# Supplementary Material: Image Animation with Perturbed Masks

Yoav Shalev Lior Wolf
Blavatnik School of Computer Science, Tel Aviv University
yoavshalev@mail.tau.ac.il, wolf@cs.tau.ac.il

## A. Architecture

Following [1], the mask generator $m$, the mask refinement network $r$ and the high-res generator $h$ have the same encoder-decoder architecture, followed by a $conv_{7\times7}$ layer and a $sigmoid$ activation. The encoder (decoder) consists of five encoding (decoding) blocks, where each encoding block is a sequence of $conv_{3\times3} - relu - batch\_norm - avg\_pool_{2\times2}$, and each decoding block is a sequence of $up\_sample_{2\times2} - conv_{3\times3} - batch\_norm - relu$. Only for the high-res generator $h$ we add skip connections from each of the encoding layers to its corresponding decoding layer, to form a U-Net architecture.

The encoder of the low-res generator $\ell$ consists of $conv_{7\times7} - batch\_norm - relu$ , followed by six residual blocks, each block consists of $batch\_norm - relu - conv_{3\times3} - batch\_norm - relu - conv_{3\times3}$. The decoder consists of two blocks, each is a sequence of $up\_sample_{2\times2} - conv_{3\times3} - batch\_norm - relu$. The decoder is followed by a $conv_{7\times7}$ layer and a $sigmoid$ activation.

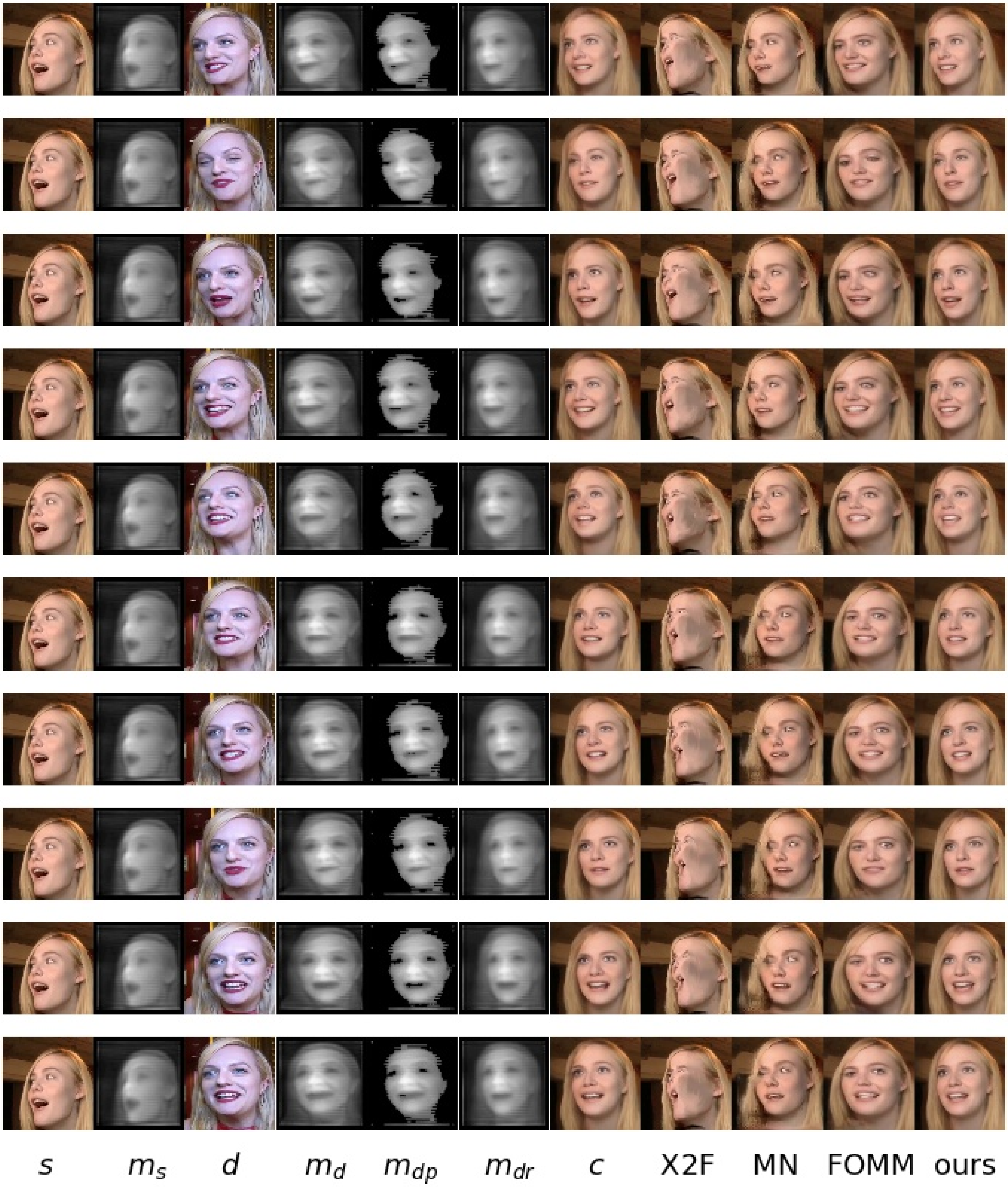
## B. Additional Qualitative Results

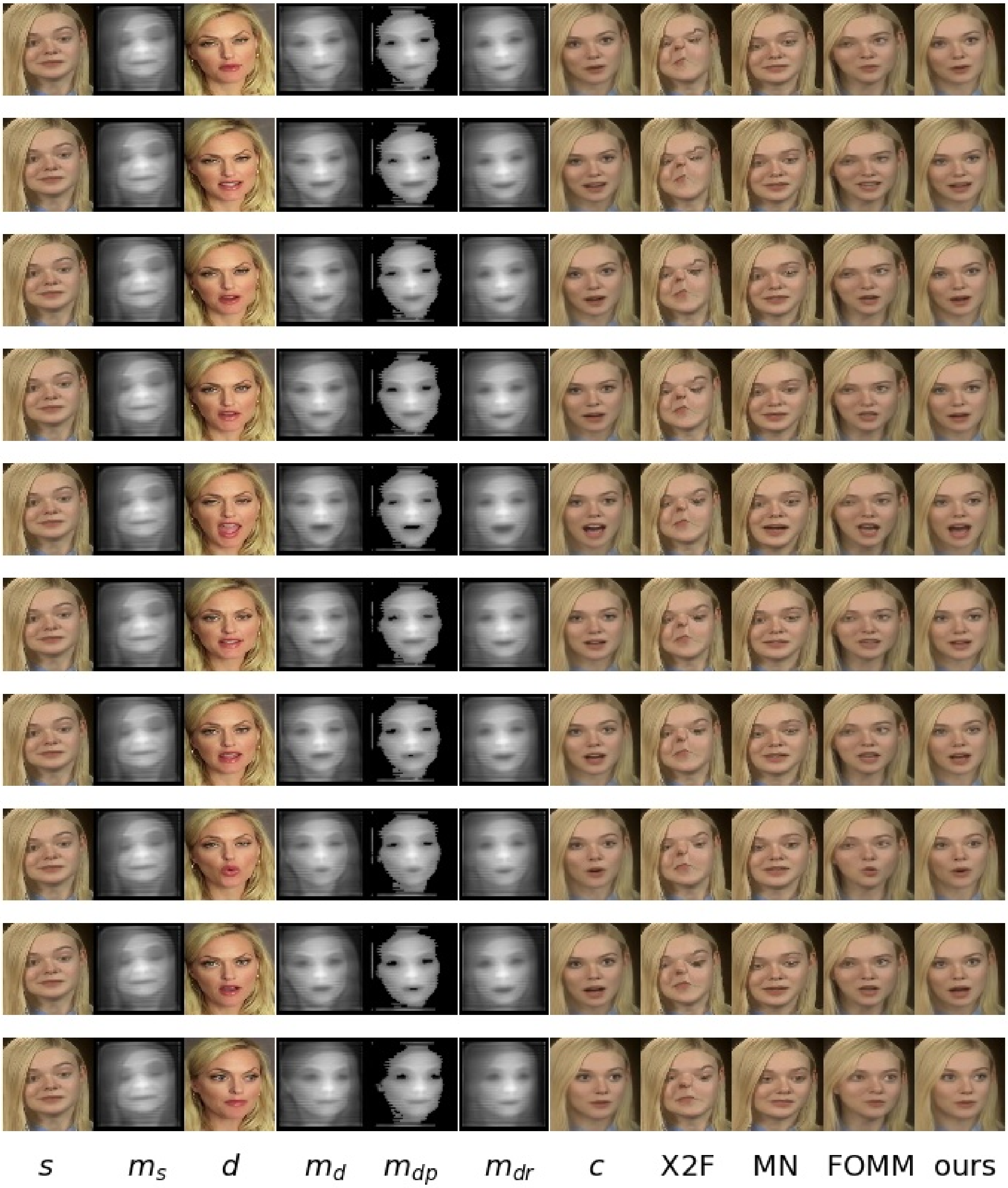
In Fig. 1, Fig. 2 and Fig. 3 we added final and intermediate results generated by our method for the VoxCeleb, Tai-Chi-HD and BAIR datasets, compared to the SOTA methods. The full videos are available at https://github.com/itsyoavshalev/Image-Animation-with-Perturbed-Masks.

