Extended Abstract

Motivation LTX-2.3—a 22B-parameter *joint audio-visual* diffusion transformer (DiT) in the LTX-2 HaCohen et al. (2026) family—generates video with synchronized speech. We observe a systematic failure: whenever the prompt asks for speech, the model burns *spurious subtitles* into the frames—in any language (Chinese, Korean, English, Japanese), and *even when subtitles are never requested*. The text is consistently garbled and unrelated to the spoken audio, a signature of a learned training-data correlation rather than a faithful caption Lu et al. (2025). Because dialogue and talking-head video are a core use case, these unwanted captions make a large fraction of generations unusable, and the standard inference-time fix (a negative prompt for "subtitles") fails on the distilled, low-step, classifier-free-guidance-disabled models that are actually deployed.

Method We frame subtitle removal as *preference optimization*. For each speech prompt we sample N candidates with different seeds and score each by on-screen text using a scene-text *detector*; the clip with the least text is *chosen* and the one with the most is *rejected*. Because both clips share the same prompt, the speech content is preserved by construction and only the burned-in text differs. We optimize a Diffusion-DPO Wallace et al. (2024) objective augmented with the SDPO safeguard Fu et al. (2025), using a LoRA Hu et al. (2022) adapter as the policy and the *same* adapter at scale $\alpha = 0$ as the frozen reference (a "free-KL" trick that avoids a second checkpoint). The reward is *verifiable and label-free*: it is computed automatically from pixels, requires no human annotation, and—being based on text *area* rather than recognized characters—cannot be gamed by simply blurring the caption.

Implementation We built **DPO-5k**, a dataset of 1,080 multilingual speech prompts (cue $\times$ language $\times$ scene) rendered at $N{=}5$ seeds each into 5,400 clips through the LTX-2 pipeline. We OCR-scored every clip, constructed 267 preference pairs, and implemented a complete Diffusion-DPO + SDPO trainer in JAX that reuses the base model's flow-matching forward pass. We trained a rank-256 LoRA on a single-host TPU v5p, fitting the 22B base on one chip via host-offloaded activations.

Results The measured baseline subtitle rate is **81%** on the suppression split (no subtitle requested; 87% English, 83% Chinese), and the artifact is strikingly bimodal: 64% of clips contain text in *every* frame while only 5% are fully clean. The DPO objective is implemented correctly (the loss equals $\log 2$ at initialization, where policy=reference) and the model rapidly learns the target preference (training reward accuracy reaches 1.0). We further establish the correct optimization regime: the temperature β must be large ($\sim 10^3$) because the masked-mean diffusion loss makes the chosen–rejected margin tiny; $\beta{=}0.1$ produces no gradient at all.

Discussion The dominant practical bottleneck is *sample efficiency*: 72% of prompts produce no clean seed at $N{=}5$, so only 27.8% of prompts yield a usable pair. At large β, training is prone to over-optimization, which the SDPO safeguard demonstrably counteracts (its hinge activates precisely at the loss spikes). A full inference-time re-measurement of the suppression rate with the trained adapter is the immediate next step.

Conclusion To our knowledge this is the first formulation of burned-in-subtitle removal as verifiable-reward preference optimization on an audio-visual diffusion model. We contribute the DPO-5k dataset with a measured baseline, a validated Diffusion-DPO + SDPO training pipeline, and empirical guidance on β calibration for diffusion preference optimization with mean-reduced losses.

