Supplementary Information

Inverse Design of Thermally Active Composite via Policy-Transferred Reinforcement Learning

Songho Lee[1†], Sukheon Kang[1†], Jisoo Nam[1], Jecheon Yu[1], Miso Kim[1], Seunghwa Ryu[1,2,*]

Affiliations

[1]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[2] KAIST InnoCORE PRISM-AI Center, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[†]These authors contributed equally to this work.

[*]Corresponding author e-mail: ryush@kaist.ac.kr (S. Ryu)

Keywords: Reinforcement learning, Policy transfer, Inverse design, Active composite, 4D Printing

S1. DETAILS OF PROPOSED RL FRAMEWORK

The reinforcement learning (RL)-based inverse design framework proposed in this study consists of a sequential design process in which a designer provides a target trajectory, which is encoded into a target-conditioned state, the agent selects an action for the column at the current step, and the surrogate-based environment returns the corresponding next state and reward (**Figure S1A**). Specifically, the present problem was formulated as a Markov decision process (MDP) by decomposing the 4 × 24 binary Thermally Active Composite (TAC) layout into 24 column-wise decisions, and the initial state starts from a layout in which all cells, except for the fixed wall, are filled with the passive material. At each step, the agent selects one of the retain/flip combinations for the four layers of the current column, and the environment predicts the deformation trajectory of the updated design to provide the next state and reward.

Here, the state consists of three channels, namely the current design, designed-area mask, and target-shape encoding (**Section 2.3**), and the action consists of 16 discrete choices defined by 4-bit switching of the four layers in the current column (**Section 2.5**). The reward is given as the negative value of the local RMSE between the predicted trajectory and the target trajectory immediately after applying the action (**Section 2.6**), and the environment was implemented as an LSTM surrogate model trained on the column-wise deformation behavior (**Section 2.7**). One episode consists of a column-wise design process of up to 24 steps.

The agent was constructed based on Dueling Double DQN for stable value learning considering long-term rewards (**Figure S1**B). To reduce the variance of Monte Carlo-based episodic updates, 3-step Temporal-Difference return and Prioritized Experience Replay (PER) were used together (**Section 2.4**). In addition, a multi-kernel CNN was used to simultaneously capture single-column features and inter-column correlations, together with adaptive average pooling and fully connected layers, and the final agent architecture adopted dueling value–

advantage heads. The major hyperparameters of the agent are summarized in **Table 1** of the main text.

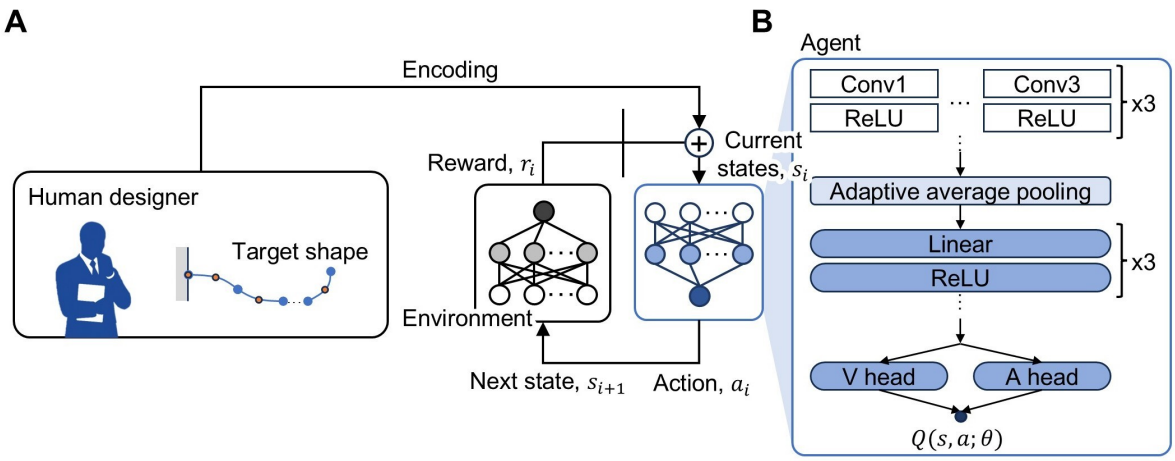


Figure S1 Overview of the proposed RL framework for sequential inverse design of TAC. A Target-conditioned sequential design loop, where the agent selects an action from the current state and the surrogate environment returns the next state and reward. **B** RL agent architecture composed of convolution layers, adaptive average pooling, fully connected layers, and a dueling head for Q-value estimation.

In the single-target setting, the agent was trained for a total of 1,000 episodes while fixing the target shape, with the exploration rate set to $\epsilon = 0.1$ and the mini-batch size set to 16. The other network and optimization settings were kept identical to those presented in **Section 2.4** and **Table 1** of the main text, and training was terminated early when the trajectory RMSE calculated using the greedy policy after each episode reached 0.1 or lower (**Section 3.1**).

In the multiple-target setting, the target shape was replaced at every episode in order to train a generalizable policy over diverse target trajectories, and training was performed for a total of 50,000 episodes. In this case, the mini-batch size was set to 64, and the exploration rate started from 0.5 and reached 0.1 in the later stage by applying an exponential decay of 0.9995. The network architecture, loss computation, replay buffer, and prioritized experience sampling were kept identical to those in the single-target setting, but unlike the single-target case, no

early-termination criterion was applied, and the final trained weights were used for evaluation (**Section 3.2**).

For the RL agent with policy transfer from the multiple-target training, only the initial model weights were initialized with the parameters of the pretrained multiple-target agent, while all other hyperparameters and training procedures were kept identical to those used in the single-target setting. In addition, in the transfer-learning evaluation, a separate set of target trajectories that had not been used in the multiple-target pretraining was used to avoid data leakage (**Section 3.3**). Unless otherwise noted, repeated RL-based results are reported as mean ± standard deviation over five random seeds.

S2. DETAILS OF OPTIMIZATION

This section describes the implementation details, hyperparameters, and objective functions used for the Genetic Algorithm (GA) and Sequential Subdomain Optimization (SSO) [1,2]. All algorithms were evaluated using the same surrogate model (LSTM-based) to ensure consistent comparison conditions.

Identical preprocessing, scaling, and termination criteria were applied to maintain fairness across optimization methods. Because the GA evaluates an entire population per generation, the x-axis in its convergence plots represents the number of samples. In contrast, SSO performs sequential evaluations for each subdomain design. Therefore, its x-axis is expressed in terms of function evaluations (FEs).

S2.1 Genetic algorithm (GA)

In this study, the Genetic Algorithm (GA) treats the grid-based design $X \in \{1,2\}^{(H \times W)}$ as a chromosome, where each individual is represented as a one-dimensional gene sequence of length $H \times W$. The first column is fixed due to the boundary condition (zero padding) and is therefore excluded from crossover and mutation operations. The initial population is generated using a uniform random distribution over $\{1,2\}$.

For each generation, the surrogate LSTM model predicts the output trajectory $\hat{y}$ corresponding to each candidate design X , and the objective function is then evaluated. A rank-based truncation selection scheme is applied to select the top n_{par} individuals as parents. Offspring are produced by performing a single-point crossover at a randomly selected cut position along the length direction, followed by a mutation that flips the material state between $\{1,2\}$ with

probability p_{mut} (excluding the fixed column). The next generation is formed using a generational replacement with elitism strategy that preserves the best-performing individuals.

The optimization terminates when either the maximum generation number is reached or the RMSE of the predicted trajectory satisfies the convergence criterion (RMSE $\leq$ 0.1). The number of function evaluations per generation equals the population size, and in the main text, the x-axis labeled "samples" is interpreted as

$$samples = n_{gen} \times n_{pop} \tag{S1}$$

In this study, two objective functions were used. Let L denote the sequence length, and let $\hat{y}_t, y_t \in R^2$ represent the predicted and target coordinates at time step t respectively.

(1) Unweighted RMSE:

$$RMSE = \sqrt{\frac{1}{L}\sum_{t=1}^{L} \| \hat{y}_t - y_t \|_2^2} \tag{S2}$$

(2) Distance-weighted RMSE:

$$wRMSE = \sqrt{\frac{1}{L}\sum_{t=1}^{L} \frac{\| \hat{y}_t - y_t \|_2^2}{t}} \tag{S3}$$

Additionally, a population-size sensitivity analysis was conducted to examine the effect of the GA population size on convergence behavior (**Figure S2**). Specifically, GA inverse design was repeatedly performed for a target with a nearly monotonic trajectory (**Figure S2A**) and for a target with a more curved trajectory (**Figure S2D**) , while varying the population size as 10, 30, 50, 70, and 90. For each case, the mean and standard deviation of the terminal sample count over 10 random seeds were compared. Here, the terminal sample count denotes the cumulative number of samples required to reach the termination criterion.