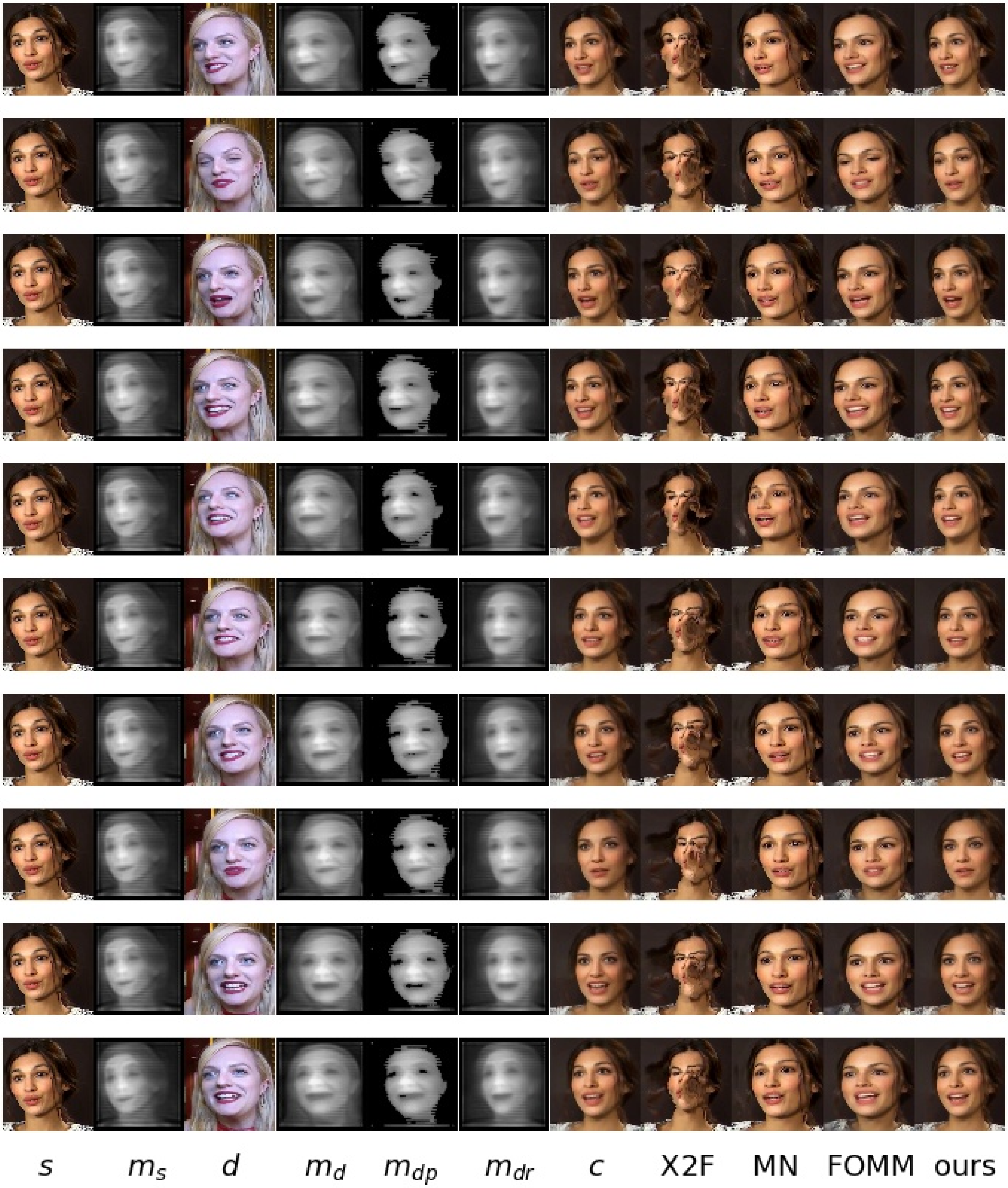
## References

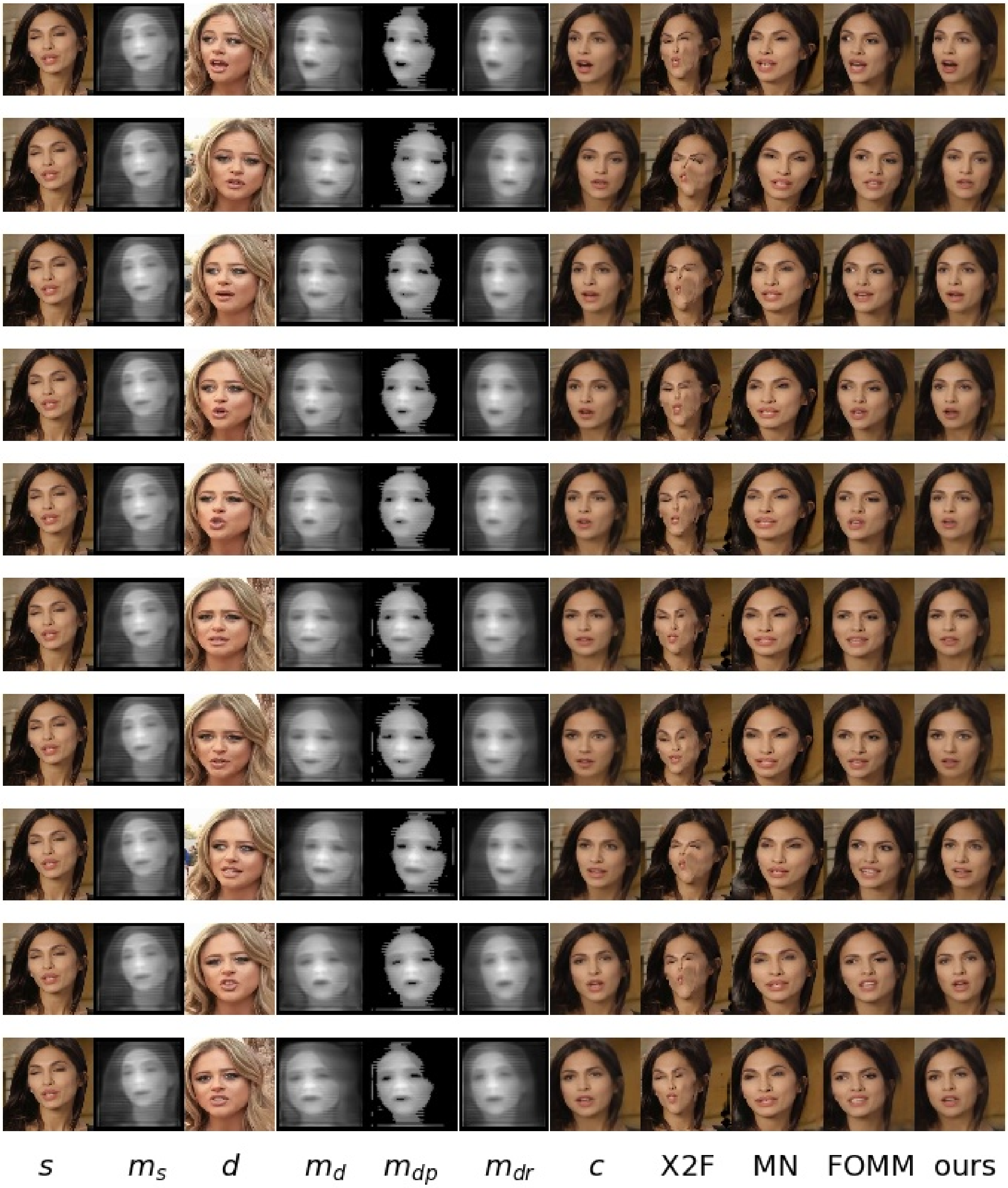
[1] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pages 7137–7147, 2019. 1

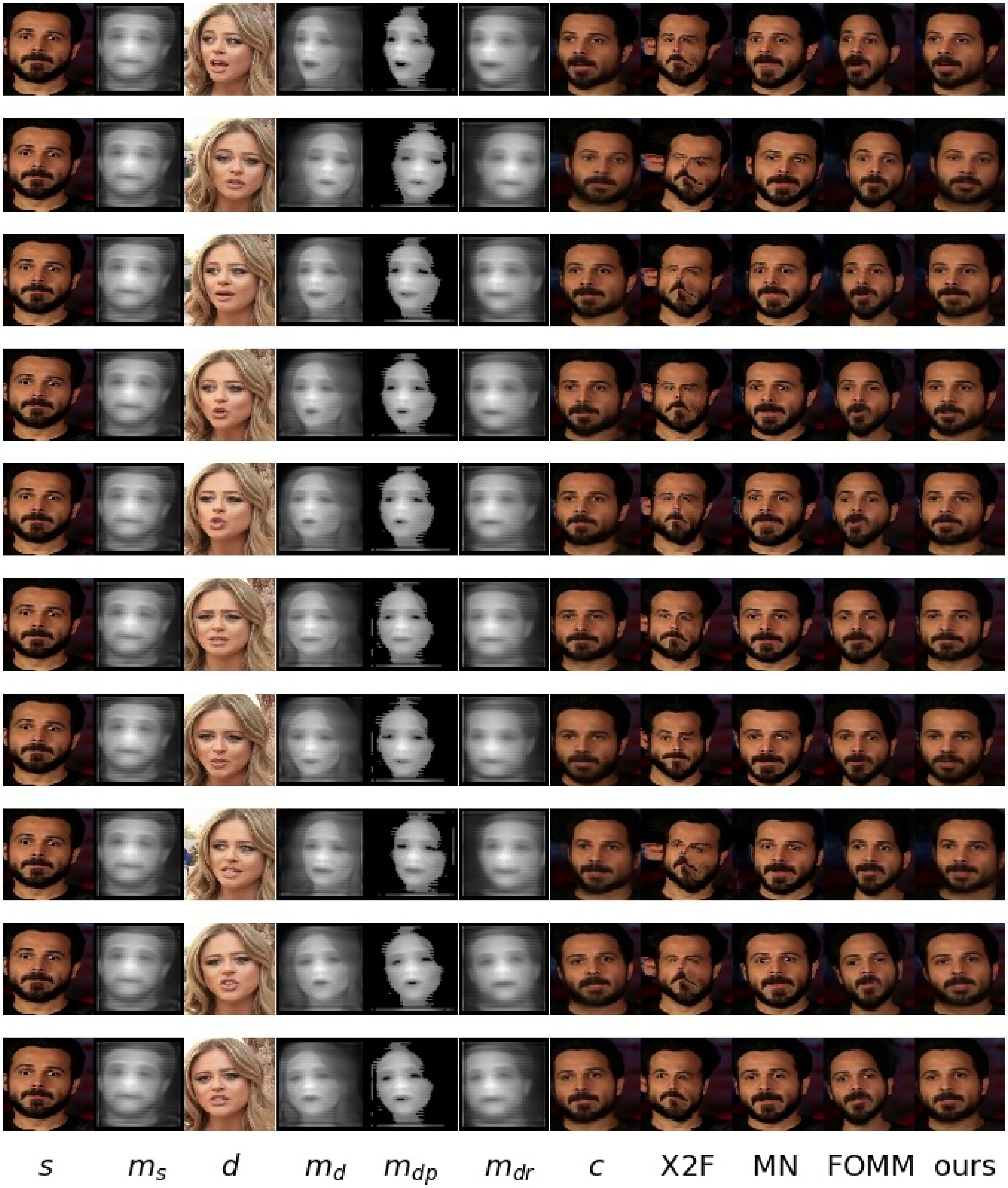
Figure 1. Final and intermediate results generated by our method for VoxCeleb, compared to the SOTA methods.