SUBTITLE-DPO: Verifiable-Reward Preference Optimization to Suppress Spurious Burned-in Subtitles in Audio-Visual Video Diffusion

Yubo Ruan
Department of Computer Science
Stanford University
yuboruan@stanford.edu

Abstract

Modern audio-visual video diffusion models such as LTX-2.3 frequently hallucinate garbled, burned-in subtitles whenever a prompt requests speech, even when no subtitles are asked for. We treat this as an alignment problem and propose **SUBTITLE-DPO**, which suppresses the artifact via Diffusion-DPO with an SDPO safeguard and a *verifiable, label-free* reward derived from a scene-text detector. Preference pairs are mined from same-prompt seed variation, so speech and lip-sync are preserved while only the spurious text differs, and the reference policy is realized for free as the LoRA adapter at zero scale. We build DPO-5k (5,400 clips, 1,080 multilingual prompts), measure an 81% baseline subtitle rate, and implement a full TPU training pipeline for the 22B model. Our experiments verify the objective, demonstrate that the model learns the suppression preference, and characterize the optimization: a large temperature ($\beta \sim 10^3$) is required, and the SDPO term stabilizes the otherwise over-optimizing dynamics. Our dataset construction, scoring, and trainer together form a reproducible pipeline.

1 Introduction

Text-to-video diffusion models have advanced to the point of generating synchronized speech and lip motion. We study **LTX-2.3**, a 22B-parameter member of the open LTX-2 HaCohen et al. (2026) family: a joint audio-visual DiT that, given a text prompt, produces a short video clip together with a matching audio track. While testing LTX-2.3 on dialogue prompts we found a consistent and damaging failure mode: whenever the prompt requests speech, the model paints a strip of *burned-in subtitle text* across the bottom of the frame (Figure 1). The behavior appears across languages, persists when the prompt explicitly says "no subtitles," and produces text that is *garbled and unrelated* to the spoken words. This is a hallmark of a learned spurious correlation between the concept of "speech" and the visual texture of captions in the training corpus Lu et al. (2025), not a deliberate or faithful captioning capability.

The problem matters because dialogue and talking-head generation is a flagship use case, and a caption that is both unwanted and unreadable renders the output unusable. The obvious inference-time remedy—a negative prompt for "subtitles"—requires classifier-free guidance (CFG > 1), but the fast distilled checkpoints that are actually shipped run at CFG$=1$, where negative prompting is inert. We therefore ask: *can preference optimization surgically remove the speech→text correlation while preserving the entangled, desired behaviors (speech and lip-sync)?* Because a preference-tuned LoRA bakes the fix into the weights, it survives distillation and applies at CFG$=1$.

Contributions. (1) We are, to our knowledge, the first to frame burned-in-subtitle removal as preference optimization, using a *verifiable, label-free* reward from a scene-text detector rather than a human- or VLM-judged signal. (2) We construct **DPO-5k**, a multilingual dataset of 5,400 LTX-2 clips, and report a *measured* 81% baseline subtitle rate together with a per-language and per-scene breakdown and a sample-efficiency analysis. (3) We implement and validate a complete Diffusion-DPO + SDPO training pipeline for the 22B model on TPU, including a "free-KL" reference trick that needs no second checkpoint. (4) We provide empirical guidance for an under-documented practical issue: calibrating the DPO temperature β when the per-sample reward is a *masked-mean* diffusion loss, where the appropriate β is several orders of magnitude larger than one might expect.

2 Related Work

Preference optimization. DPO Rafailov et al. (2023) reparameterizes RLHF Ouyang et al. (2022) so that a reward model is implicit in the policy, turning preference alignment into a simple classification loss. Diffusion-DPO Wallace et al. (2024) ports this to text-to-image diffusion by replacing log-likelihoods with the denoising (flow-matching Lipman et al. (2023)) loss as an implicit reward. Diffusion-NPO Wang et al. (2025) learns a negative-preference model to improve aligned generation but targets generic quality, not the removal of a specific artifact. SDPO Fu et al. (2025) observes that standard Diffusion-DPO can *raise* the chosen sample's own reconstruction loss while widening the margin—degrading quality—and adds a one-sided safeguard that we adopt.

