataset [17] was applied only to videos with minimal object deformation and occlusion, and was eventually deprecated by the dataset authors due to its limited applicability.

To address these challenges, we adopt a simple yet effective temporal consistency metric that quantifies frame-to-frame stability. Specifically, we compute the average difference in Intersection-over-Union (IoU) between predicted masks and ground truth masks across consecutive frames. Formally, the metric is defined as:

$$\text{Temp. Con.} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{T_n}\sum_{t=1}^{T_n}(IoU_{\text{diff}})\right], \tag{A}$$

where N is the number of samples and T_n is the number of frames in the n^{th} sample, and

$$IoU_{\text{diff}} = IoU(Pred_{t+1}, GT_{t+1}) - IoU(Pred_t, GT_t). \tag{B}$$

Lower values indicate better temporal consistency. However, it is important to interpret this metric in conjunction with prediction accuracy, as trivially empty predictions would yield a perfect consistency score of zero without meaningful segmentation.

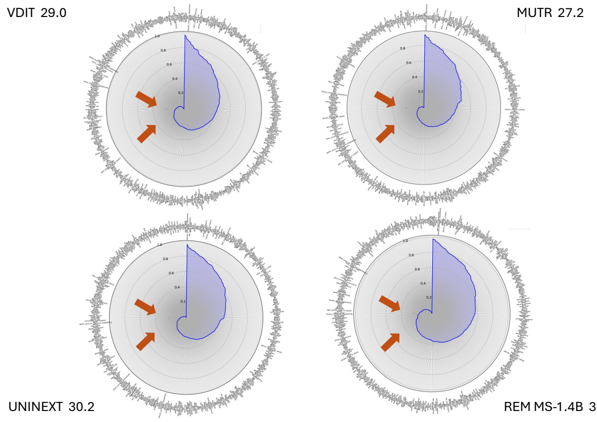

Figure D. Class-wise $\mathcal{J}$ scores (mIoU) across 454 object classes demonstrating concept coverage on BURST. As indicated by the arrows, REM is more robust on the most challenging categories compared to other methods.

We report both region similarity and temporal consistency on Ref-VPS and Ref-DAVIS (both sampled at 24 fps) in Table H. REM achieves the best temporal consistency on both object-centric and non-object-centric datasets. Interestingly, while MUTR also attains a strong consistency score on Ref-VPS, this is primarily due to its frequent output of empty masks, as reflected in its low region similarity. Conversely, UNINEXT, despite being the state-of-the-art on traditional RVOS benchmarks, shows the poorest temporal stability across both datasets.

B.8. Concept coverage plot on BURST dataset

Figure D presents the concept coverage plots on the BURST [1] dataset for VD-IT [27], MUTR [25], UNINEXT [23], and our method. As shown, REM consistently outperforms the baselines on the most challenging categories, further highlighting its strong generalization ability across a diverse range of visual concepts.

B.9. Additional evaluation on MeViS

Data efficiency. We first provide the data source for all the baseline models we compared in Table I. Our REM (Wan-14B) reaches a new state of the art on MeViS while relying on *orders of magnitude fewer* pixel-level annotations than prior work. In particular, the strongest baseline,

Method	Mask annotation	MeViS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Referformer [22]	MeViS	31.0	29.8	32.2
VISA-13B [24]	Ref-COCO/+/g, Ref-YTB, MeViS, Ref-DAVIS, ReVOS, LVVIS, Refclef, ADE20k	44.5	41.8	47.1
DsHmp [10]	MeViS	46.4	43.0	49.8
GLUS [13]	Ref-YTB, MeViS, Ref-DAVIS, ReVOS, LVVIS	51.3	48.5	54.2
REM (Wan-14B)	MeViS	57.6	54.3	60.9
REM (Wan-14B)	Ref-COCO/+/g, Ref-YTB, MeViS	**60.3**	**57.2**	**63.4**

Table I. Comparison to the state of the art on the MeViS benchmark with a comprehensive list of mask annotations used in training. Our REM (Wan-14B) reaches a new state of the art on MeViS while relying on orders of magnitude fewer pixel-level annotations than prior work.

GLUS, couples a large-scale multimodal LLM with SAM2 masks and aggregates supervision from at least six video- and image-level datasets. In contrast, the full version of REM is fine-tuned on just three datasets, yet it improves the previous best $\mathcal{J}\&\mathcal{F}$ from 51.3% to 60.3%. These results underline that preserving the generative architecture of a diffusion model transfers rich visual–language priors so effectively that only a modest amount of task-specific data is needed to surpass far heavier-supervised baselines.