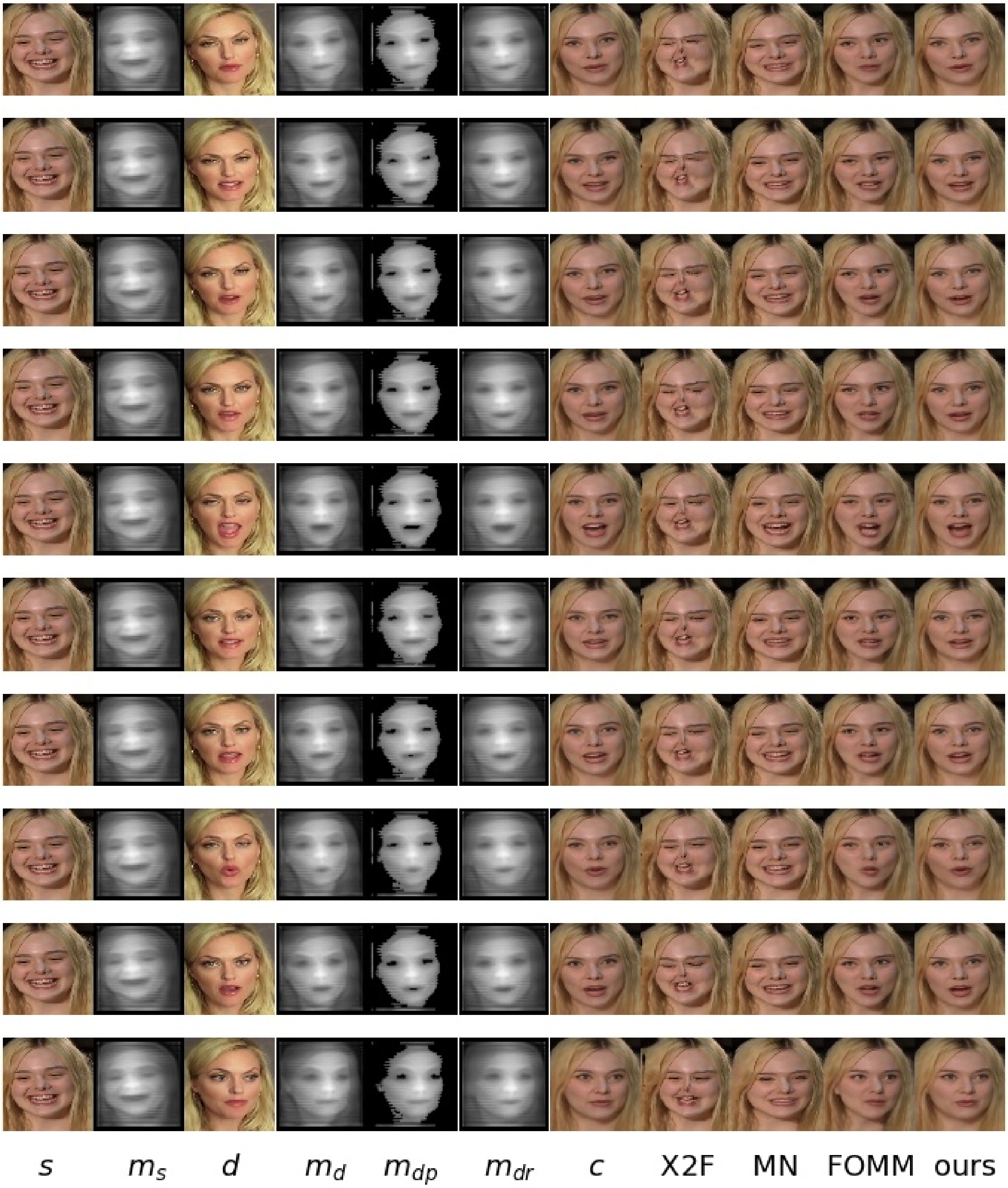


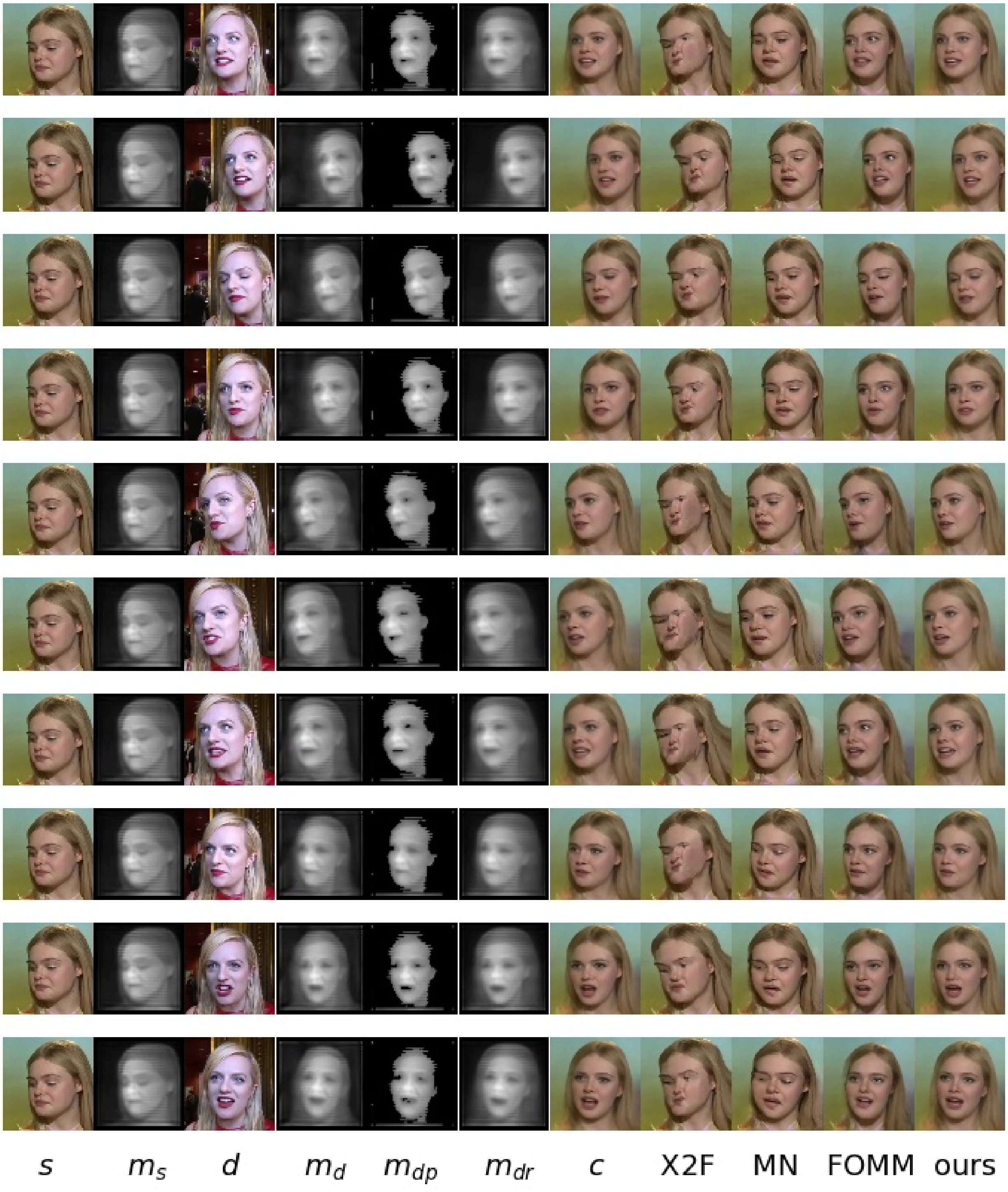

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

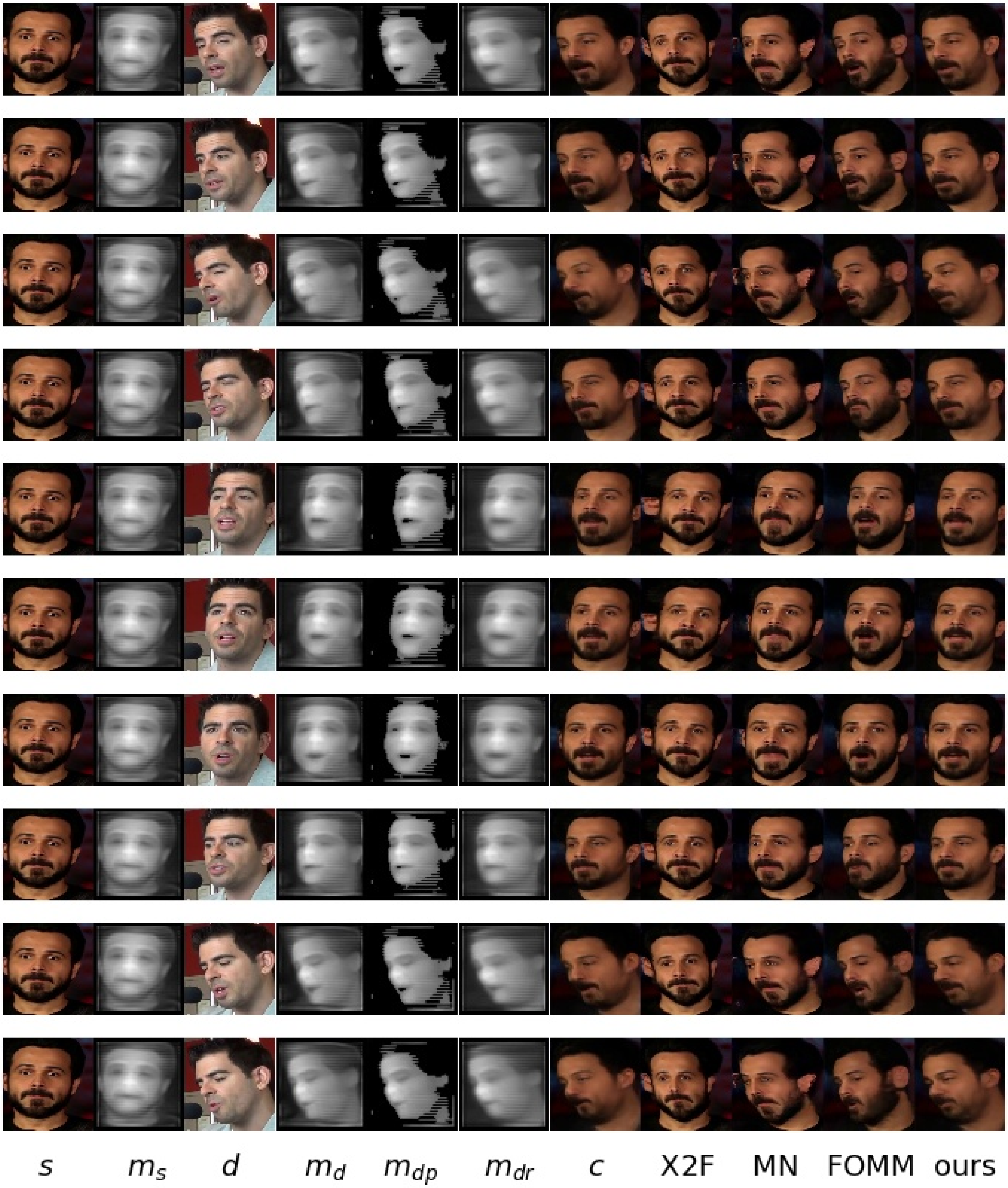
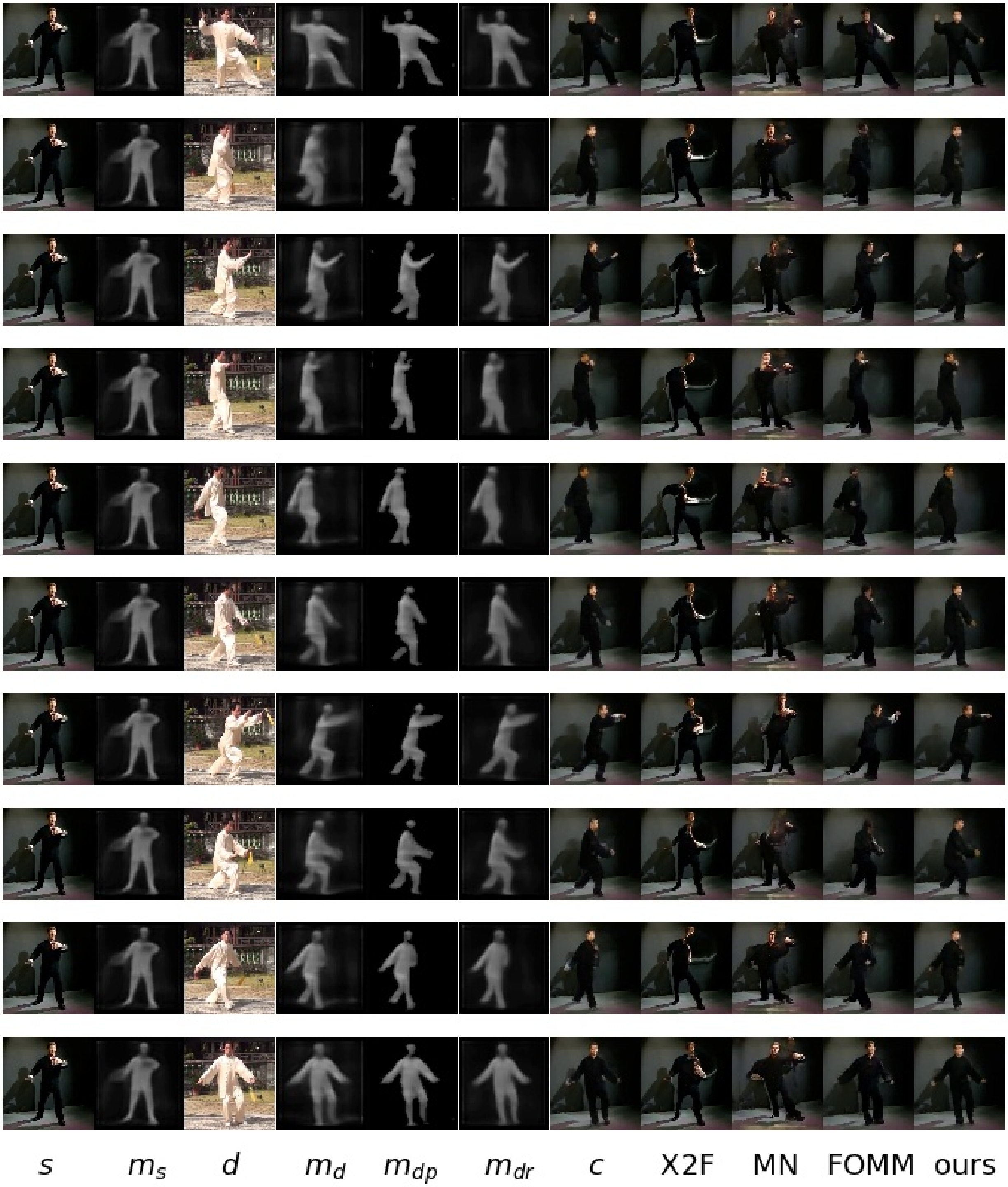
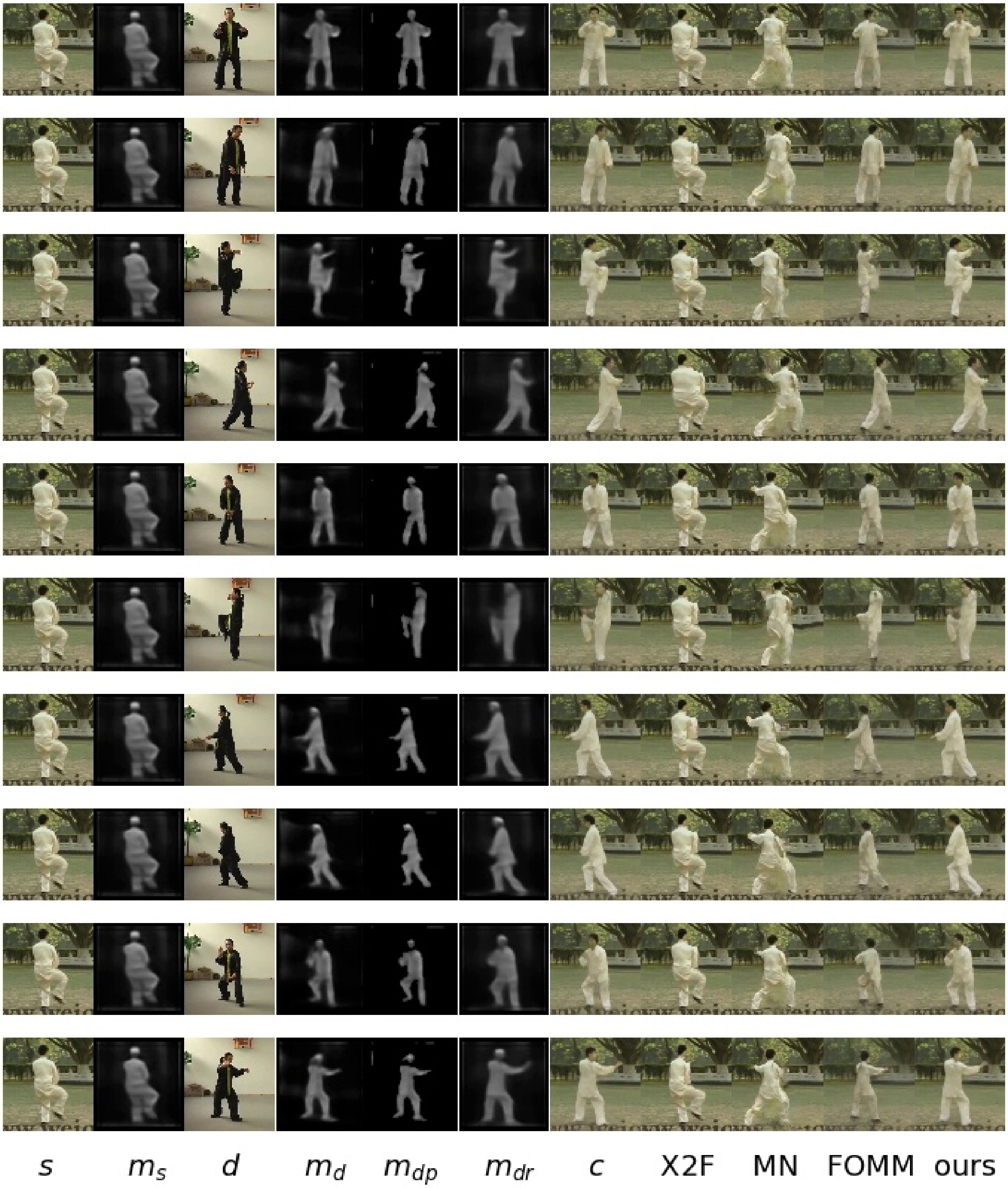