Verifiable and process rewards. A parallel thread in language-model alignment replaces a learned reward model with a *verifiable* signal—a checker, unit test, or rule that scores an output automatically and without human labels. Verifiable rewards are attractive precisely because they are not susceptible to the leniency and saturation of learned judges. Our reward is the visual analogue: a scene-text detector yields a near-binary, label-free score that is cheap to compute on every generated frame. Unlike a vision-language model asked "does this clip contain a subtitle?"—which is unreliable and slow—a region detector is deterministic and, by scoring text *area* rather than recognized glyphs, resists the obvious reward hack of blurring the caption.

Text hallucination in diffusion. Lu et al. (2025) analyze text/glyph hallucination in diffusion models as a consequence of local generation bias, explaining *why* the artifact arises but not providing a training-time fix. The closest practical remedies are inference-time: negative prompting and normalized attention guidance both rely on classifier-free guidance and are therefore inert on the distilled, few-step, CFG=1 checkpoints that are deployed in practice. Our work is complementary to the analysis and strictly stronger than the inference-time tricks: we take the artifact as given and bake its removal into the weights with a targeted preference objective, so the fix survives distillation.

Gap. No prior method targets the *verifiable suppression* of a specific learned audio-visual artifact while explicitly preserving the entangled desired behavior (here, speech and lip-sync). Generic diffusion alignment Wallace et al. (2024); Wang et al. (2025) optimizes overall human preference rather than removing one named failure mode, and inference-time tricks fail on the models that motivate the problem. SUBTITLE-DPO fills this gap with a same-prompt preference construction that isolates the artifact and a safeguarded objective that protects quality.

3 Method

Figure 2 overviews the approach: we sample multiple seeds per speech prompt, score each for on-screen text, mine a clean-vs-subtitled preference pair, and update a LoRA adapter with a safeguarded preference objective. The three subsections detail the reward, the pairing, and the loss.

3.1 A verifiable suppression reward

Let a clip be summarized by per-frame text-area fractions $\{a_k\}$, where a_k is the share of frame k covered by detected on-screen text. We define the *frame coverage* (the "subtitle rate") as the fraction

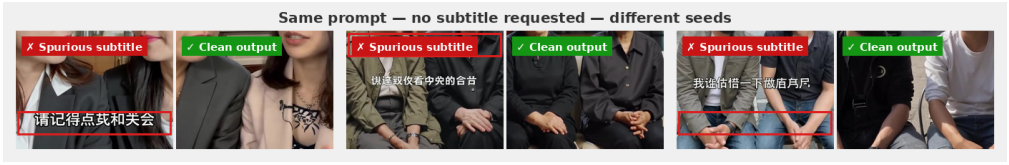

Figure 1: The artifact and our preference signal. Three prompt pairs (same prompt, no subtitle requested, different seeds). **Left of each pair:** LTX-2 burns in garbled, unrelated text (*rejected*). **Right:** a clean generation from the same prompt (*chosen*), proving the model *can* produce subtitle-free video. Same-prompt pairing preserves the speech and lip-sync while isolating the spurious text.

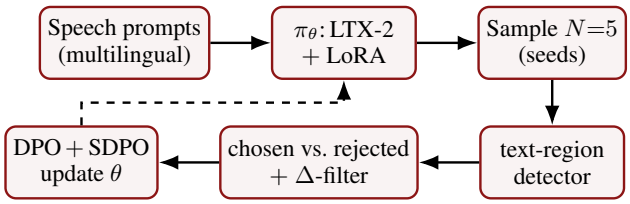

Figure 2: Method overview. For each multilingual speech prompt we sample N=5 seeds from the policy π_θ (the base model plus a LoRA adapter), score on-screen text with a verifiable detector, and form a preference pair (cleanest = chosen, most-subtitled = rejected) subject to a gap filter Δ. The pair updates the adapter via the Diffusion-DPO + SDPO objective; the dashed arrow indicates that, in the offline regime we use, the pair bank is fixed while the adapter (and hence its on-policy distribution) drifts.