For both representative targets, smaller population sizes generally resulted in lower terminal sample counts, and this trend became more pronounced for the more complex target trajectory, where the required number of samples increased as the population size increased. Based on these results, the population size was set to 30 in this study. The remaining GA hyperparameter settings were selected with reference to those used in related prior studies[1,3], and the final settings are summarized in **Table S1**.

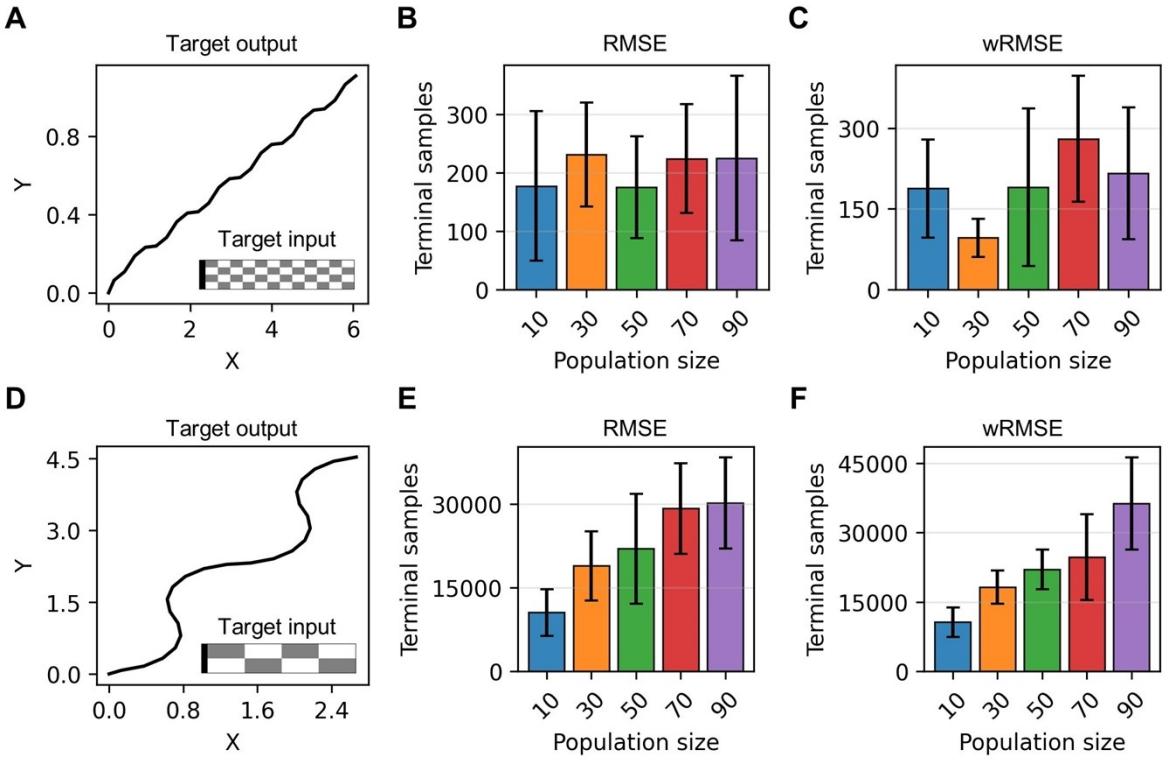


Figure S2 Population size sensitivity analysis of GA for two representative targets under RMSE- and wRMSE-based objectives. A Representative target trajectory with a nearly monotonic shape and its corresponding target input. **B** Terminal sample counts obtained by the RMSE-based GA for different population sizes. **C** Terminal sample counts obtained by the wRMSE-based GA for different population sizes. **D** Representative target trajectory with a more curved and complex shape and its corresponding target input. **E** Terminal sample counts obtained by the RMSE-based GA for different population sizes. **F** Terminal sample counts obtained by the wRMSE-based GA for different population sizes.

Table S1 Hyperparameters of GA

Item	Value
Population (n_{pop})	30
Generations (n_{gen})	10,000
Parents (n_{par})	15
Offsprings	15
Crossover	1-point cross over
Mutation probability (p_{mut})	0.15

S2.2 Sequential subdomain optimization (SSO)

The Sequential Subdomain Optimization (SSO) method performs a progressive subdomain design along the overall length W. Given a subdomain length $N_{sub} \in \{1,2,3\}$, a window size N_{window} is defined at each step as

$$N_{window} = min(N_{sub}, W - step) \tag{S4}$$

and all possible candidate blocks $\{1,2\}^{H \times N_{window}}$ are explored through exhaustive search. Each candidate block is temporarily inserted into the current design X_{curr}, after which the surrogate LSTM model predicts the corresponding trajectory $\hat{y}$. Both the local RMSE, Eq. (S5), and the global RMSE, Eq. (S6), are then computed. The selection prioritizes the candidate with the lowest local RMSE, and in the case of a tie, the one with the smaller global RMSE is chosen. From the selected optimal block, only the leftmost column is committed to the design, and the step index is

advanced by one ($step \leftarrow step + 1$). This process repeats until the entire length is designed or the global RMSE reaches the convergence threshold (≤ 0.1). The initial design is filled entirely with material '2' except for the padded column, which is fixed and excluded from the search. The number of function evaluations (FEs) corresponds to the number of surrogate model calls and is used as the x-axis in the figures in the main text.

In this implementation, the local RMSE is evaluated within the current window N_{window}. $\hat{y}_\tau, y_\tau \in R^2$ represent the predicted and target coordinates at time step t respectively.

$$RMSE_{\text{local}}(N_{window}) = \sqrt{\frac{1}{N_{window}} \sum_{\tau=t}^{t+N_{window}-1} \| \hat{y}_\tau - y_\tau \|_2^2} \tag{S5}$$

The global RMSE is computed over the entire sequence using Eq. (S6). In case of a tie, it chooses the one with the smaller global RMSE. L denotes the sequence length.

$$RMSE_{global} = \sqrt{\frac{1}{L} \sum_{\tau=1}^{L} \| \hat{y}_\tau - y_\tau \|_2^2} \tag{S6}$$

S2.3 Metrics for computational cost

To compare the computational cost in a fair and reproducible manner, RL, GA, and SSO were all evaluated under the same surrogate-based environment. Specifically, the same trained LSTM surrogate model, the same preprocessing, the same coordinate scaling, the same target representation, and the same global RMSE-based termination criterion were applied to all methods. Therefore, the differences in computational cost reported in this study can be interpreted as arising

from differences in the optimization strategy of each method, rather than from differences in the forward model or evaluation conditions.

Because the optimization structure differs across methods, the present study defined the computational cost metrics, namely "sample" and "function evaluation (FE)," according to the characteristics of each method. For RL, one sample corresponds to one completed episode. This is because one full material layout is progressively constructed over an episode through a sequential decision process and is ultimately evaluated as one inverse design result at the episode level. Meanwhile, the FE in RL was defined as the cumulative number of step-level evaluations within episodes. That is, as described in the main text, in RL, one FE corresponds to one step within an episode, and the cumulative FE is calculated as the cumulative number of steps over episodes.

For GA, one sample corresponds to one evaluated individual in the population. Because GA evaluates the entire population at each generation, the cumulative sample count is defined as the product of the number of generations and the population size. In addition, each individual corresponds to one complete design candidate, and its performance is evaluated through a full-trajectory prediction using the surrogate model. Therefore, in GA, one sample can be regarded as corresponding to one surrogate-based full-design evaluation. The comparison between RL and GA in the main text was made on the basis of sample count, since one episode in RL and one individual evaluation in GA both generate and evaluate one complete candidate design.

For SSO, the optimization is performed by sequentially exploring candidate patterns within a local subdomain window in a stepwise manner. At each step, a candidate block is temporarily inserted into the current design, and the resulting trajectory is predicted using the surrogate model. Therefore, in SSO, one FE is directly defined as one surrogate model call for one candidate subdomain state. Since SSO is inherently a sequential stepwise optimizer, the computational cost of this method was represented as the cumulative number of FEs.

For all methods, surrogate inference was included in the FE accounting because it served as the common forward evaluation. In contrast, the learning update or optimizer update itself was not counted as an FE because it does not correspond to a design evaluation. Therefore, the sample and FE metrics used in this study quantify the number of design evaluations required by each optimization method under a common surrogate environment.