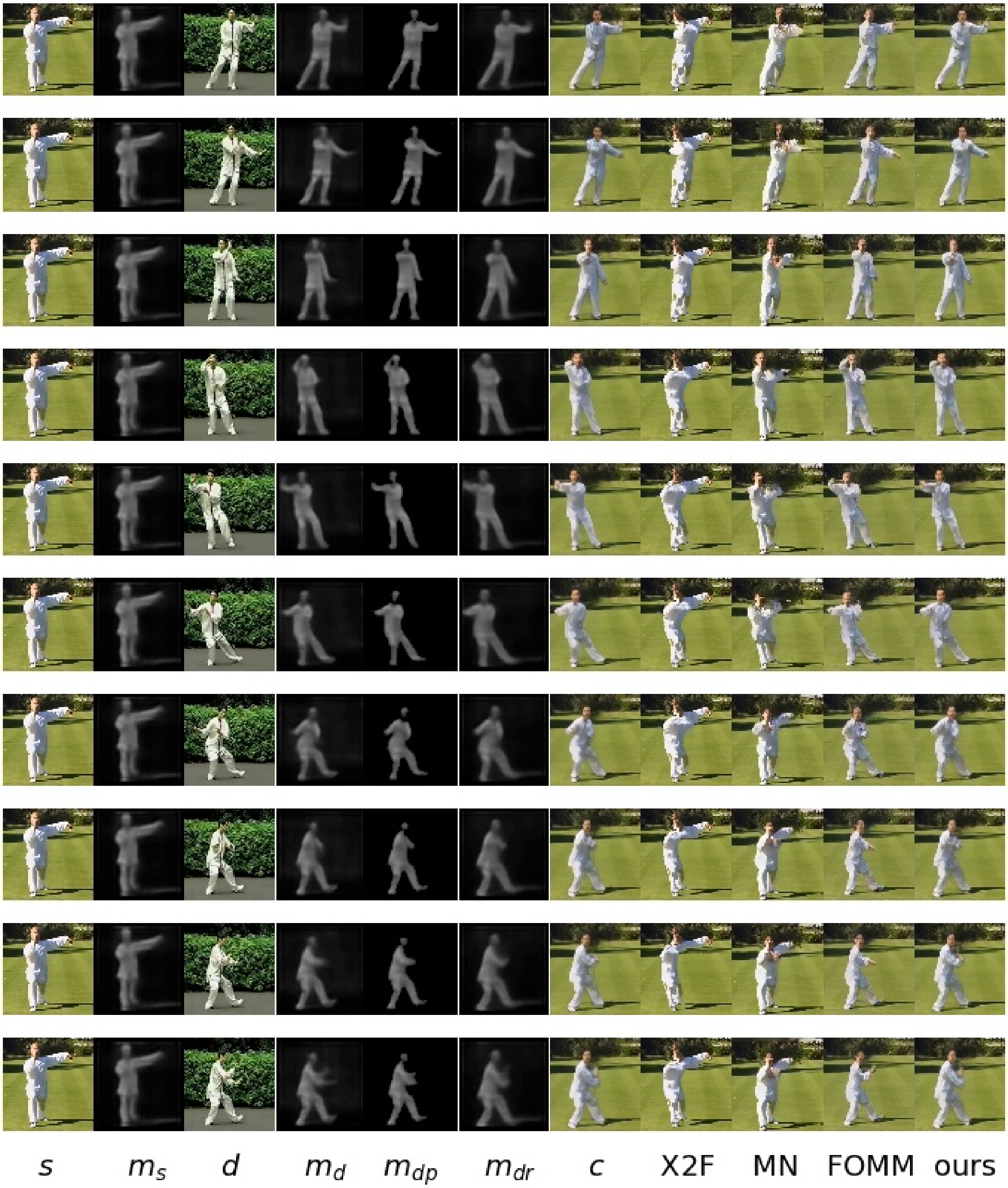
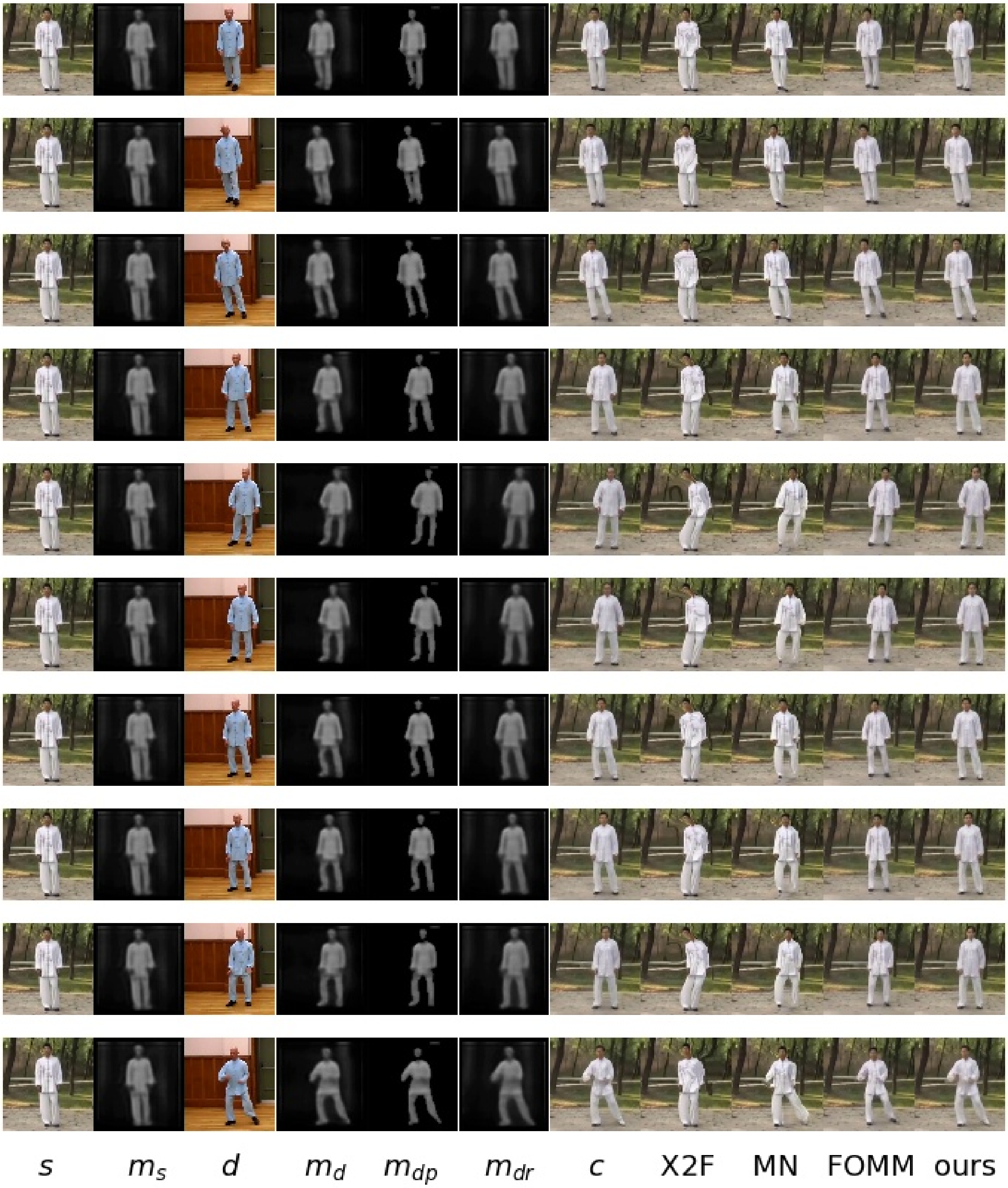

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

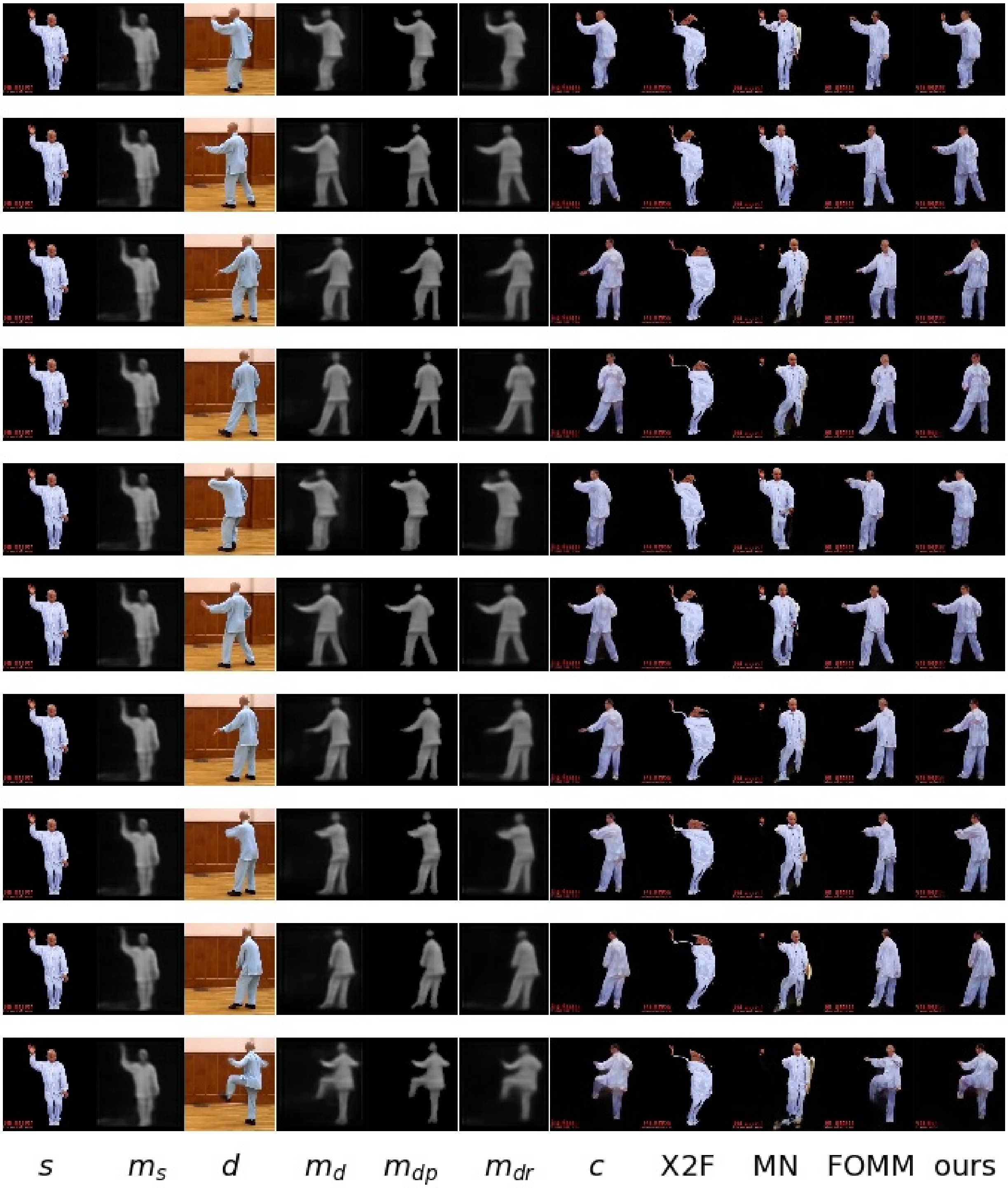
Figure 2. Final and intermediate results generated by our method for Tai-Chi-HD, compared to the SOTA methods.