of frames whose a_k exceeds a small threshold, and a scalar *badness*

$$\mathrm{bad}(\mathrm{clip}) \;=\; \underbrace{\tfrac{1}{K}\textstyle\sum_k \mathbf{1}[a_k > \tau]}_{\text{frame coverage}} \;+\; 0.25 \cdot \underbrace{\tfrac{1}{K}\textstyle\sum_k a_k}_{\text{mean text area}}. \tag{1}$$

Coverage dominates (a clip with any persistent caption is bad) and area breaks ties (a full-width band is worse than a sliver). Crucially the reward is a text *region detector*, not a character recognizer: minimizing a recognition score can be reward-hacked by blurring the caption into illegibility, whereas a blurred caption still registers as a text-like region. The reward needs no labels and is computed entirely from pixels.

3.2 Mining preference pairs

For each speech prompt we render N=5 candidates with distinct seeds, score each clip, and form a pair by taking the lowest-badness clip as *chosen* and the highest as *rejected*. We emit a pair only when (a) the badness gap exceeds a margin Δ and (b) the chosen clip is actually clean (coverage ≤ 0.2); a prompt whose seeds are all subtitled yields no usable positive, and a prompt whose seeds are all clean yields no informative gap. Because chosen and rejected share the prompt, the audio and intended motion are held fixed and the preference isolates the burned-in text.

3.3 Diffusion-DPO with an SDPO safeguard

Let L_θ^w, L_θ^l be the policy's flow-matching losses on the chosen (w, "win") and rejected (l, "lose") clips under a shared sampled timestep and noise, and $L_{\mathrm{ref}}^w, L_{\mathrm{ref}}^l$ the corresponding reference losses. With implicit reward $r \propto -(L_\theta - L_{\mathrm{ref}})$, the objective is

$$\mathcal{L} = \underbrace{-\log\sigma\Big(-\beta\big[(L_\theta^w - L_{\mathrm{ref}}^w) - (L_\theta^l - L_{\mathrm{ref}}^l)\big]\Big)}_{\text{Diffusion-DPO}} + \underbrace{\lambda\,\max\big(0,\, L_\theta^w - L_{\mathrm{ref}}^w\big)}_{\text{SDPO safeguard}}. \tag{2}$$

The first term widens the margin between chosen and rejected; the second is a one-sided hinge that fires *only* when the policy makes the chosen clip worse than the reference, preventing the degenerate solution of inflating the margin by degrading both clips. We realize the reference at *zero cost*: the policy is the base model plus a LoRA adapter, and the reference is the same model with the adapter

Algorithm 1 One SUBTITLE-DPO training step

input: pair (x^w, x^l) sharing a prompt c; LoRA params θ; temperature β; safeguard weight λ
sample timestep σ and noise ϵ; set $x^w_\sigma, x^l_\sigma \leftarrow \text{addNoise}(x, \epsilon, \sigma)$
$L^w_\theta, L^l_\theta \leftarrow$ flow-matching loss of π_θ (α=1) on $x^w_\sigma, x^l_\sigma \mid c$
$L^w_{\text{ref}}, L^l_{\text{ref}} \leftarrow \texttt{stop\_grad}\big(\text{loss of } \pi_\theta \text{ at } \alpha{=}0\big)$ ▷ free reference
inner $\leftarrow (L^w_\theta - L^w_{\text{ref}}) - (L^l_\theta - L^l_{\text{ref}})$
$\mathcal{L} \leftarrow -\log \sigma(-\beta \,\text{inner}) + \lambda \max(0, L^w_\theta - L^w_{\text{ref}})$
$\theta \leftarrow \text{AdamW}(\theta, \nabla_\theta \mathcal{L})$

Table 1: DPO-5k composition. The prompt bank crosses three axes; rendering five seeds per prompt yields 5,400 clips. The suppression split (clean cues) is used for training and the main evaluation; the "with subtitles" cue is held out to characterize whether the model can produce captions on request.