S3. ADDITIONAL SINGLE-TARGET RESULTS ON RANDOM AND HAND-DRAWN TARGETS

This section presents the single-target inverse design results for both random and hand-drawn target trajectories. All algorithms used the same surrogate model (LSTM), preprocessing procedure, and termination criterion (RMSE $\leq$ 0.1) as described in the main text. For the GA, the internal objective function was defined as either RMSE or the weighted root mean squared error (wRMSE), while for the SSO, it was defined as the local window-averaged RMSE. In the convergence plots, the x-axis represents Samples = generations × population size for GA and FEs = number of surrogate evaluations for SSO. (All reported values are given as the mean ± standard deviation over five random seeds.)

S3.1 Inverse design for randomly generated input patterns

Output trajectories were first generated from randomly created input patterns using the surrogate LSTM model, and each trajectory was then set as the target output for inverse design. The results are summarized in **Figure S3** and **Table S2**. In **Figure S3C** and **Figure S3E**, the designed output trajectories are directly compared with the target trajectory, and the insets in **Figure S3B** and **Figure S3D** more clearly highlight the early-stage convergence behavior. In particular, for the SSO results in **Figure S3D**, the solid and faint lines denote the local RMSE used for stepwise selection and the global RMSE at each step, respectively. In **Table S2**, RL (TL) achieved RMSE 0.076 ± 0.012, Samples 17.4 ± 13.012, and FEs 417.6 ± 312.277, yielding lower error and requiring fewer samples and evaluations than RL (0.088 ± 0.010, 154.2 ± 29.312, 3700.8 ± 703.491). GA (wRMSE) and GA (RMSE) also converged stably, but the required number of samples was larger than for the RL family, at 1,050 $\pm$ 637.103 and 822 $\pm$ 326.986, respectively.

For SSO, the final RMSE decreased as Nsub increased, whereas the number of function evaluations grew. Specifically, SSO (Nsub=1) attained RMSE 0.019 with 384 FEs, exhibiting the lowest computational cost. SSO (Nsub=2) and SSO (Nsub=3) each achieved RMSE 0.003, but at 5,904 and 90,384 FEs, respectively. Under this setting, the fewest FEs were required by SSO (Nsub=1), followed by RL (TL), RL, SSO (Nsub=2), and SSO (Nsub=3).

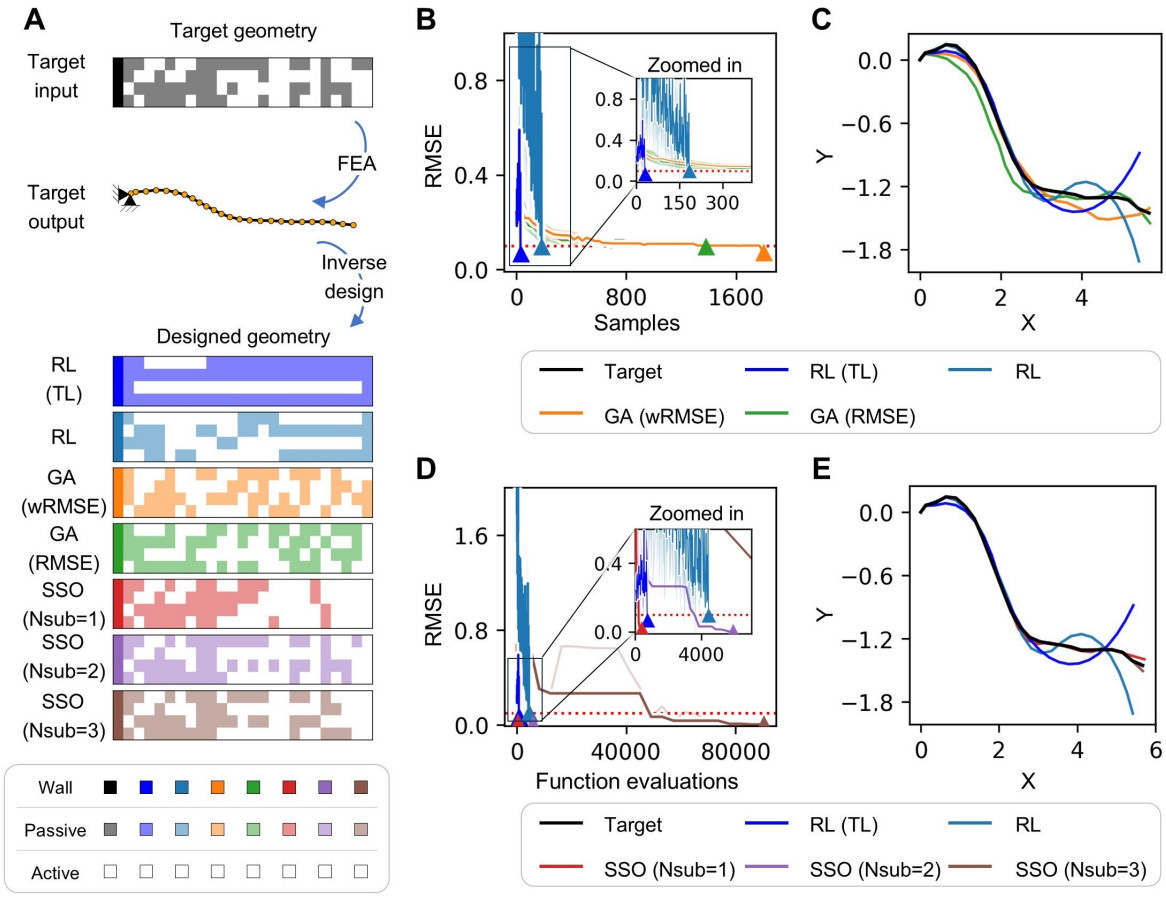


Figure S3 Inverse design on a randomly generated single target: RL (TL), RL, GA (wRMSE/RMSE), and SSO (Nsub=1-3). A Target input-output and designed inputs. **B** RMSE versus samples for RL (TL), RL, GA (wRMSE), and GA (RMSE). The red dashed line indicates the termination threshold. The inset shows the early-stage convergence. Solid lines represent the mean over five random seeds for RL (TL), RL, GA (wRMSE), and GA (RMSE), and the shaded regions represent the corresponding standard deviations. **C** Output trajectories generated from the designs of RL (TL), RL, GA (wRMSE), and GA (RMSE), compared with the target trajectory. **D** RMSE versus function evaluations for RL (TL), RL, and SSO (Nsub=1-3). The red dashed line indicates the termination threshold. The inset shows the early-stage convergence. For RL (TL) and RL, solid lines indicate the mean over five random seeds and the shaded regions indicate the corresponding standard deviations over five random seeds,

respectively, whereas for SSO (Nsub=1-3), the solid lines indicate the local RMSE used for stepwise selection and the faint lines indicate the global RMSE at each step. **E** Output trajectories generated from the designs of RL (TL), RL, and SSO (Nsub=1-3), compared with the target trajectory.

Table S2 Quantitative comparison on a randomly generated single target: RL (TL), RL, GA (wRMSE/RMSE), and SSO (Nsub=1–3). Values are reported as mean ± standard deviation over five random seeds. If averaging is not applicable, a single raw value is shown.

	RMSE	Samples	FEs
RL (TL)	0.076 ± 0.012	17.4 ± 13.012	417.6 ± 312.277
RL	0.088 ± 0.01	154.2 ± 29.312	3700.8 ± 703.491
GA (wRMSE)	0.082 ± 0.012	1,050 ± 637.103	-
GA (RMSE)	0.091 ± 0.01	822 ± 326.986	-
SSO (Nsub=1)	0.019	-	384
SSO (Nsub=2)	0.003	-	5,904
SSO (Nsub=3)	0.003	-	90,384

S3.2 Inverse Design on Hand-Drawn Target Patterns

In this section, the hand-drawn targets were generated as arbitrary trajectories from user-defined parametric functions and were therefore treated as unseen cases without pre-specified target inputs. The inverse design results for hand-drawn targets are presented in **Figure S4** and **Table S3**. In **Figure S4C** and **Figure S4E**, the output trajectories obtained from each inverse-design method are directly compared with the target trajectory. In **Figure S4D**, the SSO curves are presented using the local RMSE for stepwise selection together with the corresponding global RMSE at each step. In **Table S3**, RL (TL) reached the termination threshold with RMSE 0.091 ± 0.006, Samples 130.4 ± 64.551, and FEs 3,129.6 ± 1,549.218. Compared with RL (0.095 ± 0.004, 375.4 ± 85.845 samples, 9,009.6 ± 2,060.271 FEs), a similar error level was achieved while requiring fewer samples and evaluations. GA (wRMSE) and GA (RMSE) also

converged. However, the sample counts were higher than those of the RL family, at 750 ± 190.919 and 846 ± 424.476, respectively.