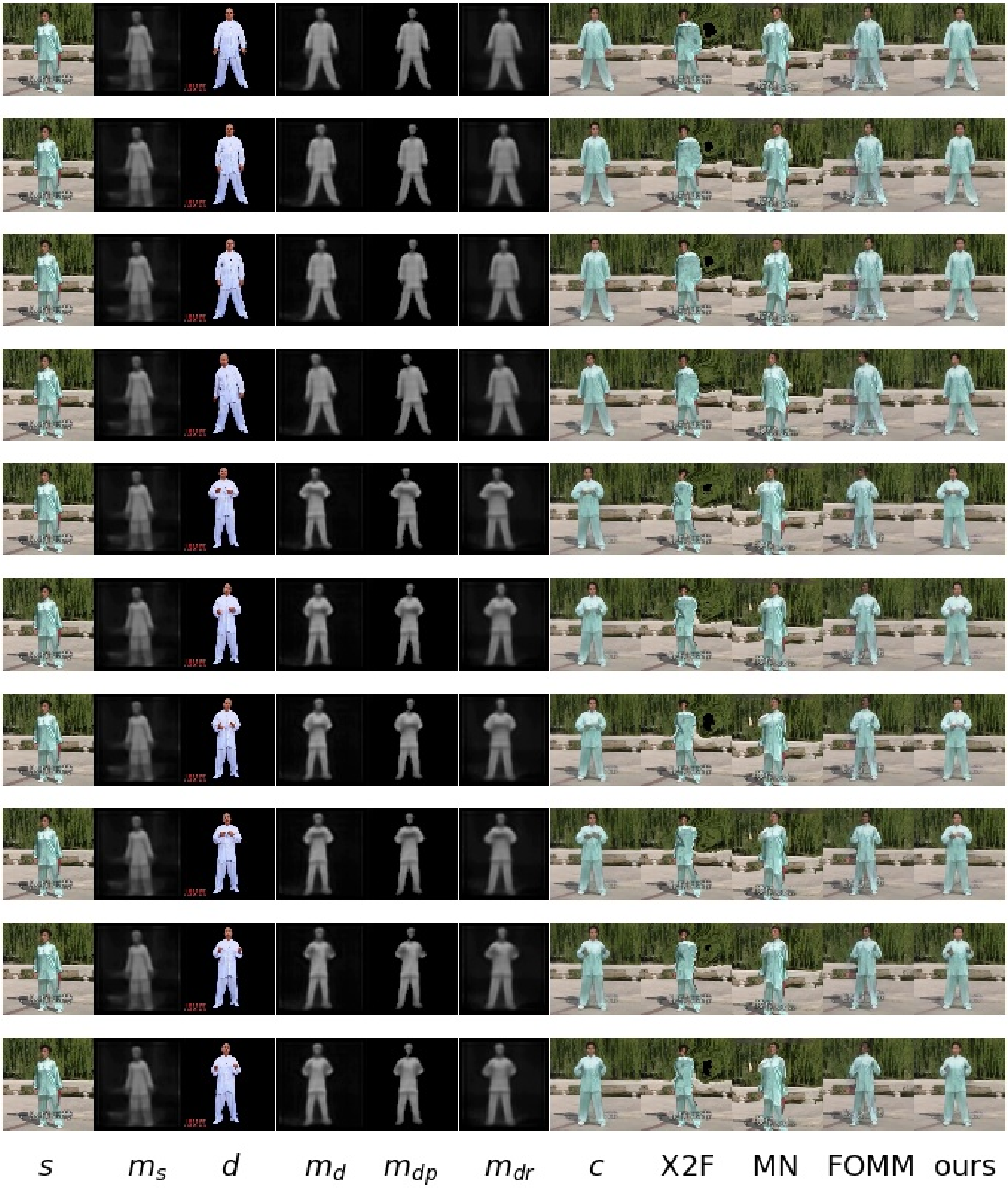


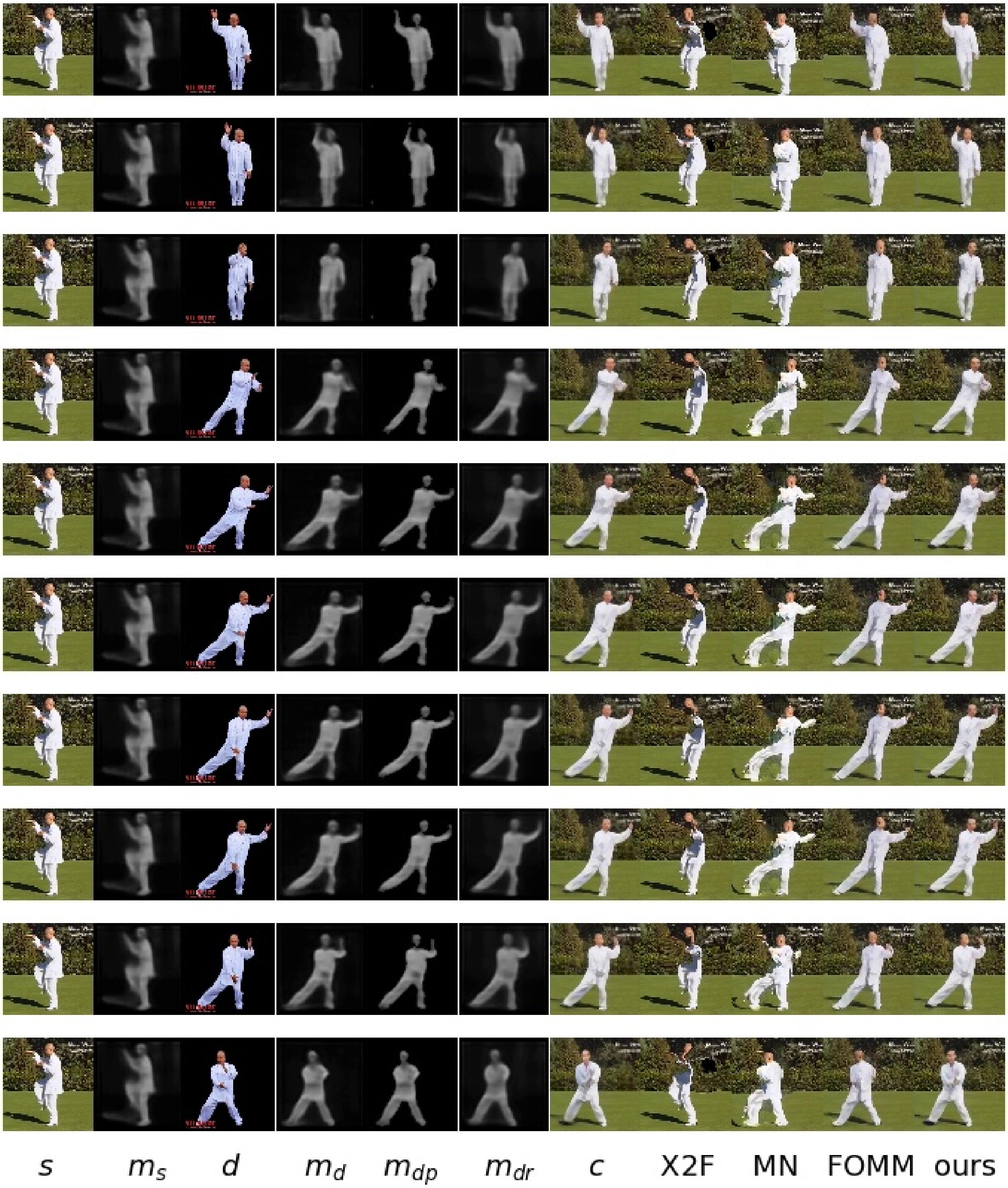

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

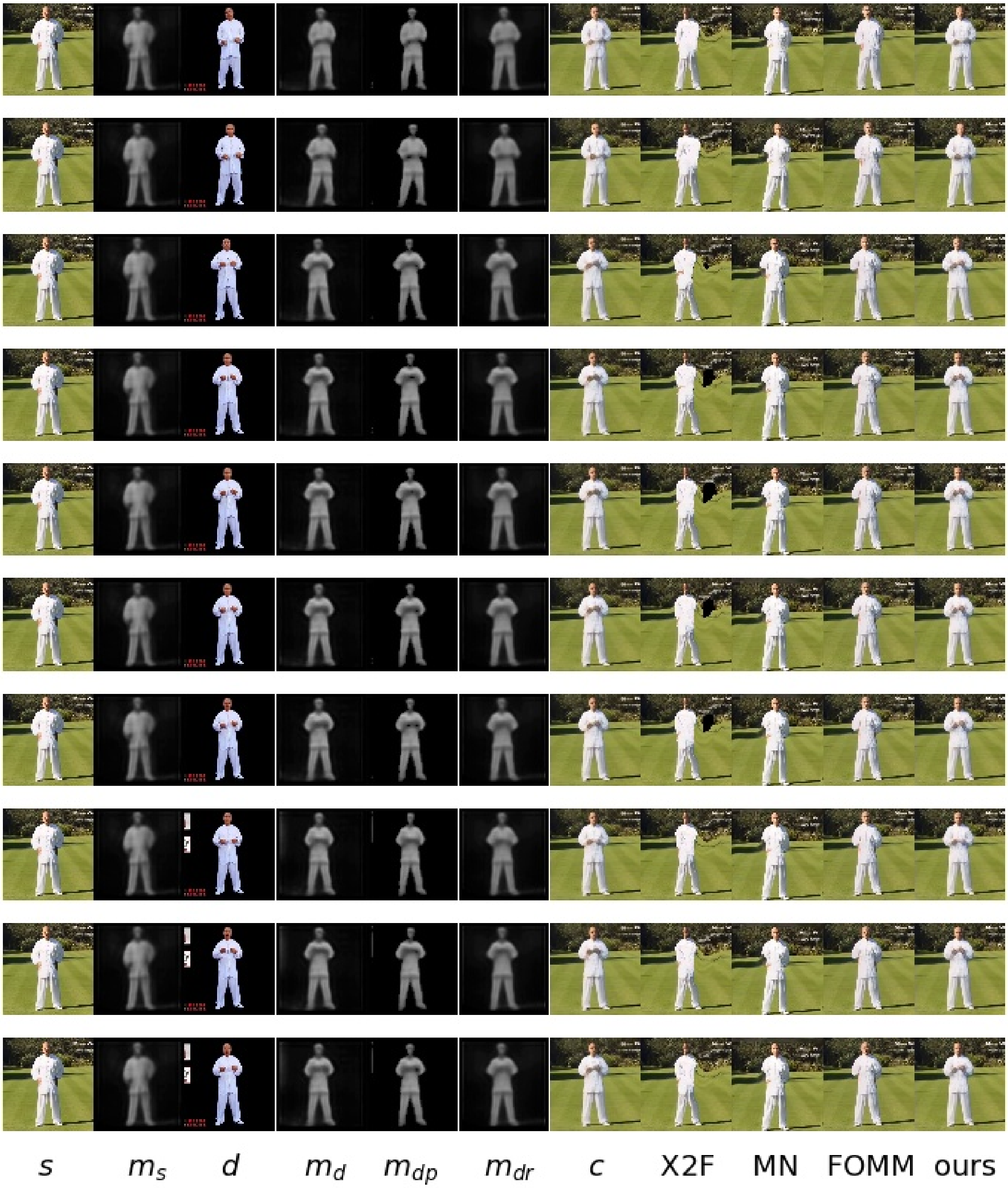

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

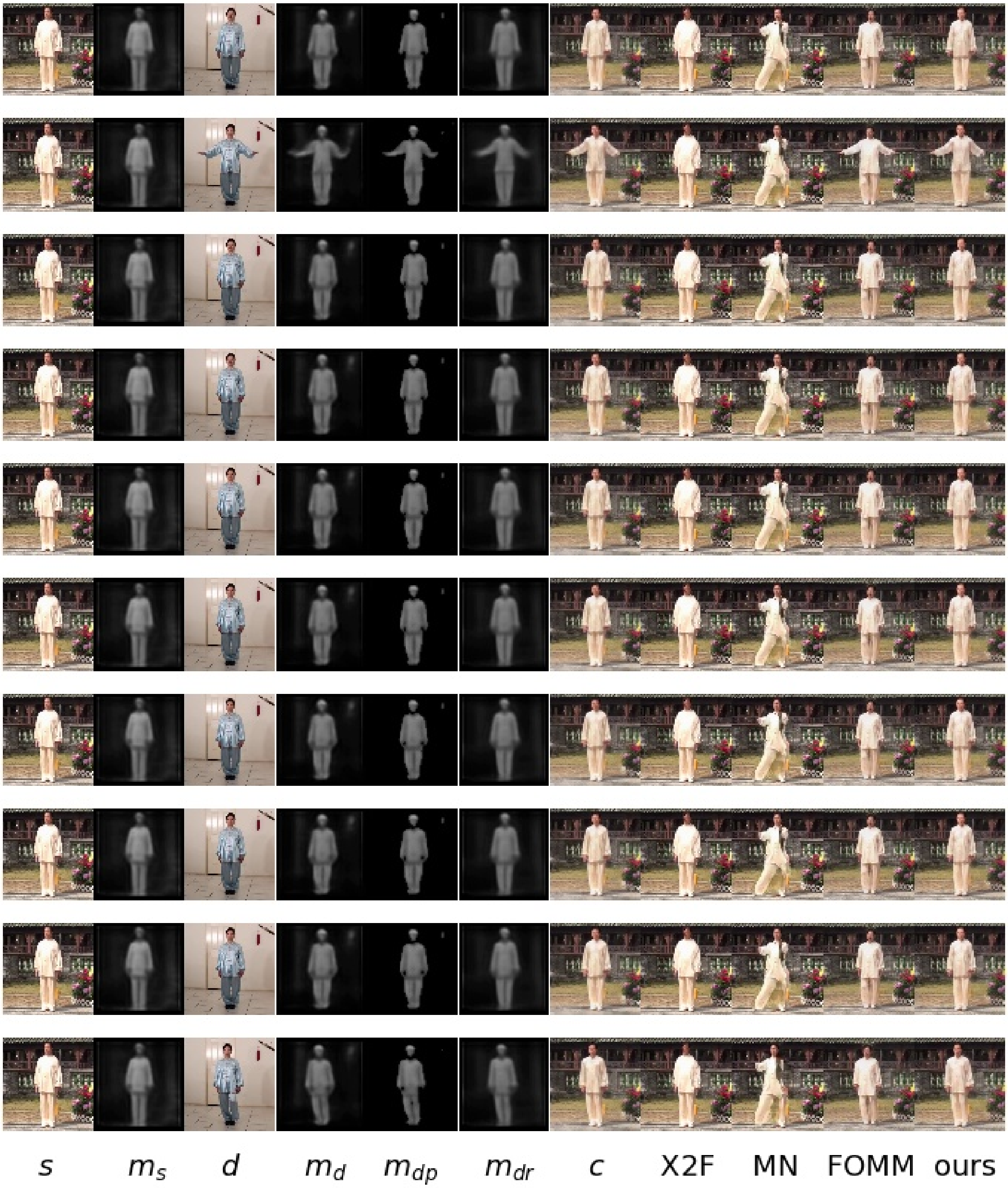