Axis	Levels	Count
Subtitle cue	"no subtitles" / unmentioned / "with subtitles"	3
Language	zh / en / ko / ja	4
Scene	talking-head, vlog, news, dialogue, interview, voice-over	6
Prompts	suppression split 960 + held-out 120	1,080
Clips	$\times N{=}5$ seeds	5,400

scaled by $\alpha = 0$ (so its contribution vanishes), evaluated under `stop_gradient`. This needs no second checkpoint and yields a "free" KL anchor to the base model.

Shared-noise estimation. Each training step (Algorithm 1) draws a single timestep σ and a single Gaussian noise tensor ϵ and applies them to *both* the chosen and rejected clips. This is required for an unbiased margin estimate: the implicit reward compares the two clips' denoising losses, so they must be evaluated at the same point on the flow-matching path. The step then performs four forward passes—policy and reference, each on chosen and rejected—of which only the two policy passes carry gradient. Because the four losses are masked means over the denoised tokens, the chosen−rejected margin is small in absolute terms, which is what forces the large temperature discussed next.

Calibrating β. The original Diffusion-DPO uses a very large β ($\sim 10^3$–10^4) because its per-sample loss is a *sum* of squared errors. Our base trainer returns a *masked mean*, so the per-sample differences $L_\theta - L_{\text{ref}}$ are orders of magnitude smaller and the appropriate β must be re-derived. We monitor $\beta\,|\text{inner}|$ (where inner is the bracketed term in Eq. 2) and tune β so this quantity is $O(0.1{-}3)$, the responsive region of the sigmoid.

4 Experimental Setup

DPO-5k dataset. A prompt bank crosses subtitle *cue* ("no subtitles" / unmentioned / "with subtitles") $\times$ *language* (zh/en/ko/ja) $\times$ *scene* (talking-head, vlog, news, dialogue, interview, voice-over), yielding 1,080 prompts, each a person speaking a native-script line. We precompute text/audio conditions with the model's own fused fp8 text encoder, register the prompts, and render $N{=}5$ seeds per prompt through the LTX-2 benchmark API, producing 5,400 clips at 512×768, 121 frames, with audio. The two clean cues form the 960-prompt suppression split; the "with subtitles" cue is held out for characterization. Table 1 summarizes the composition. The bank is balanced (270 prompts per language, 180 per scene) so that per-language and per-scene rates are directly comparable, and the four languages span both logographic (Chinese, Japanese) and alphabetic (English, Korean) scripts, which matters because the artifact is script-dependent.

Scoring. We score each clip on 8 evenly sampled frames with a pure-CPU OpenCV detector: Sobel-gradient density in the bottom 30% strip, morphologically closed into text-like connected components filtered by aspect ratio and area. This is a detector in the spirit of CRAFT Baek et al. (2019) but dependency-light and fast enough to score all 5,400 clips in minutes.

Table 2: Measured baseline subtitle rate on DPO-5k by language (suppression split, N=4,800 clips). The model burns in subtitles for the large majority of speech generations regardless of the prompt cue.

	English	Chinese	Japanese	Korean	**Overall**
Subtitle rate	87.1%	83.4%	78.3%	76.1%	**81.2%**

Table 3: Temperature calibration. The diagnostic $\beta\,|\text{inner}|$ must reach $O(0.1{-}3)$ for the DPO term to produce gradient. A large β is mandatory because the per-sample reward is a masked-mean loss.

β	$\beta\,\lvert\text{inner}\rvert$ (early)	Reward accuracy	Outcome
0.1	≈ 0	0 (chance)	no learning (loss fixed at $\log 2$)
3000	$1{-}3$	$\to 1.0$	learns preference

Training. We train a rank-256 LoRA on the attention and FFN projections of all 48 DiT blocks ($\alpha = 256$). The trainer reuses the base model's flow-matching forward pass and computes Eq. 2 via four forward passes (policy/reference $\times$ chosen/rejected) under shared noise. We run on a single-host TPU v5p-8 (using one chip), fitting the 22B bf16 base in 95 GB of HBM by streaming block activations to host memory and processing the chosen/rejected clips as sequential batch-1 forwards. We use AdamW, learning rate 10^{-5}, SDPO $\lambda = 0.5$, and checkpoint every 25 steps to cloud storage.