For SSO, accuracy improved as Nsub increased, but computational cost grew sharply. SSO (Nsub=2) and SSO (Nsub=3) achieved lower errors (RMSE 0.070 and 0.066), yet required 5,904 and 90,384 function evaluations, respectively. By contrast, SSO (Nsub=1) yielded RMSE 0.120 and did not satisfy the termination criterion (0.1). Taken together, these results indicate that, even on unseen hand-drawn targets, RL (TL) satisfied the termination criterion with fewer samples and function evaluations than RL, while maintaining competitive accuracy. This suggests that the benefits of multi-target pretraining were effectively transferred to the single-target setting.

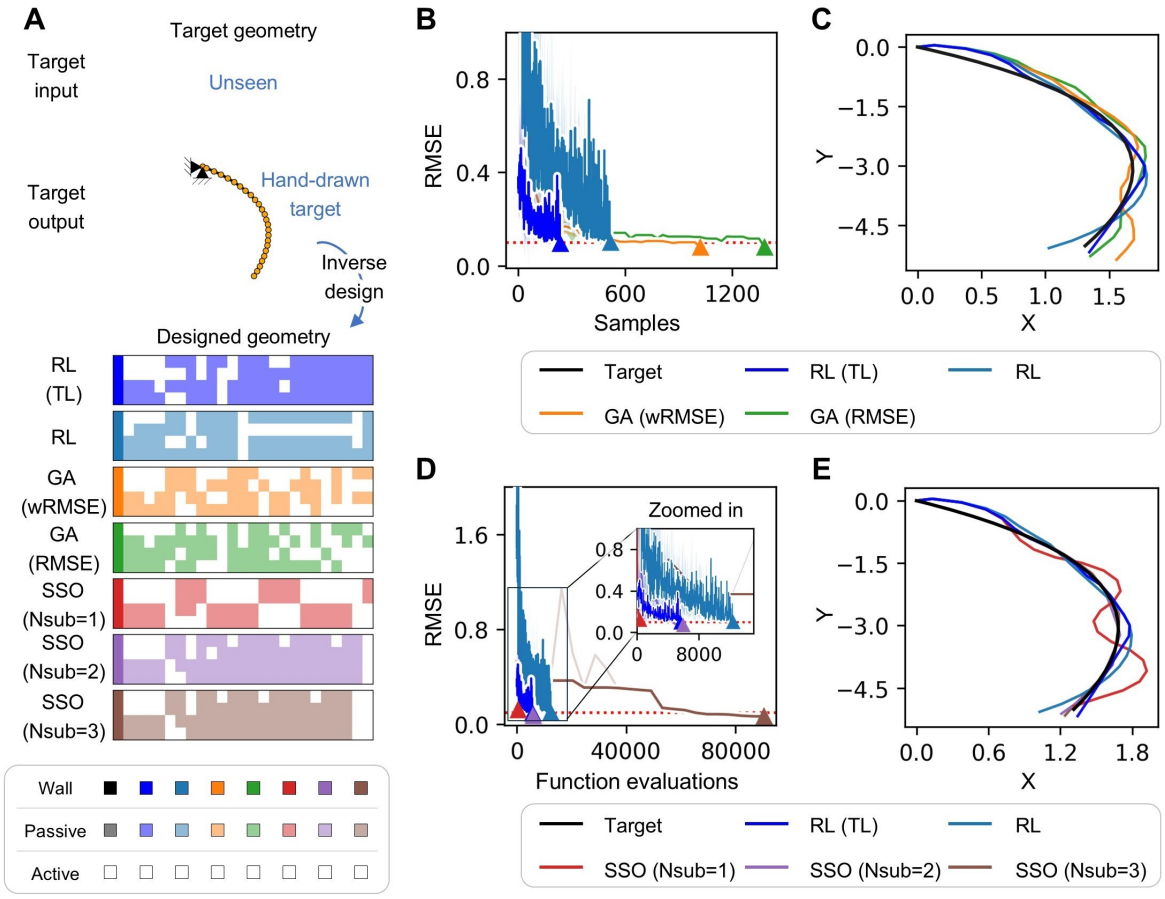


Figure S4 Inverse design on an unseen hand-drawn single target: RL (TL), RL, GA (wRMSE/RMSE), and SSO (Nsub=1-3). A Target input-output and designed inputs. **B**

RMSE versus samples for RL (TL), RL, GA (wRMSE), and GA (RMSE). The red dashed line indicates the termination threshold. Solid lines represent the mean over five random seeds for RL (TL), RL, GA (wRMSE), and GA (RMSE), and the shaded regions represent the corresponding standard deviations. **C** Output trajectories generated from the designs of RL (TL), RL, GA (wRMSE), and GA (RMSE), compared with the target trajectory. **D** RMSE versus function evaluations for RL (TL), RL, and SSO (Nsub=1-3). The red dashed line indicates the termination threshold. The inset shows the early-stage convergence. For RL (TL) and RL, solid lines indicate the mean over five random seeds and the shaded regions indicate the corresponding standard deviations over five random seeds, respectively, whereas for SSO (Nsub=1-3), the solid lines indicate the local RMSE used for stepwise selection and the faint lines indicate the global RMSE at each step. **E** Output trajectories generated from the designs of RL (TL), RL, and SSO (Nsub=1-3), compared with the target trajectory.

Table S3 Quantitative comparison on an unseen hand-drawn single target: RL (TL), RL, GA (wRMSE/RMSE), and SSO (Nsub=1-3). Values are reported as mean ± standard deviation over five random seeds. If averaging is not applicable, a single raw value is shown.

	RMSE	Samples	FEs
RL (TL)	0.091 ± 0.006	130.4 ± 64.551	3,129.6 ± 1,549.218
RL	0.095 ± 0.004	375.4 ± 85.845	9,009.6 ± 2,060.271
GA (wRMSE)	0.092 ± 0.006	750 ± 190.919	-
GA (RMSE)	0.081 ± 0.007	846 ± 424.476	-
SSO (Nsub=1)	0.120	-	384
SSO (Nsub=2)	0.070	-	5,904
SSO (Nsub=3)	0.066	-	90,384

S4. 4D PRINTING STRATEGY AND EXPERIMENTAL VALIDATION

S4.1 Resin Material Selection and Formulation

In this study, an actuation strategy based on volatilization-induced shrinkage during thermal treatment and the resulting localized strain mismatch in patterned multi-phase structures was employed. A highly volatile monomer, isobornyl acrylate (IBOA), and a long-chain urethane acrylate oligomer, aliphatic urethane diacrylate (AUD, Ebecryl 8402), were selected as the constituent materials. IBOA exhibits negligible volatilization at room temperature. However, it rapidly volatilizes during thermal treatment at 80 °C, leading to volumetric shrinkage of the structure. In contrast, AUD consists of chemically stable, flexible polyurethane chains with a long molecular backbone. As a result, AUD remains stable under thermal treatment while forming an entangled polymer network that allows the preferential volatilization of IBOA.

IBOA and AUD were first mixed at various weight ratios ranging from 40:60 to 80:20. Based on the total mass of the resulting IBOA/AUD mixture, 2,4,6-trimethylbenzoyl-diphenylphosphine oxide (TPO, 1 wt%) and Sudan I (0.1 wt%) were subsequently added as a photoinitiator and a photoabsorber, respectively. The detailed weights of each formulation are summarized in Table S4, and a photograph of the prepared resin is shown in **Figure S5A**.

For the nine resin formulations, the rheological properties of the liquid-state resins were evaluated to assess their suitability for DLP 3D printing. As shown in **Figure S5B-i**, all formulations exhibit pronounced shear-thinning behavior, characterized by a decrease in viscosity with increasing shear rate, which is favorable for DLP printing processes. The viscosities measured at a shear rate of 30 s^{-1} are compared in **Figure S5B-ii**. A systematic increase in viscosity is observed with increasing AUD content. Specifically, the viscosity of the IBOA:AUD 80:20 formulation is approximately 33 cP, whereas that of the 40:60 formulation reaches approximately 617 cP. The composition-dependent increase in viscosity

is attributed to the dominant contribution of AUD, which contains long polymer chains, in contrast to the short, monomeric nature of IBOA. Despite this increase, all formulations fall within viscosity ranges commonly reported to be suitable for DLP 3D printing.

Table S4 Compositions of the IBOA/AUD-based photocurable resins.