Single-dataset training. To isolate the contribution of our training protocol, we additionally built a variant of our Wan model by finetuning on the MeViS training set alone. Performance decreases by only 2.7 in terms of $\mathcal{J}\&\mathcal{F}$, yet it still outperforms the strongest published baseline by 6.3 points. This resilience demonstrates that our mixed training strategy endows the model with robust spatial–temporal representations that generalize even when the downstream supervision is extremely limited. We include the training details for our Wan variant in Section D.

C. Additional Qualitative Evaluations

C.1. Failure cases visualizations

A few representative failure cases of REM (MS-1.4B) on Ref-VPS are shown in Figure E. Our method suffers from object-centric bias in the most challenging scenarios (*e.g.*, light reflection and veins) and struggles with extremely fast motion (*e.g.*, the lightning strike).

C.2. Evaluation on challenging fight scenes

Fight sequences in movies and animated shows present a particularly challenging setting for referring video segmentation. These scenes are often characterized by severe and frequent occlusions, objects or characters exiting the frame, and rapid camera pose changes. Such factors cause drastic variations in appearance, demanding high temporal and semantic consistency to accurately track, re-identify, and segment the referred entities.

Our REM excels in this domain of extremely challenging samples as illustrated in Figure F. In contrast, both UNINEXT and VD-IT exhibit clear failure cases when the referred entity undergoes large occlusions or momentarily disappears from view. Notably, despite utilizing a video diffusion backbone, VD-IT fails to fully exploit the temporal consistency learned during video diffusion pre-training, whereas REM maintains robust performance under these challenging conditions.

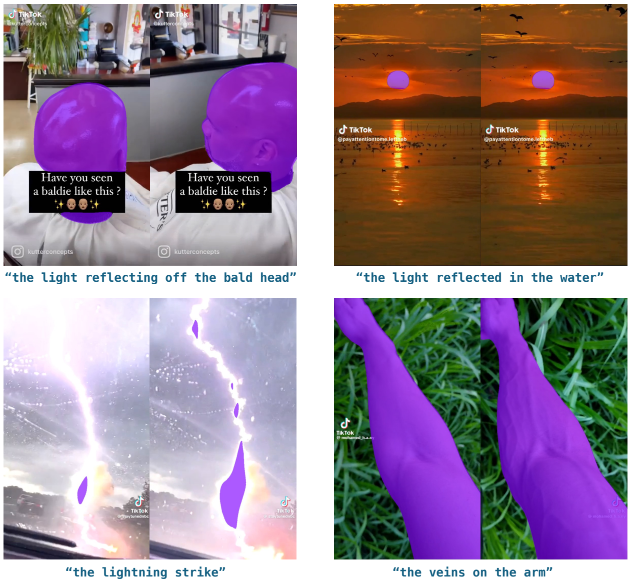
“the light reflecting off the bald head”
“the light reflected in the water”
“the lightning strike”
“the veins on the arm”

Figure E. Failure cases of REM (MS-1.4B) on Ref-VPS. Our model still exhibits some object-centric bias and struggles with extremely dynamic entities such as lightning.

C.3. Comparisons on ambiguous or overlapping scenarios

To assess how well our method handles visually ambiguous or overlapping scenarios, we present a qualitative comparison between REM and VD-IT, the strongest baseline on this benchmark, in Figure G. While many of these examples lack a single ground-truth segmentation, REM consistently produces more accurate and coherent masks in the confidently visible regions. For instance, in the first row, our method accurately segments only the clearly visible portions of lava that become apparent after being struck