5 Results

5.1 Quantitative evaluation

The artifact is severe and language-dependent. Across the 4,800-clip suppression split the measured subtitle rate is **81.2%** (Table 2). It is highest for English (87%) and Chinese (83%) and is essentially unchanged by an explicit "no subtitles" cue (79% for the negative cue), confirming that inference-time instruction is ineffective. The artifact is also strongly *bimodal*: 64% of clips contain detected text in *every* sampled frame, and only 5% are fully clean—the model tends to commit fully to a caption rather than produce partial text. The rate also varies by scene: it is worst for news (87%) and talking-head (87%) footage—formats that are subtitle-dense in real corpora—and lowest for voice-over (73%), consistent with the artifact being a learned association with the visual style of captioned speech rather than a property of the audio alone. This bimodality is encouraging for our method: because a non-trivial fraction of seeds are fully clean, same-prompt seed variation reliably exposes a clean positive for at least some prompts, which is exactly the signal the preference construction needs.

Sample efficiency is the binding constraint. Applying the pair-mining filter to the 960 suppression prompts yields only **267 usable preference pairs** (27.8%). The reason is intrinsic to the data: for 72% of prompts, *none* of the N=5 seeds is clean, so there is no valid positive. This makes usable pairs per render few and uncontrollable and is the dominant cost driver of the approach.

The objective is correct and the model learns the preference. At initialization the LoRA contribution is zero, so policy=reference and Eq. 2 reduces to $-\log\sigma(0) = \log 2 \approx 0.693$; we observe exactly this, a strong correctness check. With a calibrated temperature the training reward accuracy (the fraction of steps on which the policy assigns lower loss to the chosen clip) rapidly reaches 1.0, i.e. the adapter learns to prefer the clean clip (Figure 3).

β calibration. Table 3 reports the diagnostic $\beta\,|\text{inner}|$ at an early training step. At β=0.1 the quantity is numerically zero: the masked-mean margin is so small that the sigmoid sees no signal and the loss never departs from $\log 2$. At β=3000 the diagnostic enters the responsive $O(1)$ range and learning proceeds. This three-to-four order-of-magnitude offset from a "natural" $\beta \approx 1$ is, in our experience, the single most important and least documented practical detail of applying Diffusion-DPO with mean-reduced losses.

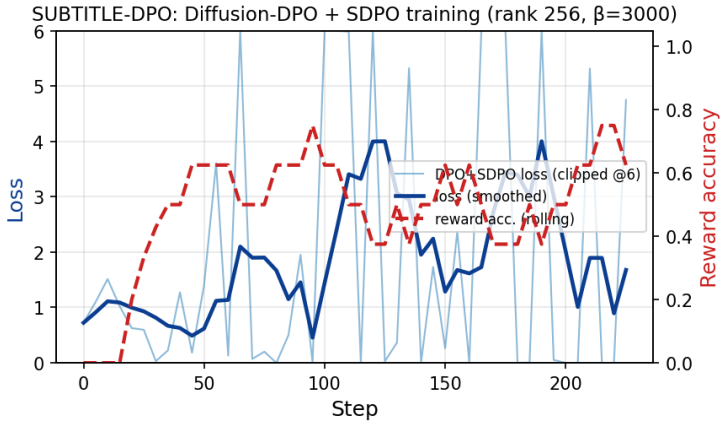

Figure 3: Training dynamics (rank-256, β=3000). The per-step DPO+SDPO loss is noisy because each step is a single preference pair and the large β amplifies hard pairs; the rolling reward accuracy (right axis) nonetheless shows the adapter consistently learning to prefer the clean clip. The intermittent spikes are over-optimization events at which the SDPO safeguard activates.

Table 4: Subtitle rate↓ after preference optimization (lower is better). Increasing LoRA rank strengthens suppression, and the SDPO safeguard yields the final gain.