(IBOA:AUD)	IBOA (g)	AUD (g)	TPO (g)	Sudan-II (g)
40:60	12.0	18.0	0.3	0.03
45:65	13.5	16.5	0.3	0.03
50:50	15.0	15.0	0.3	0.03
55:45	16.5	13.5	0.3	0.03
60:40	18.0	12.0	0.3	0.03
65:35	19.5	10.5	0.3	0.03
70:30	21.0	9.0	0.3	0.03
75:25	22.5	7.5	0.3	0.03
80:20	24.0	6.0	0.3	0.03

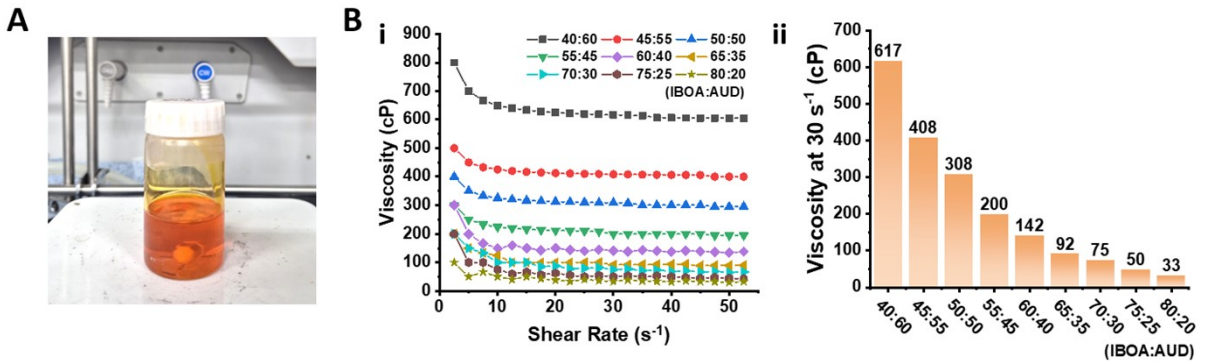


Figure S5 Resin formulations and rheological properties measurements. A Photograph of the liquid-state resin. **B** (i) Viscosity as a function of shear rate and (ii) viscosity measured at a shear rate of 30 s^{-1}.

S4.2 Fabrication, Activation and Characterization

Figure S6A illustrates the configuration of a bilayer specimen composed of active and passive phases, designed for curvature evaluation after actuation for different compositions. For clear visualization of the pattern, the schematic is shown in a proportionally scaled form. For actual

fabrication via grayscale digital light processing (g-DLP), the aspect ratio of the specimen was adjusted and the pattern was converted into a grayscale image (**Figure S6B-i**). In this image, the thick block on the left corresponds to a handle and is not included in the effective specimen length. Accordingly, as shown in **Figure S6B-ii**, the fabricated specimens have an aspect ratio (length:width) of 50:1, which is comparable to the ratio of 576:12 used in the simulations. During g-DLP printing, the exposure energy dose was fixed. The active phase was assigned a grayscale value of 255 (RGB basis), corresponding to the maximum exposure dose of 33.2 mJ/cm^2, while the passive phase was assigned a grayscale value of 135, corresponding to approximately 51% of the maximum exposure dose.

Thermal activation was carried out in an oven at 80 °C for 10 h. To prevent rigid-body motion during thermal treatment and to ensure effective deformation, the handle of the specimen was fixed in advance using an adhesive. **Figure S6C** shows the deformed shapes of the g-DLP-printed specimens after thermal activation for all nine compositions with different IBOA:AUD ratios. In each image, the dashed black circle represents the best-fit circle corresponding to the deformed geometry. The radius of this fitted circle was used to quantify the curvature (1/R) of the specimen.

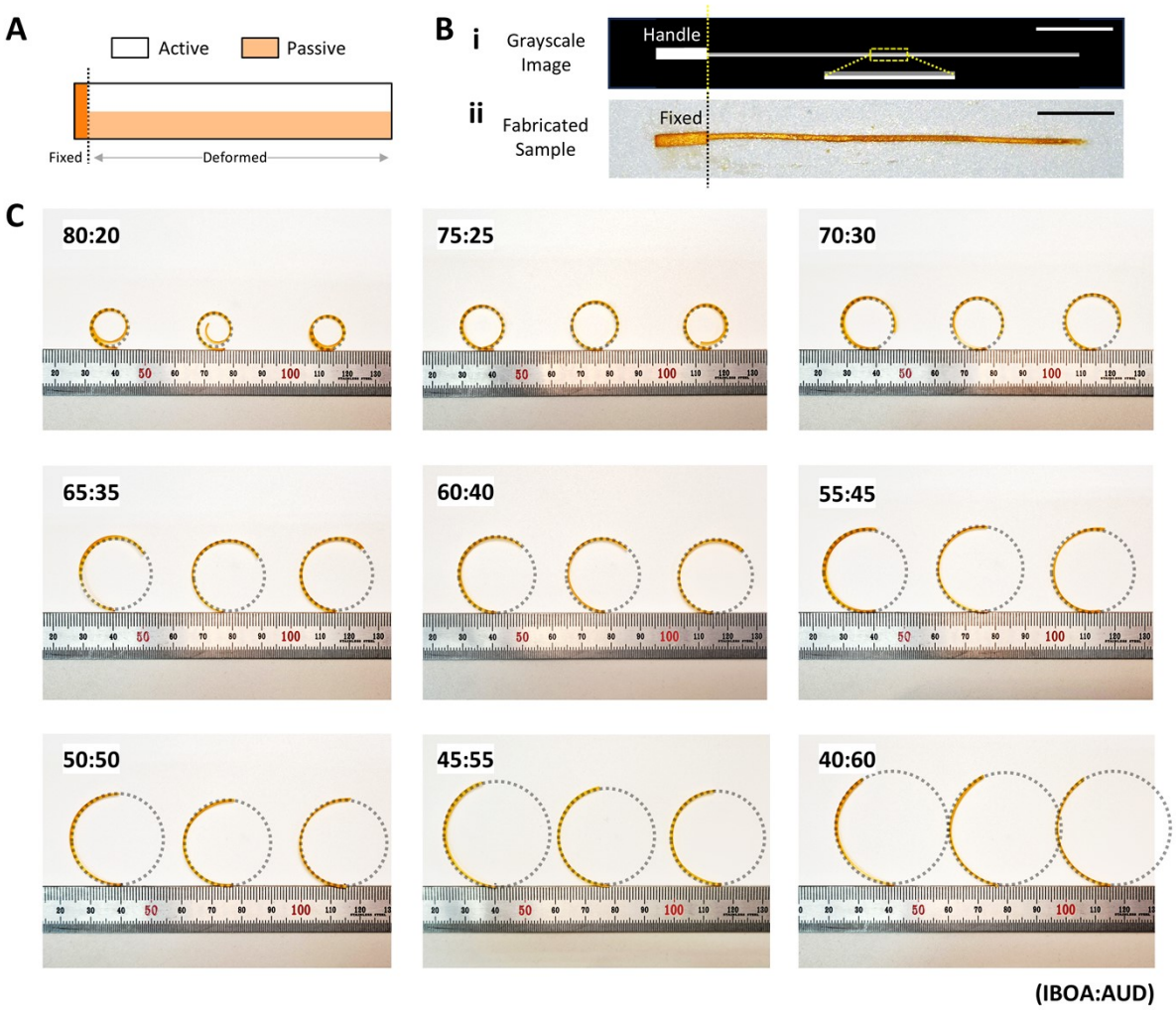


Figure S6 g-DLP printing of bilayer specimens and characterization. A Design of a bilayer specimen. **B** (i) Grayscale image used for g-DLP fabrication and (ii) the fabricated specimen. Each scale bar corresponds to 10 mm. **C** Deformed shapes of g-DLP-printed specimens after thermal activation for nine different IBOA:AUD compositions. Dashed circles are fitted circles used to calculate curvature.

To characterize the temperature-dependent mechanical properties of the two material phases and validate the material suitability of the selected actuation temperature, dynamic mechanical analysis (DMA) was additionally performed on separately fabricated specimens of g255 (active phase) and g135 (passive phase). As shown in **Figure S7**, g255 (active phase) exhibits a pronounced reduction in G' near its T_g of 50.6°C, transitioning from a glassy to a rubbery state, while g135 (passive phase) maintains relatively low G' values across the measured temperature

range owing to its lower T_g of 31.9°C. Below 60°C, g255 (active phase) maintains a substantially higher G' than g135 (passive phase). Above 60°C, both phases reach similarly low G' values, and the reduced stiffness of g255 (active phase) no longer suppresses the shrinkage-driven deformation of g135 (passive phase), allowing effective actuation to occur.