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

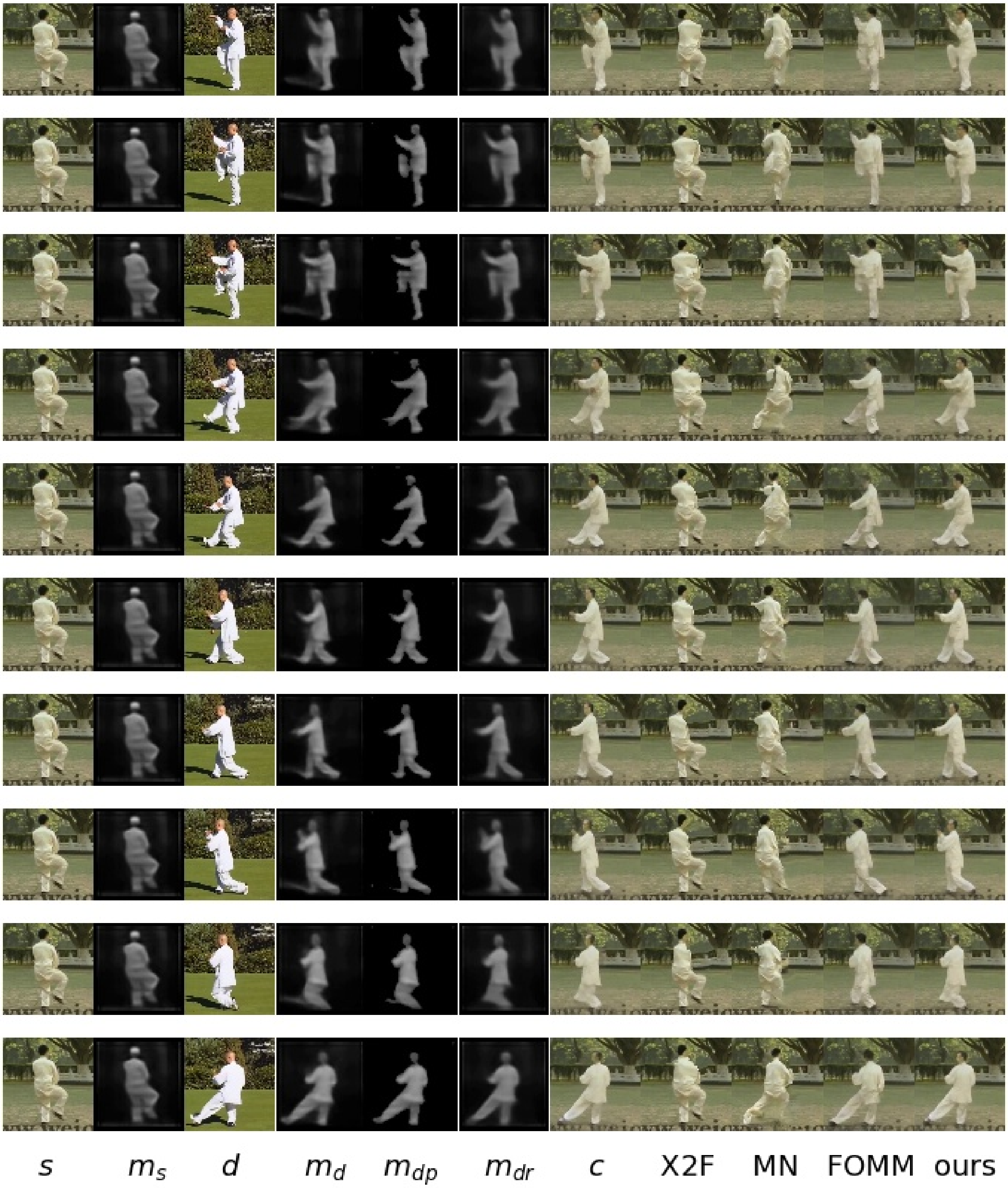

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

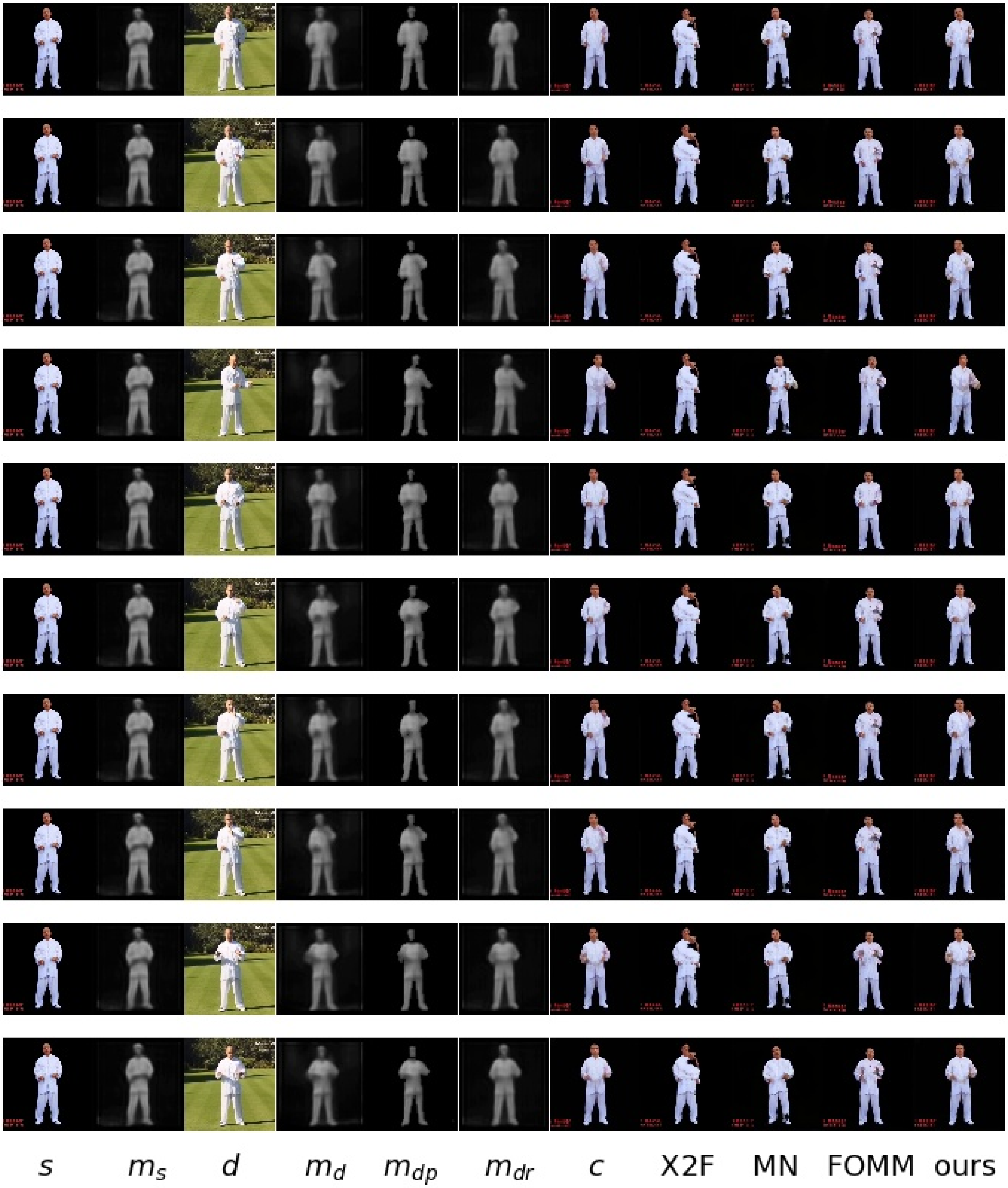
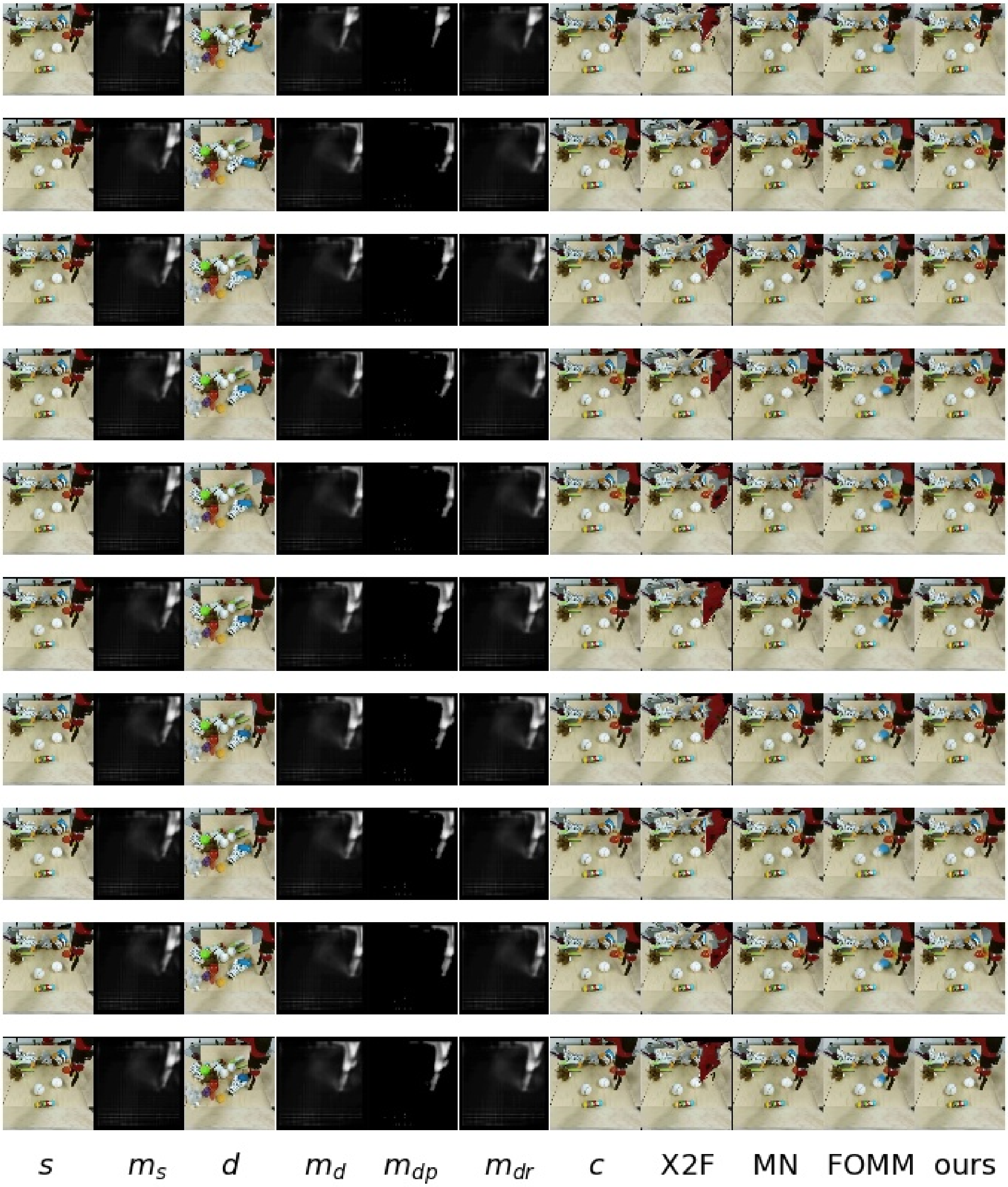
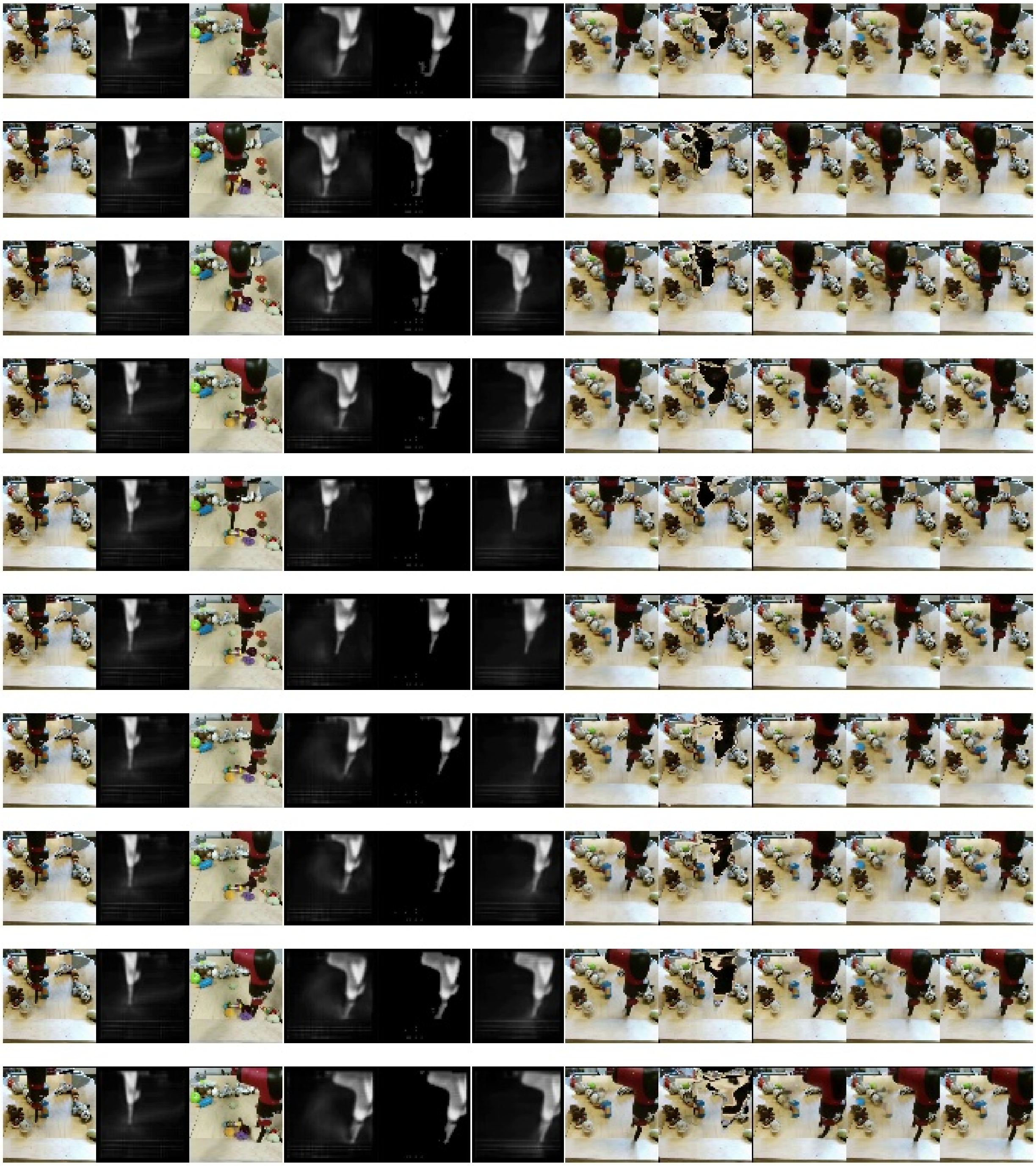