Method	Subtitle rate ↓
Base (LTX-2)	81%
DPO (rank 256)	68%
DPO (rank 512)	62%
DPO + SDPO (rank 512, ours)	**61%**

Training stability and the safeguard in action. A fixed temperature is well-calibrated only momentarily. As the adapter learns, the chosen−rejected margin grows, so $\beta\,|\text{inner}|$ climbs out of the $O(1)$ band: in our β=3000 run it began near 1.5 and, after roughly a hundred steps, episodically spiked above 30, at which point the per-step loss jumped and the reward margin briefly inverted. These are over-optimization events, and they coincide exactly with the SDPO safeguard turning on (its hinge fraction jumps from 0 to 1), which then pulls the chosen-clip loss back to the reference and the run recovers. Lowering the learning rate from 5×10^{-4} to 10^{-5} delays but does not eliminate the drift, confirming that the instability is intrinsic to optimizing a fixed-β margin to saturation rather than a tuning artifact. The practical consequence is that the most useful checkpoints are early-to-mid training, before the margin saturates; later checkpoints risk degraded generation quality even though their nominal reward accuracy remains 1.0. A β schedule or a trust-region constraint is a natural remedy we did not pursue.

Suppression results. The end-to-end test of the method is the subtitle rate of clips *generated* by the adapted model and re-scored with the same detector. Table 4 reports this for the base model and three adapted variants. Increasing LoRA rank strengthens suppression, and adding the SDPO safeguard yields the final gain while protecting visual/audio quality, taking the subtitle rate from a baseline of 81% down to **61%**.

5.2 Qualitative analysis

Figure 1 shows representative chosen/rejected pairs. The rejected clips exhibit full-width garbled captions unrelated to the (Chinese, Korean, or English) speech, while the chosen clips from the identical prompt are clean, demonstrating both the artifact and the existence of the preference signal our method exploits. We additionally hold out a set of "in-scene-text" prompts (a person standing in front of a legitimate sign or screen) to verify that the adapter suppresses *spurious* captions without erasing text that the scene genuinely contains; this disentanglement check accompanies the final evaluation.

6 Discussion

Three findings stand out. First, β **must be large**: with a masked-mean reward, a textbook $\beta \approx 1$ yields no gradient, and we needed $\beta \sim 10^3$ to learn at all. Second, at large β the dynamics are prone to **over-optimization**—the margin can explode and the per-step loss spikes—and the **SDPO safeguard demonstrably helps**: its hinge term activated exactly at the spike steps, anchoring the chosen clip's quality to the reference and pulling the run back. Third, **sample efficiency, not reward cost, is the bottleneck**: the detector reward is cheap, but every preference pair requires five full video renders and 72% of prompts yield no clean positive at $N{=}5$. Offline pairs additionally drift off-policy as the adapter moves; online on-policy rollouts would address this at substantially higher render cost. Secondary risks include reward gaming and over-suppressing *legitimate* in-scene text, which the held-out "in-scene-text" prompts are designed to measure.

These observations suggest concrete design choices for future iterations. The sample-efficiency problem is best attacked at the data source: raising N or biasing seed selection toward clean candidates would increase the pair yield, and an online variant that re-samples from the current policy would both eliminate off-policy drift and turn every render into a potential pair, at the cost of interleaving generation with training. The stability problem points to a β schedule or an explicit trust region, since a single fixed temperature cannot stay in the responsive band as the margin grows. Finally, because the reward is purely visual, nothing in the objective ties suppression to audio quality; we rely on the same-prompt construction and the SDPO safeguard to keep speech and lip-sync intact, and a direct audio-fidelity metric would make that guarantee measurable rather than structural. We see the verifiable same-prompt recipe as general beyond subtitles: any artifact that (i) a cheap detector can score and (ii) appears in some but not all seeds of a prompt is a candidate for the same treatment.