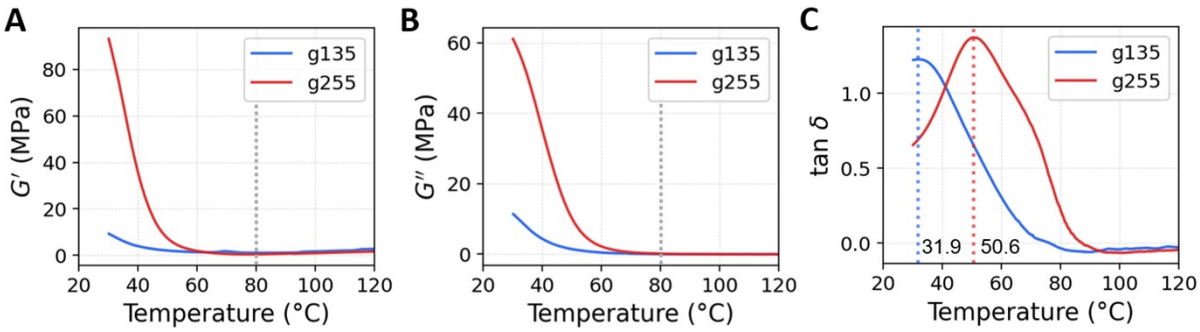


Figure S7 Dynamic mechanical analysis (DMA) results for the active (g255) and passive (g135) material phases. A Storage modulus (G'), **B** loss modulus (G''), and **C** loss tangent (tan δ) as a function of temperature. The peak of tan δ indicates a glass transition temperature (T_g) of 31.9°C and 50.6°C for the passive and active phases, respectively. Dashed vertical lines in **A** and **B** indicate the selected actuation temperature of 80°C.

To experimentally validate the actuation performance at the selected temperature, specimens fabricated at the selected 75:25 IBOA:AUD ratio were thermally actuated at 20, 40, 60, and 80°C over immersion times of 0, 2, 4, 6, 8, and 10 h (**Figure S8**). At 20°C, no deformation was observed regardless of immersion time. At 40°C and 60°C, progressive deformation developed with time, with the degree of deformation increasing with temperature. However, the deformed shapes at these temperatures exhibited non-uniform curvature, with deformation concentrated near the free end (tip) and diminishing toward the fixed end, deviating from the intended uniform curvature distribution. At 80°C, rapid and pronounced deformation was observed from as early as 2 h, with the shape approaching saturation by approximately 8–10 h. The contrast between 60°C and 80°C is visually prominent, with 60°C resulting in notably incomplete and

non-uniform deformation, confirming that 80°C is necessary to achieve spatially uniform actuation, with 10 h representing the duration at which deformation reaches saturation.

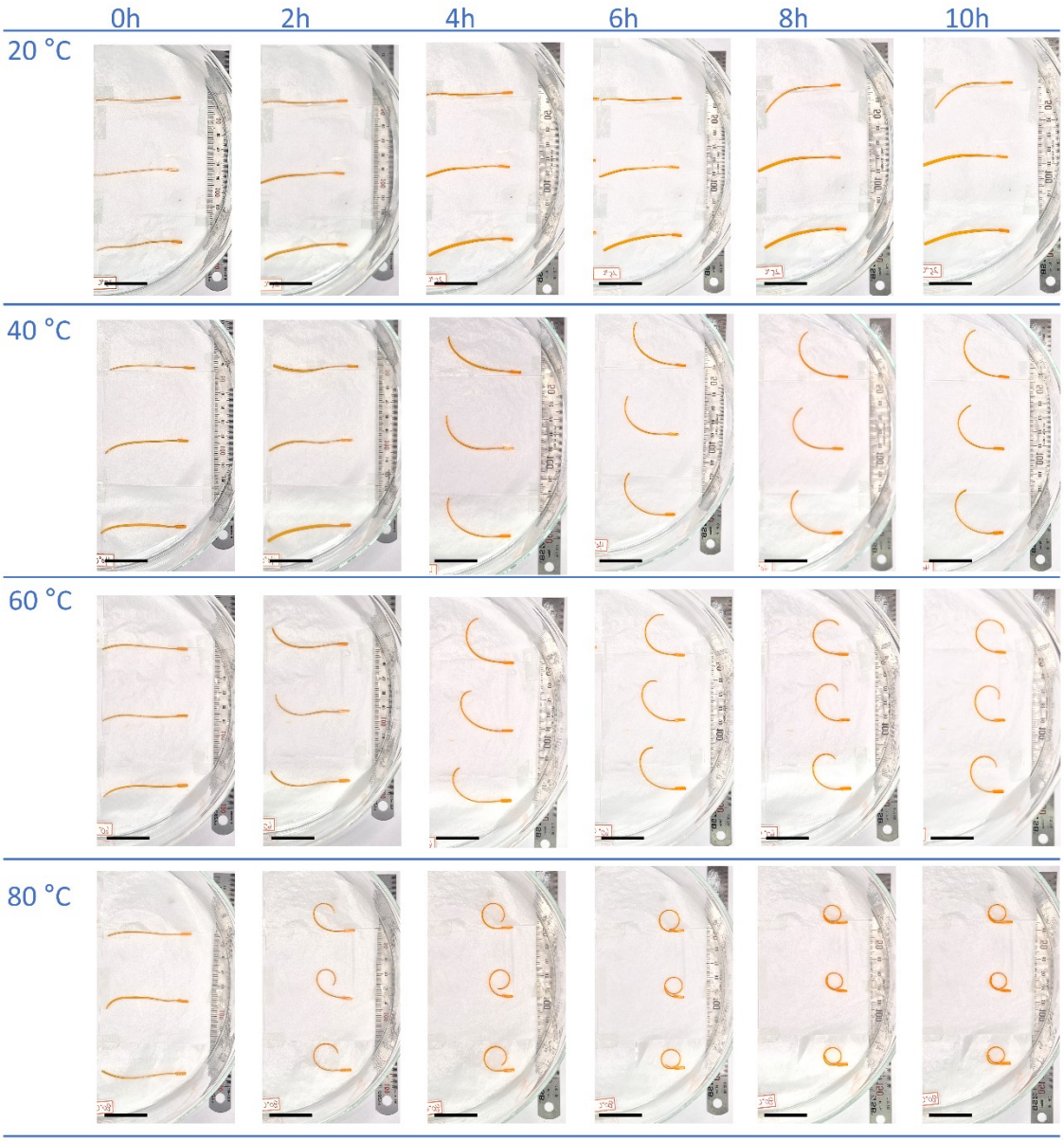


Figure S8. Temperature- and time-dependent actuation behavior of the printed specimen. Photographs of the specimen immersed in water at 20, 40, 60, and 80°C (rows) captured at 0, 2, 4, 6, 8, and 10 hours (columns). Scale bar: 30 mm.

S4.3 Experimental Methods

Materials: Isobornyl acrylate (IBOA) was obtained from Sigma-Aldrich and used as received. Aliphatic urethane diacrylate (AUD, Ebecryl 8402) was obtained from Allnex and used as received. Diphenyl(2,4,6-trimethylbenzoyl)phosphine oxide (TPO) and Sudan I dye was obtained from Sigma-Aldrich and used as received.

Resin characterizations: The viscosities of the ceramic suspensions were measured using a viscometer (DV next, Brookfield). The shear-rate range was set from 2.5 to 5.5 s^{-1}. The spindle rotation speed ranged from 10 to 250 rpm in increments of 10 rpm for each measurement and each rpm setting was measured for 30 s.

Digital light processing 3D printing: DLP printing was performed using an IMC57 from Carima, with a 405 nm UV light source, under room temperature.

S4.4 Quantitative Comparison between RL-Designed and Experimentally Fabricated Trajectories

To complement the visual comparison in Figure 8, a quantitative analysis was performed to evaluate the agreement between the RL-designed trajectories and the experimentally fabricated specimens (**Figure S9**).

Each experimental photograph was processed using HSV-based color thresholding, followed by morphological filtering and skeletonization to extract a one-pixel-wide centerline. The extracted centerline was then registered to the RL coordinate system using bounding-box-based scaling, where independent scaling was applied along the x- and y-directions.