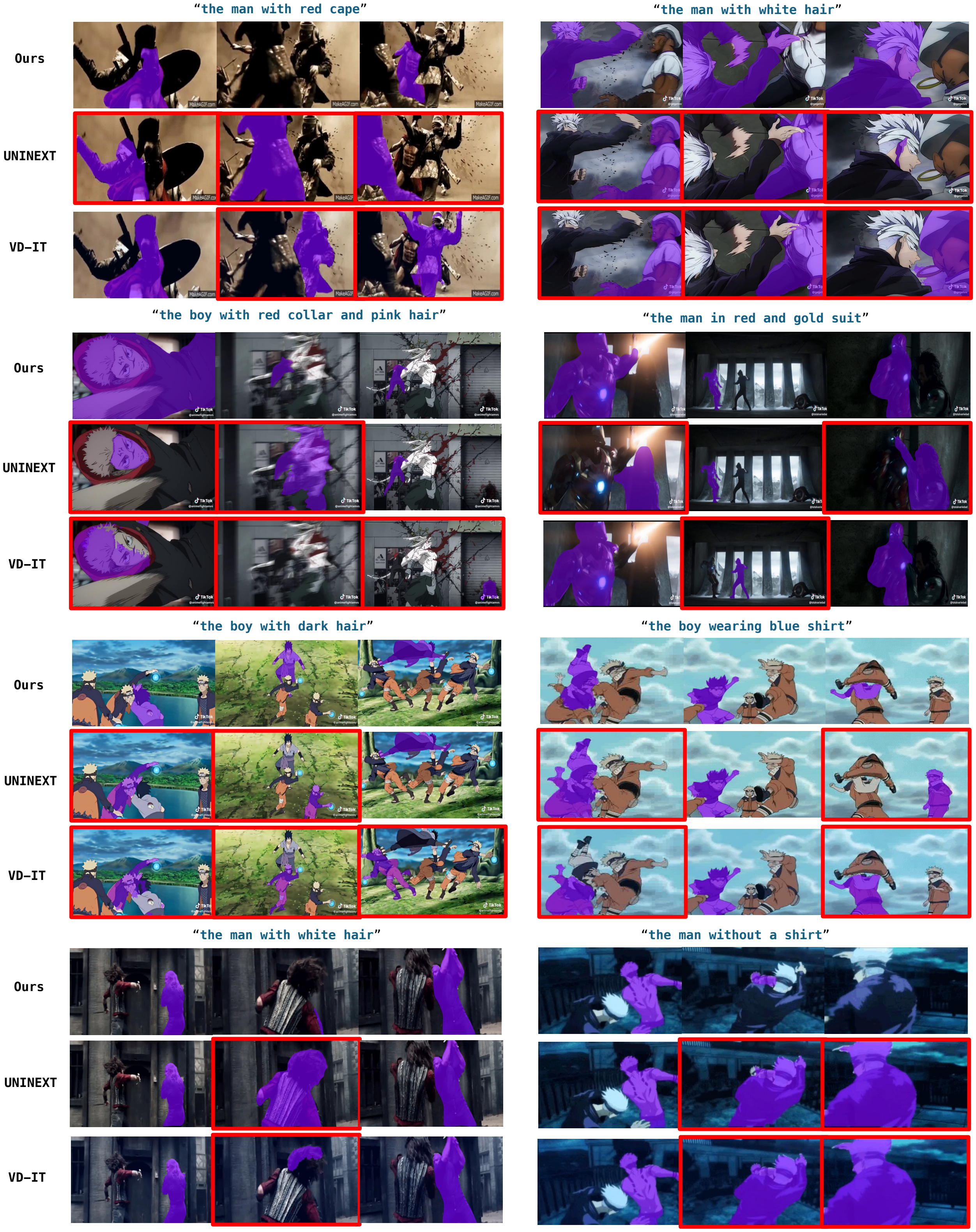
“the man with red cape”
“the man with white hair”
Ours
UNINEXT
VD-IT
“the boy with red collar and pink hair”
“the man in red and gold suit”
Ours
UNINEXT
VD-IT
“the boy with dark hair”
“the boy wearing blue shirt”
Ours
UNINEXT
VD-IT
“the man with white hair”
“the man without a shirt”
Ours
UNINEXT
VD-IT

Figure F. Qualitative comparison of REM (MS-1.4B) with state-of-the-art baselines on dynamic and challenging fight scenes. The incorrectly segmented frames are outlined in red. REM outperforms the other methods in handling frequent occlusions and POV changes. For a better illustration of the differences, please watch the full videos here.

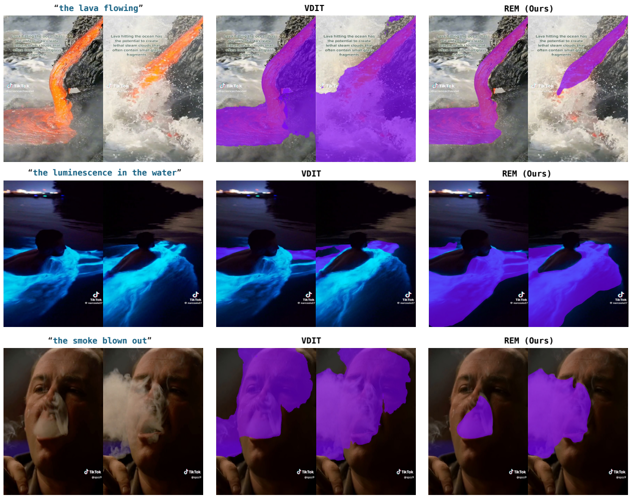

Figure G. Comparison on ambiguous or overlapping scenarios in Ref-VPS between VD-IT and REM (MS-1.4B). While no single perfect prediction exists for these samples, our method is both more precise and more consistent.

