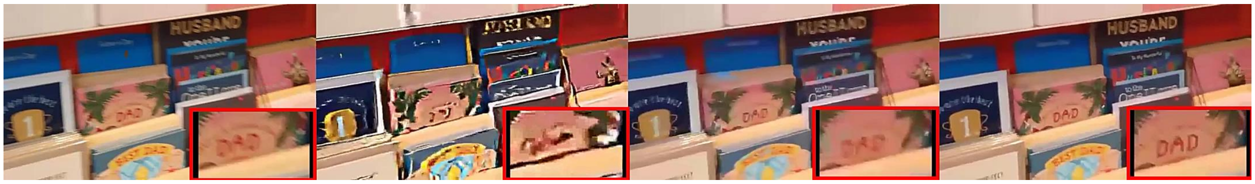

Figure 1: A real video example. From left to right: Input, BIN, TNTT* and ours. Adobe Reader with flash player is recommended to watch this video (click to play). Users may need to enable the 'Preferences->3D&Multimedia->Use Flash Player ...' option.

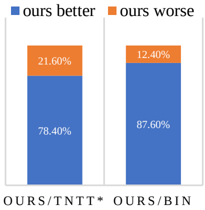

Figure 2: User study

Table 1: Runtime and model size comparison

Method	Runtime (s)	Parameters (million)
TNTT	0.33	10.7
BIN	2.22	11.4
Ours	0.73	34.4

Table 2: Ablation study on key-states restoration network

Model	PSNR	SSIM
Cascade stage-I	30.76	0.9484
Input 2 frames	30.74	0.9514
Proposed	31.72	0.9597

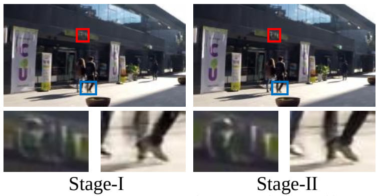

Figure 3: Comparison of outputs from different stage in deblurring network.

We sincerely thank all reviewers for their constructive comments. All concerns will be addressed in the final version.

Common concern #1: Limitations of synthetic data, the synthesis procedure is not physically correct. **(R#1-Q1;R#2-Q1)**

A: Comparing with real frames, synthetic data have two limitations: 1) during the shutter open, a recorded 'sharp ground-truth' may be blurry; 2) since the shutter close, the content missing (*i.e.,* discrete accumulation) may result in 'ghosting' artifact. However, these two flaws mainly arise when recorded a large movement, more specifically, when the relative displacement during a shutter open/close period (*e.g.,* 1/480-s) crossed two or more pixels. Our work adopted the same synthetic data and 'moderate' motion assumption with existing deblurring and blurry VFI methods. According to previous studies, synthetic data is now the best choice for simulating unavailable real training pairs.

Common concern #2: Results on real videos and related user studies should be provided. **(R#1-Q1;R#2-Q2;R#4-Q5)**

A: In our supplementary video (after 1m29s), two interpolated videos of real scenes are reported. In addition, we shot ten 10-sec real blurry videos using a telephone camera. A user study (Fig.2) collected through Amazon Mechanical Turk shows our method achieved significant improvement on real videos. For each comparison pair, a user was asked to select a better video. More than 1k responses are collected, and all videos were sorted randomly to avoid cheating. Since the space limitation, we report a short video clip in Fig.1, all video results will be released with our codes.

Common concern #3: The improvement is more significant in the 5-5 setting; is not that large in some cases. **(R#2-Q3;R#3-Q1)**

A: In our experiments, TNTT*, an improved variant implemented by ourselves, is the only model that achieved comparable results in some settings. Yet, it still faces the generalization problem that our work aims to solve. With one well-trained model, our method showed constant superiority on both synthetic data and more challenging real videos.

Common concern #4: The title issue, 'generalized' does not suit very well to the context of this paper. **(R#1-Q3;R#4-Q1)**

A:We will replace our title as 'Video Frame Interpolation without Temporal Priors' and polish the main text accordingly.

R#1-Q2: Comparing the model size and running speed of the proposed methods with existing works.

A: As shown in Table 1, we adopted the official codes, tested all methods on the same task (8x interpolation) and the same hardware. Note that BIN focuses on 2x interpolation, in our setting, it performed the interpolation repeatedly.

R#3-Q2: How reasonable the constant acceleration assumption is in practice?

A: We adopt this assumption from QVI, a SOTA method for sharp VFI (maintext-L90). In our setting, this assumption only needs to hold for two consecutive blurry frames (around 1/20-s for a 30fps video). According to our experiments, the derived curves can handle most cases. For challenging cases, we hope to relax this assumption in our future work.

R#3-Q3: Pls expand on the limitation of this work. (The trajectory prior can be only used to refine one optical-flow.)

A: When we employ the trajectory prior to refine the calculated pixel displacement $\hat{\boldsymbol{S}}_{\boldsymbol{23}}$, it is based on an assumption that estimated optical-flows $\boldsymbol{f_{0\rightarrow1}}$ and $\boldsymbol{f_{1\rightarrow2}}$ are accurate. However, they may exist errors. In future work, we hope to introduce a new trainable module to extract more accurate displacements (or optical-flow) for our interpolation.

R#4-Q2: How Eq.2 is derived; and how Eq.4 can be degraded to Eq.1?

A: Eq.2 is derived from the equation sets: $\{\boldsymbol{s}_{12} = \boldsymbol{v_1}t_1 + \frac{1}{2}\boldsymbol{a}t_1^2\}$ and $\{\boldsymbol{s}_{01} = \boldsymbol{v_0}t_0 + \frac{1}{2}\boldsymbol{a}t_0^2; \boldsymbol{s}_{23} = \boldsymbol{v_2}t_2 + \frac{1}{2}\boldsymbol{a}t_2^2; \boldsymbol{v_1} = \boldsymbol{v_0} + \boldsymbol{a}t_0; \boldsymbol{v_2} = \boldsymbol{v_1} + \boldsymbol{a}t_1; t_0 = t_2\}$. Eq.4 can be degraded to Eq.1 when we set $\lambda = 1$ and substitute $2\boldsymbol{S_{12}} - \boldsymbol{S_{01}}$ for $\boldsymbol{S_{23}}$ according to Eq.3. Since space limitation, detailed derivation will be provided in the final version.

R#4-Q3: More discussions on the restoration network are required.

A: In our restoration network, the first stage mainly focuses on figuring out the frame sequence with correct temporal order. The second stage aims to refine the output of the first stage using the proposed second-order residual structure. As shown in Fig. 3, there exists a severe artifact in stage-I's output. In addition, the newly added ablation study (Table 2) shows that, even employing the same amount of parameters, the model simply repeats stage-I's architecture (*i.e.,* cascade stage-I) performs inferior to our proposed restoration network.

R#4-Q4: What is the temporal ambiguity; why utilize 4 frames but not 2?

A: For a single blurry frame, temporal ambiguity means there exist two possible outputs of the start/end states. Generally, two or more consecutive frames are required to decide the temporal order. In both TNTT and our work, 4 input frames are employed to reduce the ambiguity and improve deblurring results. A new ablation study is provided in Table 2.

R#4-Q5: Writing issues.

A: Thanks for pointing our typos. It should be B_1 and B_2 in L-143. We will carefully proofread and polish our draft.