The agreement between the RL-designed trajectory X and the experimental centerline Y was quantified using the symmetric bidirectional Chamfer distance in root-mean-square (RMS) form CD_{RMS}, defined as follows:

$$CD_{RMS} = \sqrt{\frac{\frac{1}{|X|}\sum_{x \in X} d(x,Y)^2 + \frac{1}{|Y|}\sum_{y \in Y} d(y,X)^2}{2}} \tag{S7}$$

where $d(x,Y)$ denotes the nearest-neighbor distance from point x to the point set Y, defined as

$$d(x,Y) = \min_{y \in Y} \|x - y\| \tag{S8}$$

In addition, a normalized error $E_{norm}(\%)$ was defined as

$$E_{norm}(\%) = \frac{1}{|X|}\sum_{x \in X} d(x,Y)/L \times 100 \tag{S9}$$

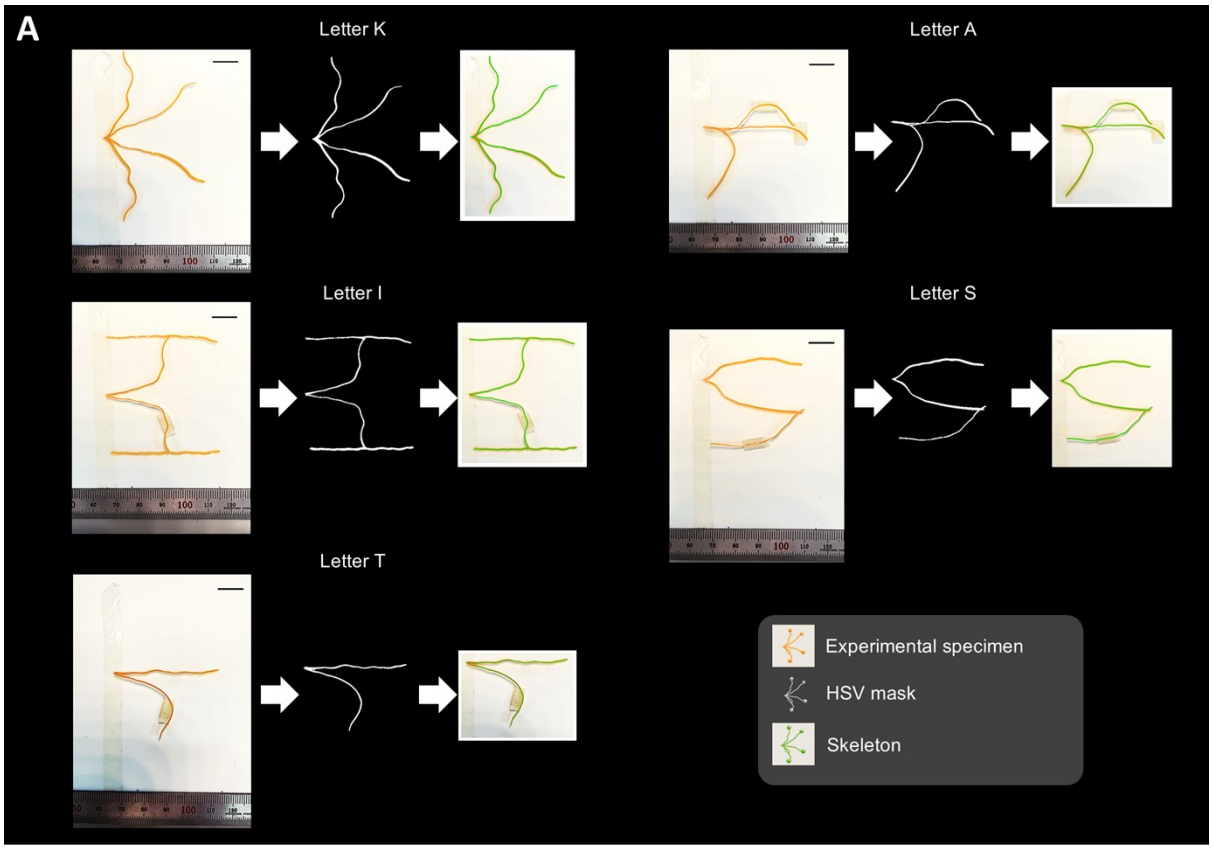
where L is the characteristic length of the RL trajectory, defined as the x-extent of the assembled trajectory.

Figure S9 shows the registered experimental centerlines (gray) overlaid with the RL-designed trajectories (green), and **Table S5** summarizes the results. $E_{norm}(\%)$ remain below 5% for all letters, indicating good agreement between the RL-designed trajectories and the fabricated specimens.

Table S5 Quantitative comparison between RL-designed trajectories and experimentally fabricated specimens.

Letter	CD_{RMS} **(RL units)**	$E_{norm}(\%)$
K	0.287	4.56
A	0.109	1.53
I	0.103	1.21
S	0.328	4.90

T	0.098	1.13


A
Letter K
Letter A
Letter I
Letter S
Letter T
Experimental specimen
HSV mask
Skeleton

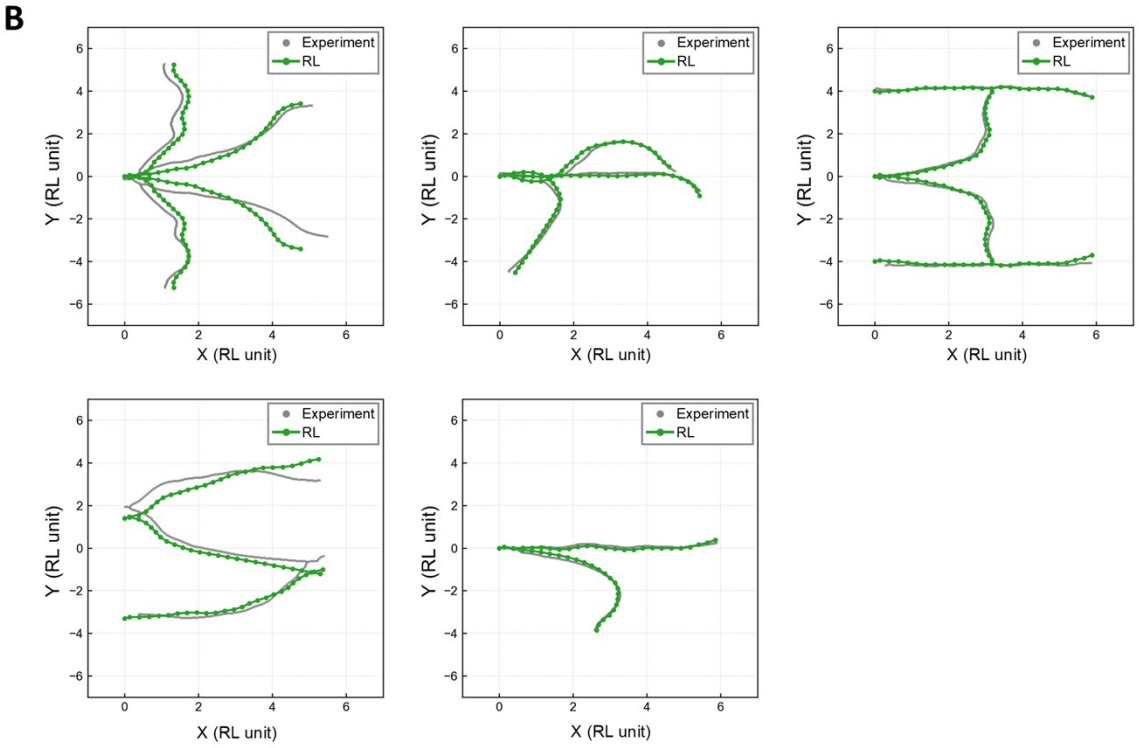

B
Experiment
RL
Y (RL unit)
X (RL unit)

Figure S9 Quantitative comparison between RL-designed and experimentally fabricated trajectories. A Processing pipeline: experimental photograph → HSV mask → skeletonized centerline (green overlay). **B** Registered experimental skeleton (gray points) overlaid with RL-designed trajectories (green) in the RL coordinate system for letters K, A, I, S, and T.

Branch-wise repeatability assessment was additionally performed for representative letters K and I. These two letters encompass a broad range of curvature, as their branches exhibit geometrically distinct profiles spanning from nearly straight to highly curved segments. For each branch, three independent fabrications were conducted to evaluate variance across repeated trials.

The image processing pipeline follows the same HSV-based skeletonization and bounding-box registration procedure described in Section S4.4 above. The algorithm-based automated registration minimizes measurement uncertainty in trajectory extraction and enables more objective accuracy evaluation compared to manual visual alignment. Since the evaluation is conducted on individual branches rather than the full assembled letter, an additional rotation transformation was applied prior to registration to align each branch to its corresponding RL trajectory orientation. The normalized mean error was then computed for each trial using Eq. (S9), enabling quantitative and objective assessment of shape agreement between the RL-designed trajectories and the experimentally fabricated specimens.

Figure S10A shows the results of three repeated fabrications for each of the two branches, (i) K_branch1 and (ii) K_branch2, of letter K. Figure S10B shows the results of three repeated fabrications for each of the two branches, (i) I_branch1 and (ii) I_branch2, of letter I. The leftmost column of each row presents experimental photographs of the repeatedly fabricated specimens, and the subsequent three columns sequentially show the overlay of the experimentally extracted centerline (gray) projected onto the RL-designed trajectory (green)

for each specimen. **Table S6** summarizes the mean and standard deviation of the Sym. Chamfer RMSE and normalized error for each branch. Even for the two branches of K, which exhibit complex multi-curvature geometry, the normalized errors were $3.01 \pm 0.57\%$ and $2.08 \pm 0.25\%$, respectively, while the branches of I showed errors within approximately 1%. All branches maintained normalized errors below 3%, and the low standard deviations across trials confirm stable reproducibility across varied curvature profiles, demonstrating that the fabrication process yields consistent results regardless of trajectory complexity.