Benchmark	Type	Training Samples	Testing Samples
Ref-COCO [26]	Image	320K	-
Ref-YTB [19]	Video	12,913	2,096
MeViS [8]	Video	1,712	140
Ref-DAVIS [12]	Video	-	90
BURST [1]	Video	-	2,049
VSPW [15]	Video	-	343

Table J. Details about the benchmarks we used for training and evaluation.

by a wave, whereas VD-IT incorrectly includes the entire wave. In the second row, REM reliably segments all regions of glowing water, while VD-IT detects only a few scattered patches. These results demonstrate our model's robustness in ambiguous settings and its capacity to avoid over-segmentation.

D. Implementation Details

Benchmark details and baseline models We report the details about the benchmarks we used in Table J. We quote the results of all the baseline models on Ref-YTB, Ref-DAVIS, and MeViS from their original papers. For the zero-shot evaluation of BURST, VSPW, and Ref-VPS, we report the numbers by running the official checkpoints of MUTR[1], UNINEXT[2], VD-IT[3], and GLUS[4].

Training details. Our approach builds upon two state-of-the-art text-to-video diffusion architectures: ModelScope [21] and Wan [20]. Additional video diffusion backbones are evaluated in Section B. ModelScope comprises 1.4 billion parameters and extends Stable Diffusion [3] with temporal modules. We adopt a two-stage training protocol following Zhu et al. [27]: in Stage I, we fine-tune only the spatial weights on Ref-COCO image-text pairs[26] for one epoch; in Stage II, we fine-tune all network weights for 40 epochs using Ref-YTB video–text examples [19] supplemented with 12K Ref-COCO images converted into pseudo-videos following Wu et al. [22]. By contrast, Wan employs a unified diffusion transformer that jointly models spatial and temporal information, without dedicated temporal modules [16]. Accordingly, we train this variant in a single stage on the combined Ref-COCO and Ref-YTB datasets for 80k steps, with half of the steps trained with images and half trained with videos. Throughout training, the text encoder and VAE remain frozen.

Unless otherwise stated, all models are trained and evaluated at a resolution of 512 × 512. We use AdamW [14] for optimization with a constant learning rate of 1e-6. The training batch size is 4 for ModelScope and 8 for Wan, and for each sample, we randomly load an 8-frame video clip for ModelScope and a 17-frame video clip for Wan. We train our model using eight NVIDIA 80GB A100 GPUs, and it takes about 1 week to finish the whole training process.

For MeViS, we train our MS-1.4B variant by finetuning our Stage I checkpoint jointly on MeViS and Ref-YTB for 37 epochs. We achieve our best results on MeViS by finetuning the Wan-14B checkpoint trained on Ref-COCO and Ref-YTB, for an additional 8 epochs on MeViS.

Evaluation details. We follow the standard evaluation protocol for Ref-YTB, Ref-DAVIS, and MeViS. For BURST [1] and VSPW [15], neither of them contains referring text for the segmented entities. We automatically generate referring expressions using only the category information of the mask entity as "the <class>" (*e.g.*, *the hat*). For VSPW, we conduct our evaluation on the validation set, which has 66 different stuff categories. In the case of BURST, we evaluate the combined validation and test set, which contains 454 classes and a total of 2,049 sequences. For inference and evaluation, we follow the standard VSPW protocol for our RVS evaluation. For BURST, we predict masks for all the original frames, and compute the metrics for the annotated ones provided by the dataset. For Ref-VPS, to ensure high-quality performances, we perform inference at the original 24 FPS and compute evaluation metrics on the annotated frames at 6 FPS.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. BURST: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023. 5, 8

[1] https://github.com/OpenGVLab/MUTR
[2] https://github.com/MasterBin-IIAU/UNINEXT
[3] https://github.com/buxiangzhiren/VD-IT
[4] https://github.com/GLUS-video/GLUS

[2] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019. 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 4, 8

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 3

[5] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 4

[6] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 4

[7] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 3

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 8

[9] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023. 4

[10] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, 2024. 6

[11] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 3

[12] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2019. 8

[13] Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. GLUS: Global-local reasoning unified into a single large language model for video segmentation. In *CVPR*, 2025. 6

[14] Ilya Loshchilov, Frank Hutter, et al. Decoupled weight decay regularization. In *ICLR*, 2019. 8

[15] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021. 8

[16] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 8

[17] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017. 5

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[19] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 8

[20] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 8

[21] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 4, 8

[22] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022. 6, 8

[23] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 4, 5

[24] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *ECCV*, 2024. 6

[25] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *AAAI*, 2024. 5

[26] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 8

[27] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *ECCV*, 2024. 5, 8