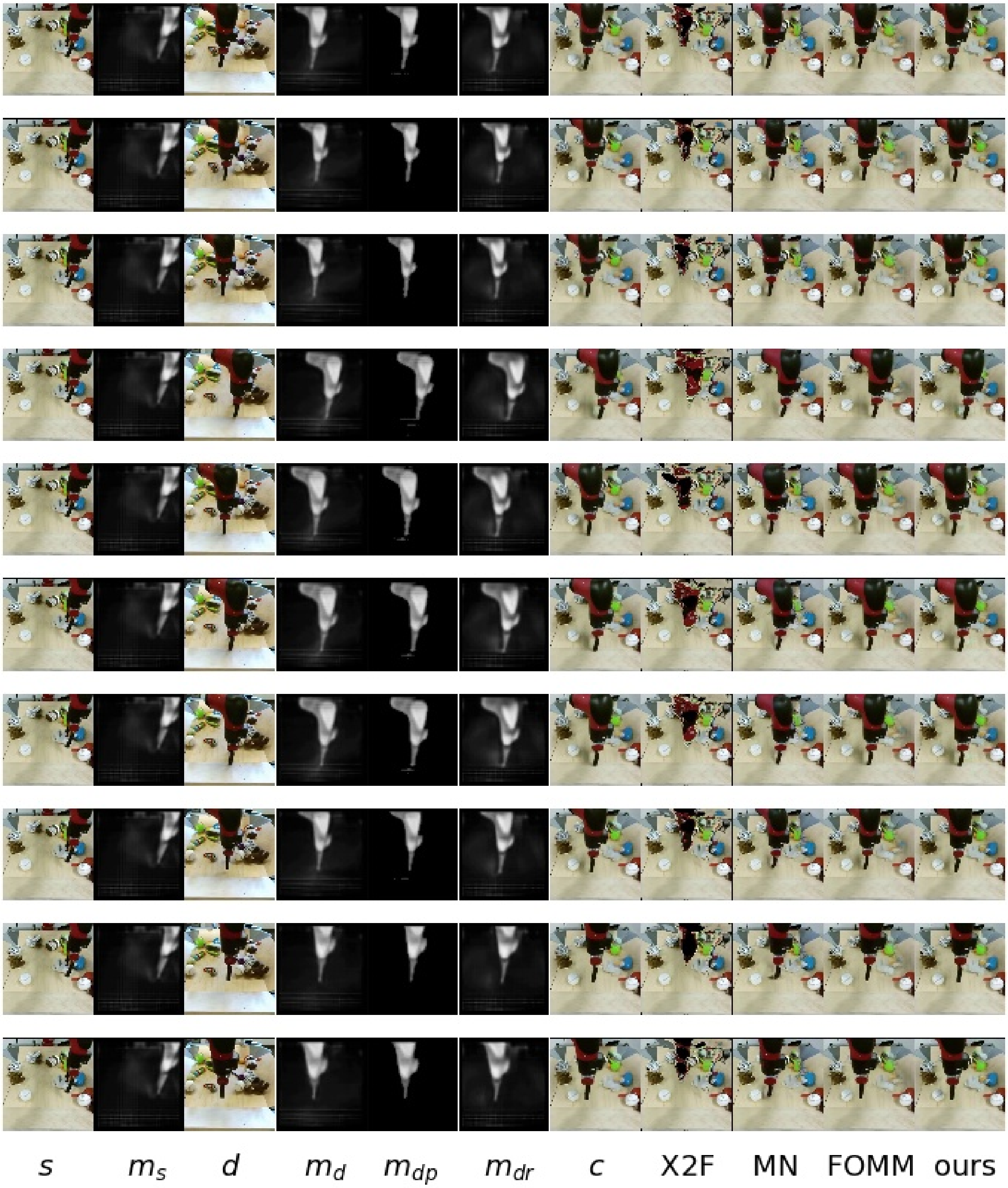

$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours

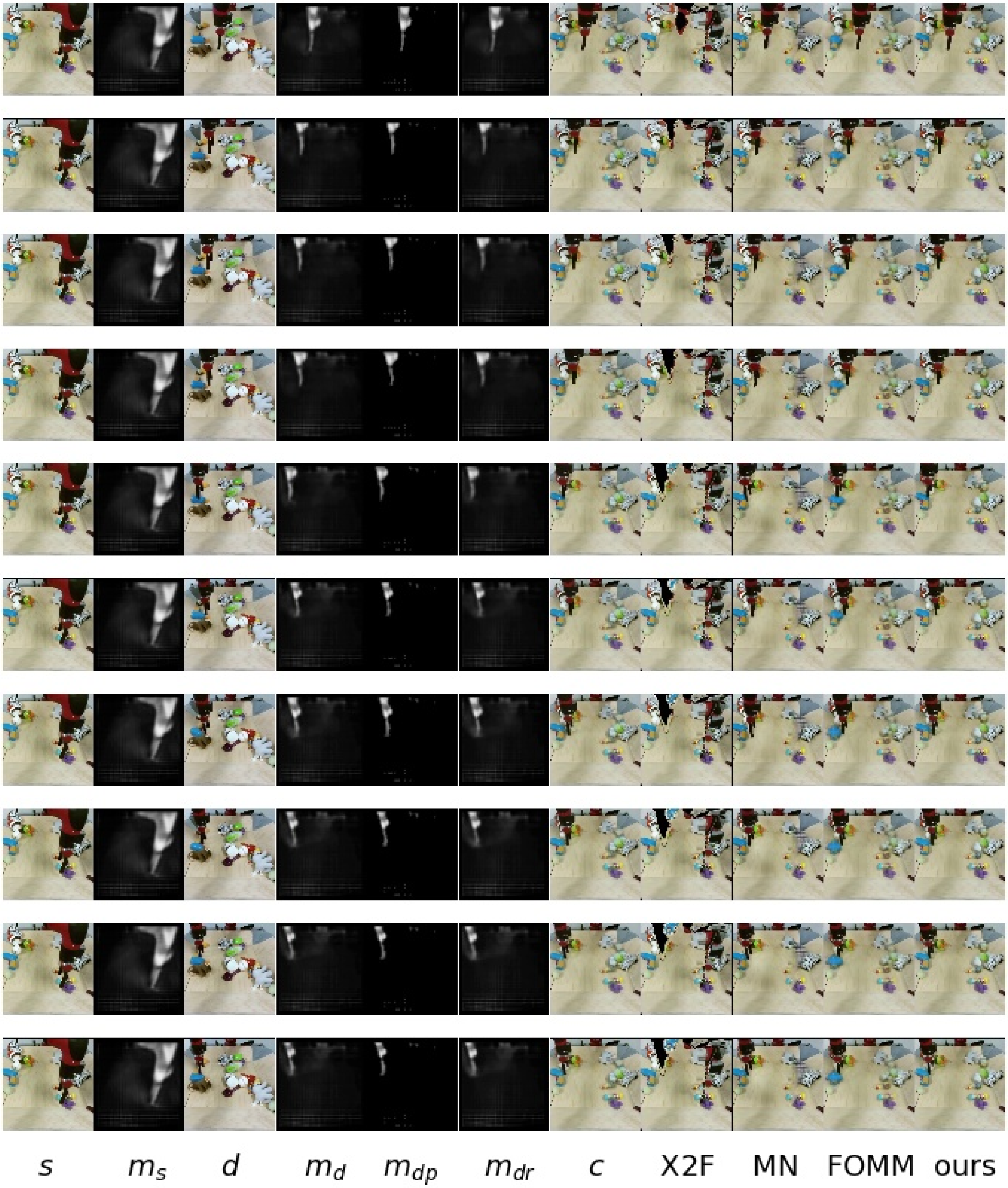
Figure 3. Final and intermediate results generated by our method for BAIR, compared to the SOTA methods.