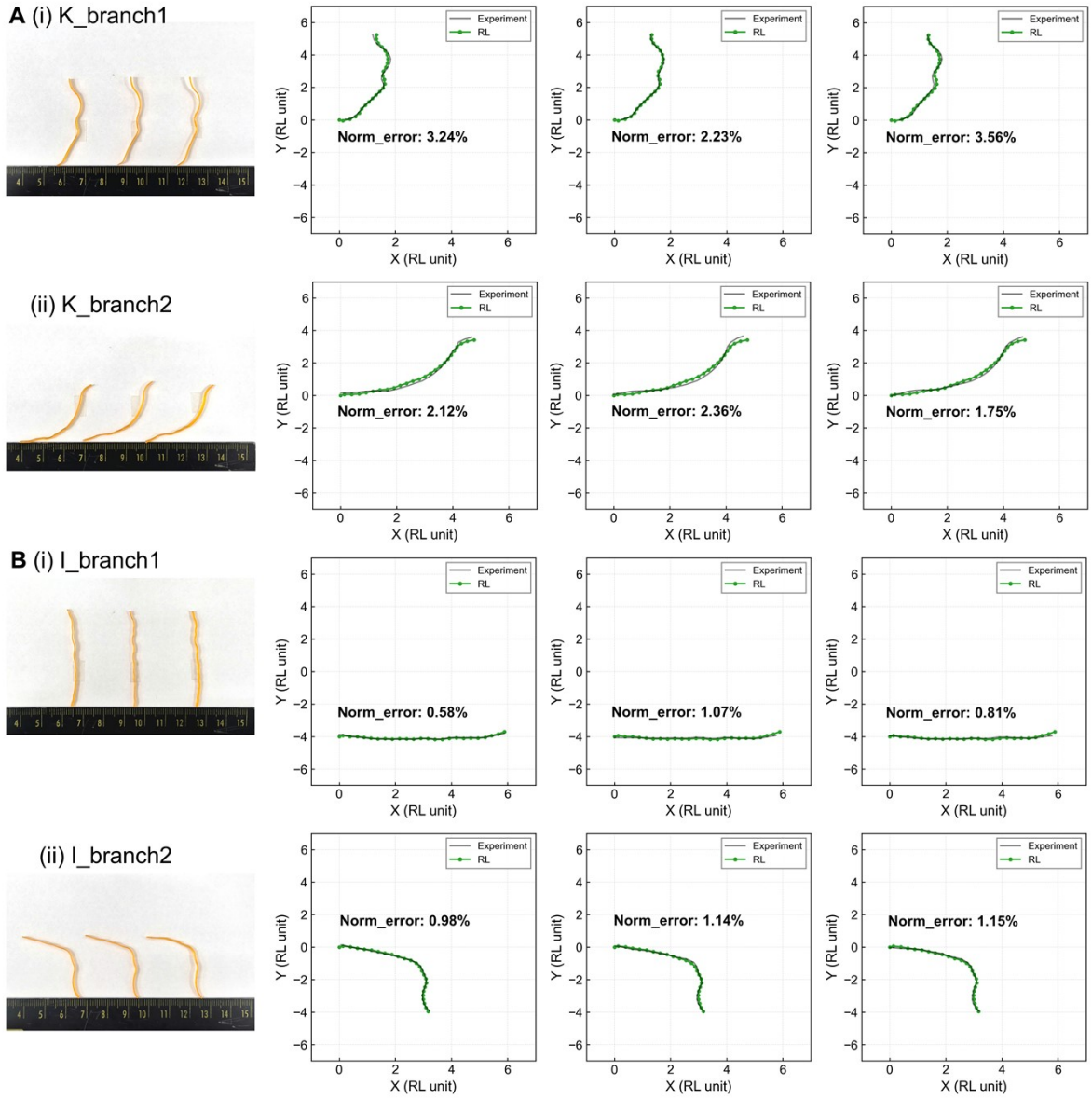


Figure S10 Quantitative comparison between RL-designed and experimentally fabricated trajectories on a per-branch basis. A Letter "K" consisting of two branch elements (K_branch1, K_branch2). **B** Letter "I" consisting of two branch elements (I_branch1, I_branch2). For each branch, three repeated fabrications (columns) are shown to assess reproducibility. Left: experimental photographs of the fabricated structures. Right: overlay of the experimentally extracted centerline (gray) and the RL-designed trajectory (green) in the RL coordinate system. $E_{norm}(\%)$ is indicated in each panel.

Table S6 Quantitative repeatability assessment of experimentally fabricated trajectories on a per-branch basis

Branch	CD_{RMS} **(RL units)**	$E_{norm}(\%)$
K_branch1	0.085 ± 0.008	3.01 ± 0.57
K_branch2	0.130 ± 0.012	2.08 ± 0.25
I_branch1	0.078 ± 0.009	0.82 ± 0.20
I_branch2	0.069 ± 0.001	1.09 ± 0.08

S5. LIMITATIONS AND FUTURE DIRECTIONS

Although the proposed RL-based inverse-design framework effectively demonstrates the feasibility of sequential inverse design for thermally active composites, it still has several limitations. First, the present framework was developed for a static configuration problem under fixed boundary and thermal conditions. In such a static setting, the advantages of reinforcement learning, including policy adaptability, online updating, and continual learning, cannot be fully exploited. Because the present framework is formulated as a target-conditioned sequential decision process, it can in principle be extended to settings with time-varying targets or constraints, where the strengths of RL would become more directly beneficial.

A second limitation lies in the current design discretization. In the present study, a relatively coarse 4×24 binary grid was adopted as a tractable benchmark for comparing RL, GA, and SSO under the same combinatorial design space. While this setting was set to demonstrate the feasibility of the proposed framework, the limited number of rows constrains the diversity of through-thickness material layouts and can therefore restrict the fidelity of more complex shape transformations. Increasing the number of rows or adopting higher-resolution printing approaches would enable richer deformation modes and more accurate realization of intricate target geometries. However, such extensions would also substantially enlarge the combinatorial design space and increase the complexity of the sequential decision problem. Therefore, additional methodological development of the RL framework will be required to efficiently address higher-resolution design problems.

A third limitation lies in computational efficiency. Although RL exhibits higher sample efficiency by requiring fewer function evaluations, the total wall-clock time can be longer than that of GA or SSO because of the neural-network training overhead at every update step. This

trade-off between sample efficiency and computation time becomes critical for scaling to larger design domains.

Fourth, during multi-target training, several cases failed to meet the termination criterion (RMSE $\leq$ 0.1), most notably trajectories involving sharp reversals, as exemplified by Case 3 in **Figure 6**. This limitation is attributed mainly to data sparsity in rare deformation patterns such as reverse-bending trajectories. As a result, such target regimes were insufficiently represented during training episodes, which likely limited the ability of the RL agent to learn a sufficiently balanced policy over the full target distribution.

Future improvements may therefore be achieved by defining the design space more systematically so that such rare deformation modes are sufficiently included, constructing the dataset by sampling that space in a more uniform and representative manner, fine-tuning the RL agent's hyperparameters, and employing continuous-action algorithms (e.g., Proximal Policy Optimization or Soft Actor-Critic) or hierarchical/goal-conditioned architectures to improve precision and stability.

In summary, the current RL framework effectively demonstrates the feasibility of sequential inverse design for thermally active composites, but its scalability, design-resolution extensibility, and dynamic applicability remain open challenges. Nevertheless, this study provides a foundation for extending the framework toward broader design spaces, finer structural discretization, more representative datasets, and dynamic inverse-design problems, ultimately enabling more robust and practical applications in complex engineering systems.

References

1 X. Sun, L. Yue, L. Yu, H. Shao, X. Peng, K. Zhou, F. Demoly, R. Zhao and H. J. Qi, *Adv. Funct. Mater.*, DOI:10.1002/adfm.202109805.

2 X. Sun, L. Yu, L. Yue, K. Zhou, F. Demoly, R. R. Zhao and H. J. Qi, *J. Mech. Phys. Solids*, DOI:10.1016/j.jmps.2024.105561.

3 T. Poltue, C. Zhang, F. Demoly, K. Zhou and H. J. Qi, *Advanced Intelligent Systems*, DOI:10.1002/aisy.202500916.