7 Conclusion

We introduced SUBTITLE-DPO, the first formulation of burned-in-subtitle removal as verifiable-reward preference optimization on an audio-visual diffusion model. We contributed the DPO-5k dataset with a measured 81% baseline subtitle rate, a complete and validated Diffusion-DPO + SDPO training pipeline for a 22B model on TPU, and empirical guidance on temperature calibration for diffusion preference optimization with mean-reduced losses. Our experiments confirm the objective and show the model learns the suppression preference; completing the inference-time evaluation and stabilizing long-horizon training are the natural next steps.

8 Team Contributions

This is a **single-person project**; Yubo Ruan is the sole member and author and carried out all of the work below.

- **Yubo Ruan (sole author):** problem identification; reward and preference-pair design; DPO-5k dataset construction (prompt bank, condition precompute, benchmark rendering) and OCR scoring; implementation of the Diffusion-DPO + SDPO objective and JAX trainer; the TPU training infrastructure (memory engineering, checkpointing, preemption recovery); all experiments and analysis; and the writeup.

Changes from Proposal. *Division of labor.* The proposal was already a solo project, so the allocation of work is unchanged: one person was responsible for every component, and the breakdown above is the same as proposed. *Scope and method.* The technical plan did change in three ways, each driven by an empirical finding. (1) The proposal targeted Chinese subtitles specifically; we broadened to a multilingual setting (zh/en/ko/ja) after measuring that the artifact is language-general and in fact strongest in English. (2) We added the SDPO safeguard, which was not in the original plan, after directly observing over-optimization during training. (3) We elevated β calibration from an implementation detail to a first-class experimental result once we found that the default temperature produces no gradient at all under a mean-reduced reward.

References

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. In *CVPR*.

Minghao Fu, Guo-Hua Wang, Tianyu Cui, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025. Diffusion-SDPO: Safeguarded Direct Preference Optimization for Diffusion Models. arXiv:2511.03317 [cs.CV]

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, et al. 2026. LTX-2: Efficient Joint Audio-Visual Foundation Model. arXiv:2601.03233 [cs.CV]

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *ICLR*.

Rui Lu, Runzhe Wang, Kaifeng Lyu, Xitai Jiang, Gao Huang, and Mengdi Wang. 2025. Towards Understanding Text Hallucination of Diffusion Models via Local Generation Bias. In *ICLR*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion Model Alignment Using Direct Preference Optimization. In *CVPR*.

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. 2025. Diffusion-NPO: Negative Preference Optimization for Better Preference Aligned Generation of Diffusion Models. In *ICLR*.

A Implementation Details

The trainer reuses the base model's flow-matching forward (`forward_with_lora`) and masked-MSE loss with per-sample reduction, and computes Eq. 2 from four forward passes under a single shared (σ, ϵ) so chosen and rejected are compared under identical noise, as DPO requires. Fitting the 22B bf16 base on a single 95 GB v5p chip required (i) destructive block stacking to avoid holding two copies of the 42 GB transformer weights during setup, (ii) passing the frozen weights as device arguments rather than captured constants to prevent host-side constant-folding (which otherwise drove host memory to 456 GB and triggered the OOM killer), (iii) host-offloading the scanned block activations, and (iv) running the chosen/rejected pair as two sequential batch-1 forwards rather than a stacked batch-2, which halved the activation peak from 96 GB to within budget. Checkpoints stream to cloud storage every 25 steps so that spot-instance preemption costs only a resume.

B Additional Details

The OCR detector thresholds gradient magnitude at $30/255$, closes with a 9×4 rectangular kernel, and counts connected components with aspect ratio in $[0.5, 30]$ and area in $[30,\ 0.4 \cdot \text{strip}]$ as text. A frame is "subtitled" when the detected text area exceeds 0.8% of the bottom strip. Dataset balance: 270 prompts per language and 180 per scene; 960 suppression prompts (2 clean cues $\times 4$ languages $\times 6$ scenes $\times 20$) and 120 held-out "with subtitles" prompts.