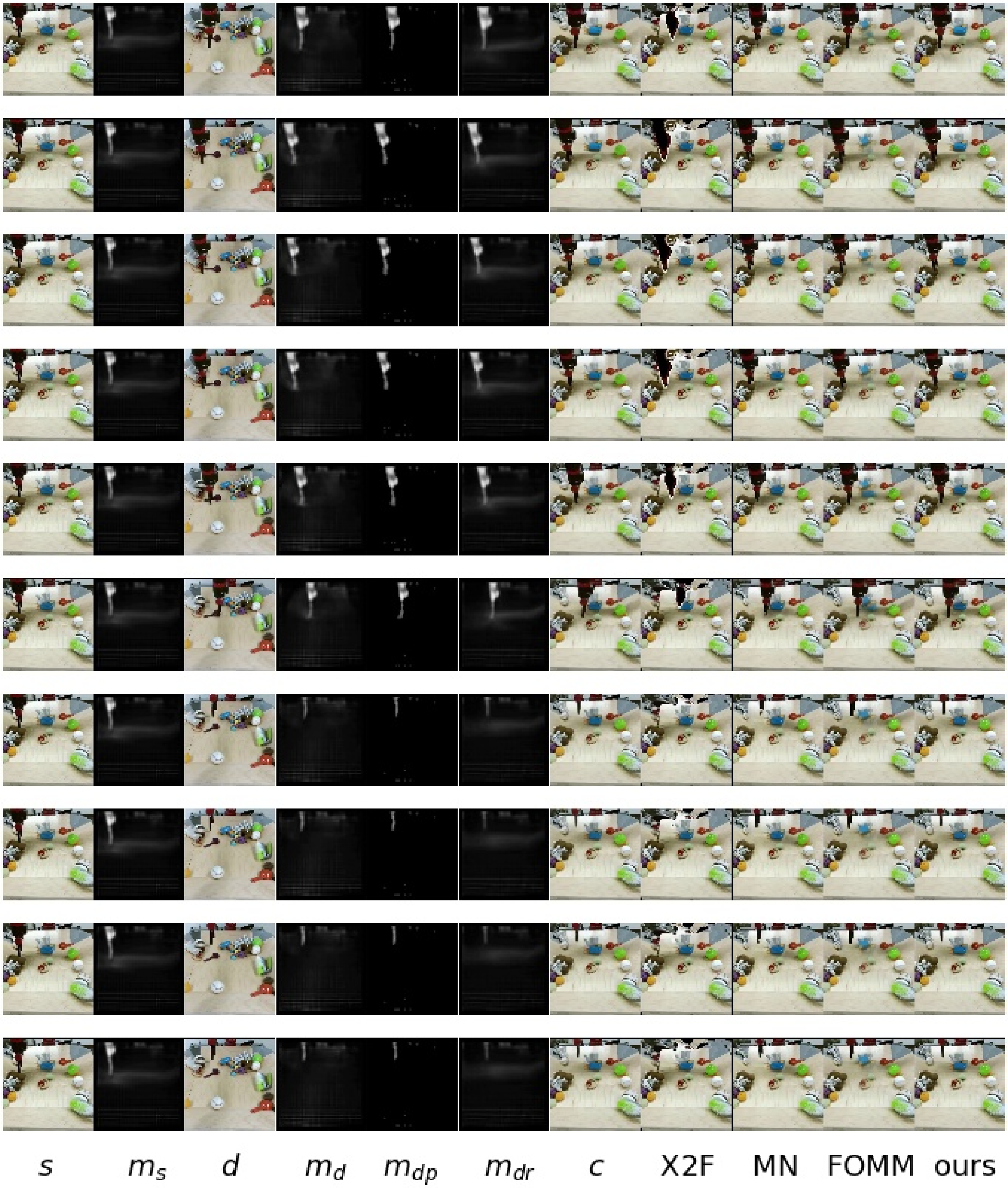


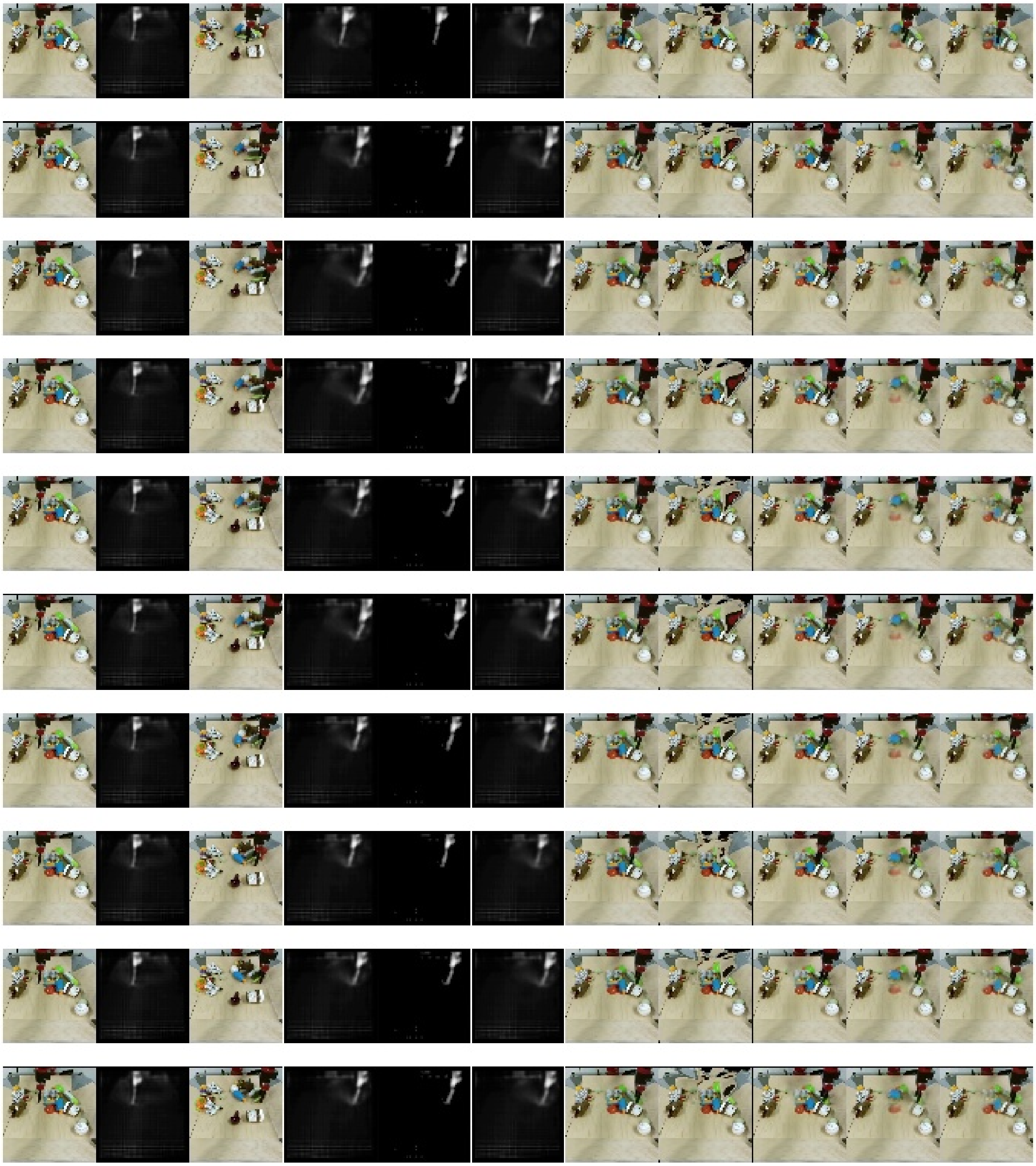


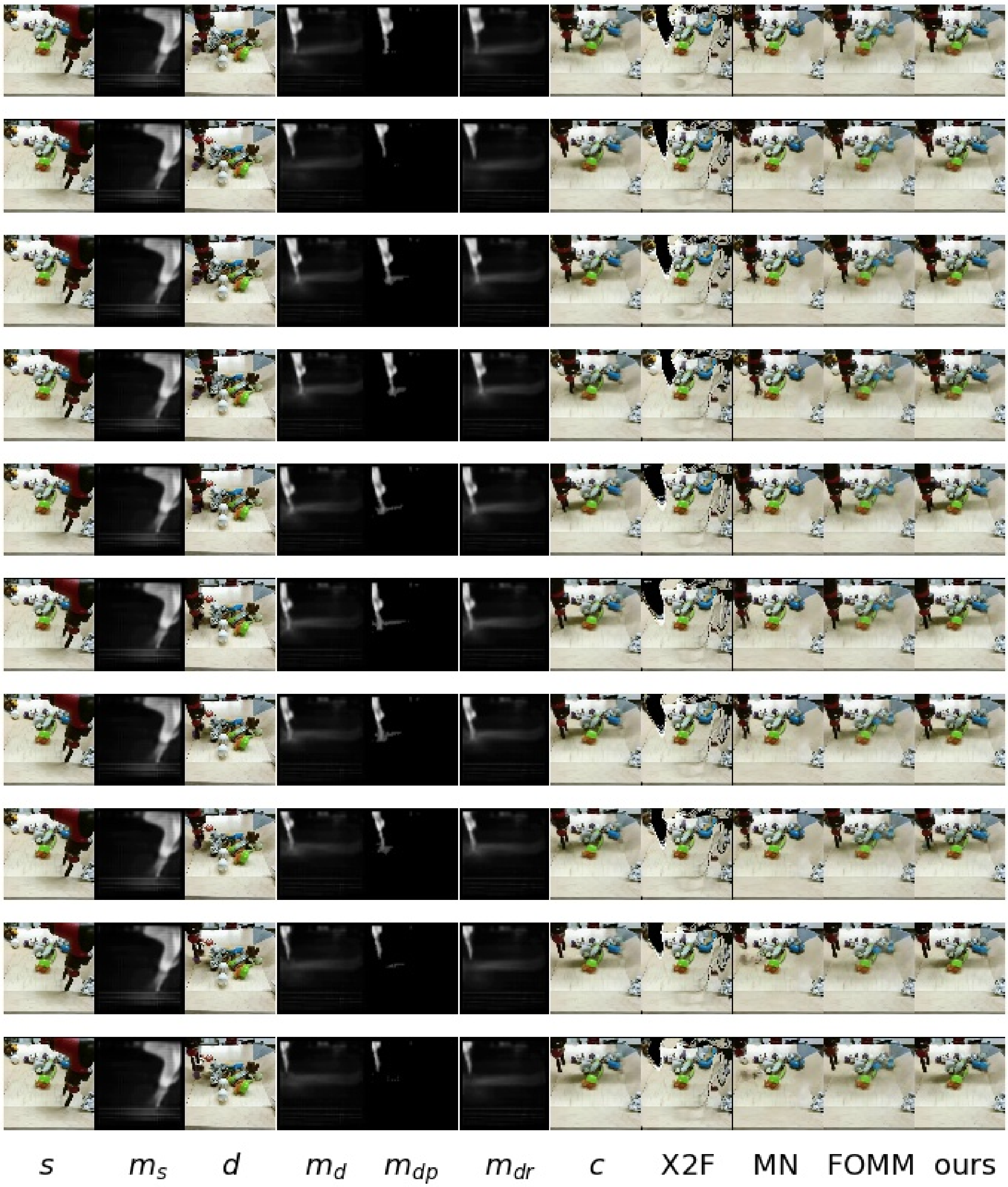
$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours


$s$
$m_s$
$d$
$m_d$
$m_{dp}$
$m_{dr}$
$c$
X2F
MN
FOMM
ours